Universidad Politècnica de Catalunya

Progama de Doctorat:

Automàtica, robòtica I visió

Tesi Doctoral

SENSORIMOTOR EXPLORATION: CONSTRAINT AWARENESS AND SOCIAL REINFORCEMENT IN EARLY VOCAL DEVELOPMENT

Juan Manuel Acevedo Valle

Director de tesi:
Dr. Cecilio Angulo Bahón

Setembre de 2018

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Sensorimotor Exploration: Constraint Awareness and Social Reinforcement in Artificial Vocal Development

Doctoral Program:

Automatic Control, Robotics and Computer Vision

Juan M. Acevedo-Valle

Supervisor:

Dr. Cecilio Angulo Bahón

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy
by the
UNIVERSITAT POLITÈCNICA DE CATALUNYA
in the
Knowledge Engineering Research Group
Intelligent Data Science and Artificial Intelligence Research Center
Automatic Control Department

Barcelona, September 2018

A mi familia

(Para leer en forma interrogativa)
Has visto,
verdaderamente has visto
la nieve, los astros, los pasos afelpados de la brisa...
Has tocado,
de verdad has tocado
el plato, el pan, la cara de esa mujer que tanto amás...
Has vivido
como un golpe en la frente,
el instante, el jadeo, la caída, la fuga...
Has sabido
con cada poro de la piel, sabido
que tus ojos, tus manos, tu sexo, tu blando corazón,
había que tirarlos
había que llorarlos
había que inventarlos otra vez.

- Julio Cortázar

Acknowledgement

First of all, I am grateful to my advisor, Prof. Dr. Cecilio Angulo, for believing in this project and supporting it from the beginning till the end of the journey. I am also thankful to Prof. Dr. Verena Hafner and the members of her team for the knowledge they shared with me during my stage at the Humboldt University of Berlin. I am also grateful to Dr. Diego Pardo and Prof. Dr. Jonas Buchli for inviting me to collaborate with them at the ADLR-lab at ETHZ. An special mention and acknowledgement to the international reviewers of this thesis Prof. Dr. Bruno Lara and Dr. Guido Schillaci. I am also thankful to Dr. Clément Moulin-Frier for sharing with me his expertise in the field of developmental robotics at the beginning of this project.

I am thankful to the GREC research group's members for all the discussions during our seminars and meetings, specially Karla and Jenn. I express my gratitude to many friends, cousins, and colleagues that at different moments collaborated with knowledge, emotional support, and fun. I will be always in debt with Karla and Victor, when I was homeless after my stages abroad they opened the doors of their homes to me, they helped me in so many ways. Together with Julio and Noe you filled these years with laughter, encouragement and inspiration, during the last six years, thank you guys. I am also grateful to my Mexican crew. I am also grateful to Claudi, his friendship through these years has encouraged me in so many ways and enriched my knowledge.

To Anna, Daria, Hubert and Marcin (my wonderful polish crew) for being always there, for being family, for all the Christmas, Easters, New years's eves, and summers we spent together, from pseudo-snowboarding in the Polish mountains, till sailing the Adriatic sea and, literally, almost dying in the pursuing of happiness and adventures.

Finally, I express my gratitude to Mexico and CONACyT for providing the resources to reach this milestone in my professional life. My gratitude is also for the Swiss National Centre of Competence and Research and Fundació "la Caixa" for funding my stages in Zurich and Berlin, respectively.

El estado actual de mi vida, un estado más bien observable, no podría explicarse sin el amor y apoyo recibidos a lo largo de estos años de parte de mi familia, a ellos les dedico cada uno de los esfuerzo hechos para concluir este proyecto, y les agradezco muchas cosas más. A mi

papá, Lorenzin, que me dejó tantas enseñanzas y regalos, y a veces en sueños no deja de aconsejarme. A mi mamá, Chalo, que con sus oraciones y su hambre infinita de sacar a sus hijos adelante ha sido la principal benefactora de todos estos años. A mis hermanos, Elena y Herminio, a quienes admiro por sus éxitos profesionales y su calidad humana. No perdieron cualquier oportunidad de acompañarme con visitas cortas, o estancias largas a lo largo de estos años. Gracias por los viajes, por el apoyo, y por mantenerme cerca de casa a pesar de la distancia.

Aún con todas las experiencias y conocimientos cosechados, de la felicidad inegable, los años fuera de casa sin duda representan un costo muy alto, por lo que cada minuto lejos ha tenido que valerlo. Sin duda agradezco a grandes amigos que se han mantenido cercanos a la familia y nos han apoyado en momentos llenos de alegrías y sobre todo en momentos díficiles: Don Gustavo, Doña Tere, Sergio Rodríguez. En este sentido gracias a mi hermano Ismael y a su mamá.

Agradezo a Alejandra, por haberme acompañado y dado tanta fuerza a lo largo de la última etapa de este proyecto. Por embarcarse conmigo en una de las aventuras más grandes y hermosa de nuestras vidas. Por sostenerme en mis tambaleos y por llenar mi vida de alegría, redefiniendo tantas palabras y lugares.

Juan Manuel Acevedo Valle
Barcelona, September 2018

Abstract

This research is motivated by the benefits that knowledge regarding early development in infants may provide to different fields of science. In particular, early sensorimotor exploration behaviors are studied in the framework of developmental robotics. The main objective is about understanding the role of motor constraint awareness and imitative behaviors during sensorimotor exploration. Particular emphasis is placed on prelinguistic vocal development because during this stage infants start to master the motor systems that will later allow them to pronounce their first words.

Previous works have demonstrated that goal-directed intrinsically motivated sensorimotor exploration is an essential element for sensorimotor control learning. Moreover, evidence coming from biological sciences strongly suggests that knowledge acquisition is shaped by the environment in which an agent is embedded and the embodiment of the agent itself, including developmental processes that shape what can be learned and when.

In this dissertation, we firstly provide a collection of theoretical evidence that supports the relevance of our study. Starting from concepts of cognitive and developmental sciences, we arrived at the conclusion that spoken language, i.e., early vocal development, must be studied as an embodied and situated phenomena. Considering a synthetic approach allow us to use robots and realistic simulators as artifacts to study natural cognitive phenomena. In this work, we adopt a toy example to test our cognitive architectures and a speech synthesizer that mimics the mechanisms by which humans produce speech.

Next, we introduce a mechanism to endow embodied agents with motor constraint awareness. Intrinsic motivation has been studied as an important element to explain the emergence of structured developmental stages during early vocal development. However, previous studies failed to acknowledge the constraints imposed by the embodiment and situatedness, at sensory, motor, cognitive and social levels. We assume that during the onset of sensorimotor exploratory behaviors, motor constraints are unknown to the developmental agent. Thus, the agent must discover and learn during exploration what those motor constraints are. The agent is endowed with a somesthetic system based on tactile information. This system generates a sensor signal indicating if a motor configuration was reached or not. This information is later used to create a somesthetic model to predict constraint violations.

Finally, we propose to include social reinforcement during exploration. Some works studying early vocal development have shown that environmental speech shapes the sensory space explored during babbling. More generally, imitative behaviors have been demonstrated to be crucial for early development in children as they constraint the search space during sensorimotor exploration. Therefore, based on early interactions of infants and caregivers we proposed an imitative mechanism to reinforce intrinsically motivated sensorimotor exploration with relevant sensory units. Thus, we modified the constraints aware sensorimotor exploration architecture to include a social instructor, expert in sensor units relevant to communication, which interacts with the developmental agent. Interaction occurs when the learner production is 'enough' similar to one relevant to communication. In that case, the instructor perceives this similitude and reformulates with the relevant sensor unit. When the learner perceives an utterance by the instructor, it attempts to imitate it.

In general, our results suggest that somesthetic senses and social reinforcement contribute to achieving better results during intrinsically motivated exploration. Achieving less redundant exploration, decreasing exploration and evaluation errors, as well as showing a clearer picture of developmental transitions.

Keywords:

Developmental Robotics, Artificial Vocal Development, Sensorimotor Exploration, Constraint Awareness, Social Reinforcement, Incremental Learning, Gaussian Mixture Models, Speech Technologies, Language, Sensorimotor Contingencies

Resumen

La motivación principal de este trabajo es la magnitud que las contribuciones al conocimiento en relación al desarrollo infantil pueden aportar a diferentes campos de la ciencia. Particularmente, este trabajo se enfoca en el estudio de los comportamientos de autoexploración sensorimotora en un marco robótico e inspirado en el campo de la psicología del desarrollo. Nuestro objetivo principal es entender el papel que juegan las restricciones motoras y los reflejos imitativos durante la exploración espontánea observada en infantes. Así mismo, este trabajo hace especial énfasis en el desarrollo vocal-auditivo en infantes, que les provee con las herramientas que les permitirán producir sus primeras palabras.

Trabajos anteriores han demostrado que los comportamientos de autoexploración sensorimotora en niños, la cual ocurre en gran medida por motivaciones intrínsecas, es un elemento importante para aprender a controlar su cuerpo con tal de alcanzar estados sensoriales específicos. Además, evidencia obtenida de estudios biológicos sugiere tajantemente que la adquisición de conocimiento es regulada por el ambiente en el cual un agente cognitivo se desenvuelve y por el cuerpo del agente *per se*. Incluso, los procesos de desarrollo que ocurren a nivel físico, cognitivo y social también regulan que es aprendido y cuando esto es aprendido.

La primera parte de este trabajo provee al lector con la evidencia teórica y práctica que demuestran la relevancia de esta investigación. Recorriendo conceptos que van desde las ciencias cognitivas y del desarrollo, llegamos a la conclusión de que el lenguaje, y por tanto el habla, deben ser estudiados como fenómenos cognitivos que requieren un cuerpo físico y además un ambiente propicio para su existencia. En la actualidad los sistemas robóticos, reales y simulados, pueden ser considerados como elementos para el estudio de los fenómenos cognitivos naturales. En este trabajo consideramos un ejemplo simple para probar las arquitecturas cognitivas que proponemos, y posteriormente utilizamos dichas arquitecturas con un sintetizador de voz similar al mecanismo humano de producción del habla.

Como primera contribución de este trabajo proponemos introducir un mecanismo para construir robots capaces de considerar sus propias restricciones motoras durante la etapa de autoexploración sensorimotora. Ciertos mecanismos de motivación intrínseca para exploración sensorimotora han sido estudiados como posibles conductores de las trayectorias de desarrollo observadas durante el desarrollo temprano del habla. Sin embargo, en previos

estudios no se consideró que este desarrollo está delimitado por restricciones debido al ambiente, al cuerpo físico, y a las capacidades sensoriales, motoras y cognitivas. En nuestra arquitectura, asumimos que un agente artificial no cuenta con conocimiento de sus limitantes motoras, y por tanto debe descubrirlas durante la etapa de autoexploración. Para tal efecto, el agente es proveído de un sistema somatosensorial que le indica cuando una configuración motora viola las restricciones impuestas por el propio cuerpo.

Finalmente, como segunda parte de nuestra contribución proponemos incluir un mecanismo para reforzar el aprendizaje durante la autoexploración. Estudios anteriores demostraron que el ambiente lingüístico en que se desarrolla un infante, o un agente artificial, condiciona sus producciones vocales durante la autoexploración o balbuceo. En este trabajo nos enfocamos en el estudio de episodios de imitación que ocurren durante el desarrollo temprano de un agente. Basados en estudios sobre la interacción entre madres e hijos durante la etapa prelingüística, proponemos un mecanismo para reforzar el aprendizaje durante la autoexploración con unidades sensoriales relevantes. Entonces, a partir de la arquitectura con autoconocimiento de restricciones motores, construimos una arquitectura que incluye un instructor experto en control sensorimotor. Las interacciones entre el aprendiz y el experto ocurren cuando el aprendiz produce una unidad sensorial relevante para la comunicación durante la autoexploración. En este caso, el experto percibe esta similitud y responde reformulando la producción del aprendiz como la unidad relevante. Cuando el aprendiz percibe una acción del experto, inmediatamente intenta imitarlo.

Los resultados presentados en este trabajo sugieren que los sistemas somatosensoriales y el reforzamiento social contribuyen a lograr mejores resultados durante la etapa de autoexploración sensorimotora motivada intrínsicamente. En este sentido, se logra una exploración menos redundante, los errores de exploración y evaluación disminuyen, y por último se obtiene una imagen más nítida de las transiciones entre etapas del desarrollo.

Resum

La motivació principal d'aquest treball és la magnitud que les contribucions al coneixement en relació al desenvolupament infantil poden aportar a diferents camps de la ciència. Particularment, aquest treball s'enfoca en l'estudi dels comportaments d'autoexploració sensorimotora en un marc robòtic i inspirat en el camp de la psicologia del desenvolupament. El nostre objectiu principal és entendre el paper que juguen les restriccions motores i els reflexos imitatius durant l'exploració espontània observada en infants. Així mateix, aquest treball fa especial èmfasi en el desenvolupament vocal-auditiu en infants, que els proveeix amb les eines que els permetran produir les seves primeres paraules.

Treballs anteriors han demostrat que els comportaments d'autoexploració sensorimotora en nens, la qual ocorre en gran mesura per motivacions intrínseques, és un element important per aprendre a controlar el seu cos per tal d'assolir estats sensorials específics. A més, evidències obtingudes d'estudis biològics suggereixen que l'adquisició de coneixement és regulada per l'ambient en el qual un agent cognitiu es desenvolupa i pel cos de l'agent *per se*. Fins i tot, els processos de desenvolupament que ocorren a nivell físic, cognitiu i social també regulen què és après i quan això és après.

La primera part d'aquest treball proveeix el lector amb les evidències teòrica i pràctica que demostren la rellevància d'aquesta investigació. Recorrent conceptes que van des de les ciències cognitives i del desenvolupament, vam arribar a la conclusió que el llenguatge, i per tant la parla, han de ser estudiats com a fenòmens cognitius que requereixen un cos físic i a més un ambient propici per a la seva existència. En l'actualitat els sistemes robòtics, reals i simulats, poden ser considerats com a elements per a l'estudi dels fenòmens cognitius naturals. En aquest treball considerem un exemple simple per provar les arquitectures cognitives que proposem, i posteriorment utilitzem aquestes arquitectures amb un sintetitzador de veu similar al mecanisme humà de producció de la parla.

Com a primera contribució d'aquest treball proposem introduir un mecanisme per construir robots capaços de considerar les seves pròpies restriccions motores durant l'etapa d'autoexploració sensorimotora. Certs mecanismes de motivació intrínseca per exploració sensorimotora han estat estudiats com a possibles conductors de les trajectòries de desenvolupament observades durant el desenvolupament primerenc de la parla. No obstant això,

en previs estudis no es va considerar que aquest desenvolupament és delimitat per restriccions a causa de l'ambient, el cos físic, i les capacitats sensorials, motores i cognitives. A la nostra arquitectura, assumim que un agent artificial no compta amb coneixement dels seus limitants motors, i per tant ha de descobrir-los durant l'etapa d'autoexploració. Per a tal efecte, l'agent és proveït d'un sistema somatosensorial que li indica quan una configuració motora viola les restriccions imposades pel propi cos.

Finalment, com a segona part de la nostra contribució proposem incloure un mecanisme per reforçar l'aprenentatge durant l'autoexploració. Estudis anteriors han demostrat que l'ambient lingüístic en què es desenvolupa un infant, o un agent artificial, condiciona les seves produccions vocals durant l'autoexploració o balboteig. En aquest treball ens enfoquem en l'estudi d'episodis d'imitació que ocorren durant el desenvolupament primerenc d'un agent. Basats en estudis sobre la interacció entre mares i fills durant l'etapa prelingüística, proposem un mecanisme per reforçar l'aprenentatge durant l'autoexploració amb unitats sensorials rellevants. Aleshores, a partir de l'arquitectura amb autoconeixement de restriccions motors, vam construir una arquitectura que inclou un instructor expert en control sensorimotor. Les interaccions entre l'aprenent i l'expert, ocorren quan una producció sensorial de l'aprenent durant l'autoexploració és similar a una unitat sensorial rellevant per a la comunicació. En aquest cas, l'expert percep aquesta similitud i respon reformulant la producció de l'aprenent com la unitat rellevant. Quan l'aprenent percep una acció de l'expert, immediatament intenta imitar-lo.

Els resultats presentats en aquest treball suggereixen que els sistemes somatosensorials i el reforçament social contribueixen a aconseguir millors resultats durant l'etapa d'autoexploració sensorimotora motivada intrínsecament. En aquest sentit, s'aconsegueix una exploració menys redundant, els errors d'exploració i avaluació disminueixen, i finalment s'obté una imatge més nítida de les transicions entre etapes del desenvolupament

Contents

List of Figures

List of Tables

List of Algorithms

Nomenclature

Throughout this doctoral dissertation, all column vectors are denoted by bold lower case, e.g., $\mathbf{x}$. Matrices and datasets (represented as matrices) are denoted by bold uppercase, e.g., $\mathbf{A}$. Scalars are denoted by non-bold lower case, e.g., B. Real numbers are denoted by $\mathbb{R}$. Sensor and motor spaces are denoted using Euler script style, e.g., $\mathcal{S}$. The most relevant variables enlisted through this work are enlisted below.

α	Forgetting factor
c	Competence function
e_{av}	Average evaluation error
F_*	Speech signal formant frequencies
$\mathbf{m}$	Motor command
ma	Moving average
M_*	Model
$M_{\mathcal{SM}}$	Sensorimotor model
$M_{\mathcal{SS}}$	Somesthetic model
M_{IM}	Interest model
μ	Mean of a normal distribution
$\mathcal{N}$	Normal distribution
π	Prior probability of a normal distribution
p	Pain signal
$\mathbf{s}$	Sensory outcome
$\mathbf{s}_g$	Sensory goal
$\mathbf{S_{KLD}}$	Symmetrized Kullback-Leibler divergence
Σ	Covariance Matrix of a normal distribution
ucr	Undesired motor configuration ratio

List of Acronyms

ASR	Automatic Speech Recognition
BIC	Bayesian Information Criterion
CB	Canonical Babbling
DNN	Deep Neural Network
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
GMR	Gaussian Mixture Regression
HMM	Hidden Markov Models
HRI	Human-Robot Interaction
iGMM	Incremental Gaussian Mixture Model
KDE	Kernel-Distribution Estimation
KLD	Kullback-Leibler divergence
MTSP	Motor Theory of Speech Perception
MUAC	Maximal Use of Available Controls
NLP	Natural Language Processing
PACT	Perception-for-Action-Control Theory
PCA	Principal Component Analysis
SPT	Speech to Text Conversion
TSS	Text to Speech Synthesizing
wNN	weighted k-Nearest Neighbor

Chapter 1

Introduction

"People are mistaken when they think that technology just automatically improves. It does not automatically improve. It only improves if a lot of people work very hard to make it better, and actually it will, I think, by itself degrade, actually."
— Elon Musk

Despite an optimistic view about the future of robotics, robots have still not pervaded our daily life. A number of scientific issues are yet to be solved for robots to be able to efficiently behave in open and uncertain environments. Modern technological approaches have aimed at solving some of the critical issues to develop more complex robotic systems through understanding and modeling key cognitive processes in humans. Among the most challenging fields associated with robotics, one could find computer vision, navigation, motion control, and human-robot interaction.

During the last years, the robotic industry has rapidly grown. This growth hasbeen restricted to services (e.g., surgical robots), exploration and surveillance (e.g., autonomous submarines and drones), military and manufacturing. In fact, robotic systems have become a strategic element for those industries. Even though robots have been endowed with some autonomy to make decisions, those decisions are restricted to structured environments where sources of uncertainty are scarce, and humans are generally not physically involved in the task at hand.

More recently, there has been an increasing interest in using robots for domestic, rescuing tasks, more complex medical procedures, among other applications. In the years to come, robots are going to be necessary to solve important problems in our societies. Just consider the example of elderly people in Japan, by 2035 a third of the country's population will be 65 or older. Thus, the robotic industry has started to develop the technologies to fulfill the healthcare and nursing requirements that this situation will generate, opening also a door to a big business considering that other countries will face similar problems (Forster, 2018). Such applications require a robot to perform daily life human-like activities. Some robots have been developed for those areas, but they are neither effective nor efficient performing in unstructured environments and interacting with humans. We consider these handicaps as the main reason to prevent robots from "overrun" our homes, workplaces, streets, leisure spaces, among other places. From our perspective, if robots are intended to work side-by-side with humans, they require at least to satisfy three general aspects. First, they must operate safely for humans and themselves. Second, robots must efficiently perform the work they have been created to fulfill. Third, they must be able to efficiently interact (regarding the task they are designed for) with other agents (humans, animals, and other robots) becoming social artifacts.

Researchers are aware that, in the most complex scenarios of interaction, robots should be endowed with human-like communication mechanisms. Consequently, artificial speech and natural language technologies have been widely investigated. Recently Google Duplex was presented (Hyken, 2018), a human-like talking computer able to make calls to arrange a haircut appointment or book a table at a restaurant without the interlocutors suspecting that they are talking to a computer. Many questions may arise from Google later developments, technical, ethical and even philosophical. The important technical question at hand is *What would be the result of a call if the topic of conversation drifts apart from the original purpose of the call?* In general, what we observe in the release of Google Duplex is that the purposes of the calls are structured. Thus, the possibilities are constrained to a certain degree, therefore facilitating the development of a system performing well. To answer more questions about this new system we will have to wait until the app or technical reports will be granted access.

In spite on any criticism, Google results are impressive, but much of this advance in human-like conversational machines must be acknowledged to all the scientific community that has been working broadly on Speech to Text Conversion (STP) technologies, Natural Language

Processing (NLP) technologies and Text to Speech Synthesizing (TSS) technologies during decades. Advanced Automatic Speech Recognition (ASR) systems have emerged as a popular solution to solve the challenge of providing artificial agents with orders through spoken commands. The body of work on NLP and TSS also includes a broad collection of solutions. ASR, NLP and TSS systems with different capabilities have been successfully implemented in robots, computers, smartphones and other devices. However, they lack many relevant features of human language which make them only available for structured interactions.

On the other hand, besides satisfying a broad range of needs in the industry, advances in artificial intelligence and robotics, computers and robots have become essential tools used as means for studying the human mind. Machine learning techniques, fast robot-prototyping, and complex simulators have fostered the appeal of artificial agents for studying the mechanisms of cognitive development. In this sense, robots are built at least because of two different reasons: as useful artifacts or as scientific tools as discussed later in this work (Mirolli and Parisi, 2011).

In this work, we use the advances in machine learning, artificial bioinspired agents, and developmental psychology studies to investigate observed phenomena during the early development of infants. During this investigation, we use artificial developmental agents. In general, our study contributes to constructing an approach in which a robot can learn the relation between its motor actions and sensory consequences efficiently. Moreover, we emphasize the study of the emergence of developmental stages during early vocal development.

This work is a journey through the study of intrinsically motivated learning algorithms inspired by the developmental processes observed in infants. Nowadays, these algorithms are broadly used to learn sensorimotor coordination skills. The journey starts with the studies in Moulin-Frier et al. (2013), where intrinsically motivated sensorimotor learning was used to study the emergence of developmental stages in the course of prelinguistic vocal development. Inspired by Moulin-Frier et al. (2013), we extended the study of prelinguistic vocal development to understand the role of somesthetic senses in early development.

Following biological evidence, in Acevedo-Valle et al. (2015, 2018) we argued that somesthetic senses are an important element that should be considered during sensorimotor exploration as they are a good candidate to deal with motor constraints. Motor constraints, and constraints in general, are an important element that shapes the development of cognitive skills

as discussed later in this work. Finally, in Acevedo-Valle et al. (2017a, 2018), we argued that using a simple imitation scenario to reinforce sensorimotor exploration using intrinsic motivations and somesthesis improves the exploration performance. Social reinforcement is a crucial source of opportunities for early sensorimotor development, especially during early vocal development as discussed extensively later. It is hard to think of an infant learning to pronounce a word or even a simple language directed syllable without the guidance of an adult.

This introductory chapter establishes the area of robotics where the contribution of this work lies. In the following, the chapter is divided into six sections. First, in Section 1.1 we establish a general link between cognitive and developmental sciences with robotics. Next, Section 1.2 remarks the relevance of studying speech and language development, not with the aim of manufacturing more complex interaction artifacts, but also as a mean to contribute to the study of the human mind. Sections 1.3-1.5 clarifies the motivations, objectives and contributions of this work. Finally, Section 1.7 describes the structure of this thesis dissertation.

1.1 From Cognitive Sciences to Developmental Robotics

As defined in the dictionary, *cognition is the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses* (Oxford Dictionaries, 2018). In this section, we study, how sciences study cognition have contributed to building adaptive intelligent robots. In Figure 1.1, we describe the concepts that link cognitive science and developmental science with different approaches to Robotics. In recent years, the idea of *embodied cognition* has become popular among sciences that, from different perspectives, study cognition (Wilson and Golonka, 2013). As defended in Galantucci et al. (2006), cognition, like all the products of evolution, cannot be understood in isolation. Instead, understanding cognition requires comprehending that it is embedded in a meaningful ecological context and embodied systems (Liberman and Mattingly, 1985).

As a cognitive system, the human mind could be studied as a dynamical system. The state of this complex system is determined by the interaction between several building blocks, i.e., memory, attention, motor control, perception, emotions, among other. Regarding embodiment and cognitive system's situatedness, at least motor control and perception might be

considered as systems that together build up our mechanisms to experience the environment. As mentioned in Pfeifer et al. (2007), autonomous robot design could notably benefit from the available knowledge of biological science and self-organization theories. As indicated in Figure 1.1, one of the most relevant sciences to be considered are cognition and developmental psychology. The existing knowledge about human cognition, and natural cognition in general, has been a constant inspiration for the development of intelligent machines, leading to the concept of cognitive robotics.

Different definitions could be found for cognitive robotics. In Mirolli and Parisi (2011) it is simply defined as the study of cognitive phenomena by their modeling in robotic systems. In this case, cognitive robots are considered as scientific tools, which is observed in Figure 1.1 as a feedback loop from robotics' approaches to cognitive sciences. On the other hand, cognitive robotics can be seen as an approach to achieve robots with the key characteristic of adaptive anticipatory interaction (IEEE RAS, 2017). In IEEE RAS (2017) cognitive robotics is defined as a science combining research coming from adaptive robotics, cognitive science and artificial intelligence, and often exploits models based on biological cognition. It is also emphasized that, as a form of embodied cognition, cognitive robotics exploits the robot's physical morphology, kinematics, and dynamics, as well as the environment in which it is operating. The definition in IEEE RAS (2017) is consistent with the direct trajectory shown in Figure 1.1.

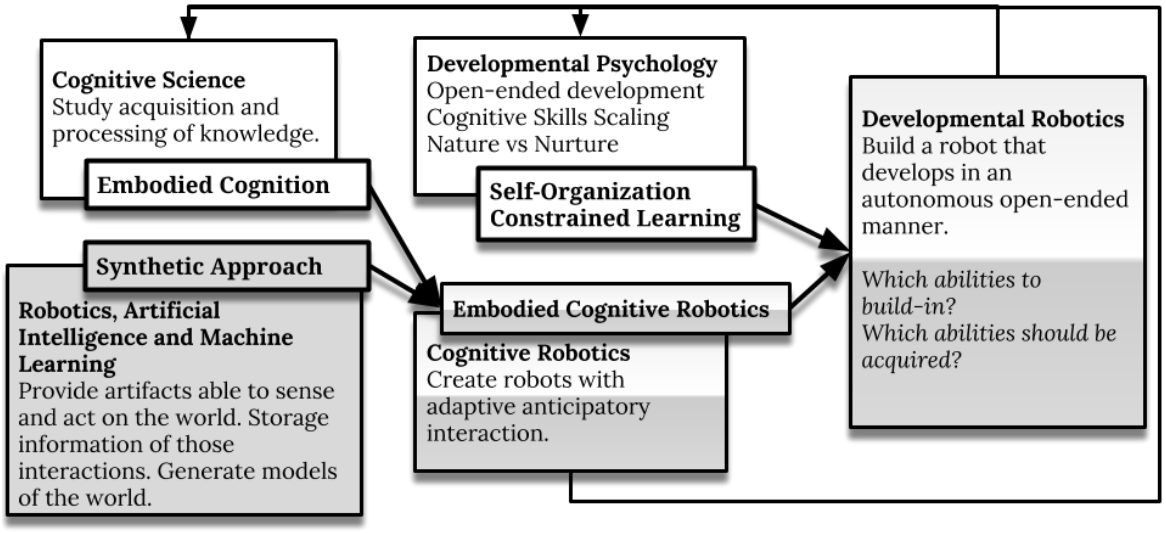

FIGURE 1.1: From cognitive sciences to developmental robotics.

So far, in the sense of studying cognition, for example using evolutionary approaches (Ferrell and Kemp, 1996), the scalability of the approaches aimed at generating agents with complex cognitive skills built over simpler cognitive skills is null. Until a few years ago, roboticist had predominately addressed basic cognitive phenomena, like sensory-motor coordination, perception, and navigation, but none of those provide clues regarding how to scale the system to high complex cognitive systems as humans (Mirolli and Parisi, 2011).

Indicated in Figure 1.1, as a part of a new trend to use a synthetic approach to study cognition as an embodied phenomena, the use of robots as scientific tools has also spread in recent years (Asada et al., 2001, 2009, Mirolli and Parisi, 2011, Pfeifer and Scheier, 1999). Robots can be used as tools to understand reality analytically. The synthetic approach consists of building systems that reproduce observed phenomena and obtain candidates to explain that phenomenon (Mirolli and Parisi, 2011). A critical question that robots may help to answer is *How natural cognition scales from basic cognitive skills, to complex cognitive skills?* In the field of developmental psychology, Piaget's provided evidence to show that a mature adult's cognitive skills are the result of evolving physical and mental skills that build during infancy, usually occurring in a clear sequence of defined stages (Ferrell and Kemp, 1996). This developmental process is not random; the existence of structured developmental stages suggests that development is the result of physical changes and a constant interaction with the environment.

Developmental psychology is a pillar of a relatively new approach in robotics: developmental robotics. The other pillar of this approach is cognitive robotics and its strong linkage with embodied cognition. As indicated in Figure 1.1, developmental psychology and cognitive robotics are merged to take advantage of the embodied nature of development to exploit embodied cognitive robotics. Thus, allowing to integrate interesting concepts into the robotics framework and enriching the synthetic approach. Developmental robotics aims at understanding and modeling the role of developmental processes in the emergence of complex behaviors, including social ones (Asada et al., 2009, Cangelosi et al., 2010). Developmental robotics has emerged as an interesting solution to achieve the scalability of intelligent artificial systems. The developmental approach has been shown to be suitable to scale basic cognitive phenomena to emulate the complex cognitive processes occurring for example within the human mind (Ferrell and Kemp, 1996).

One of the advantages for researchers working on developmental robotics is the plenty of available psychological studies, which provides an important amount of cues for the improvement of intelligent artificial agents. The growth of developmental robotics would not be possible if the developmental psychologists would not have been studying human infants for more than a century (Demiris and Meltzoff, 2008). Moreover, the improvement of experimental techniques in psychological studies has been a door of opportunities to determine what capabilities infants are born with and how these capabilities develop over time and with experience (Demiris and Meltzoff, 2008, Gopnik et al., 2001).

Studies about the human mind are the best examples for developmental robotics, but it is important to acknowledge that humans develop in an autonomous open-ended manner through lifelong learning (Oudeyer et al., 2007). Despite some advances, no robot has the capacity of developing in an autonomous open-ended manner. Hence, if a roboticist wants to build a robot that emulates at least some human capabilities, developmental robotics has a niche of action where the aim is to build robots with the capacity of developing in an autonomous open-ended manner to achieve characteristic of adaptive anticipatory interaction. The contributions of this work lay on the niche defined by the concept of developmental robotics.

Looking at developmental robotics as a branch forked from cognitive robotics. It is aimed at generating complex social robots with human-like cognitive and physical skills imitating the natural developmental mechanisms produced by evolution (Asada et al., 2009). Moreover, it also has the objective of understanding how human beings develop from harmless infants to functional adults capable of solving complex cognitive tasks. Developmental roboticists attempt to achieve these objectives building intelligent artificial agents endowed with physical and cognitive mechanisms.

Developmental roboticists should acknowledge that developmental psychology theories are based on the hypothesis that adult-level skills and competencies are not innate, but they emerge through life (Ferrell and Kemp, 1996). Thus, these roboticists must consider that not all the skills and competencies of a robot must be innate, but at least some of them emerge from the cumulative learning and interaction with its environment. However, in a developmental approach, as mentioned in Lungarella et al. (2003), an autonomous robot should be endowed with an initial set of values and drives to build complex cognitive skills on, i.e., motivations or needs to act and interact with the environment in order to achieve

an adaptive behavior. Just as an infant, a robot must build its complex cognitive skills inventing, discovering, and constructing its cognitive structures during its early development through cognitive and developmental evolution intertwined with a constant evolution of mental structures and physical competencies (Ferrell and Kemp, 1996).

Just as developmental psychologists (Gopnik et al., 2001) do, roboticist working using the developmental approach face an important question, an open debate in natural sciences that confronts two important concepts: nature and nurture. In the analysis of any developing system, the initial conditions play a crucial role. In the case of infants, nature is defined by the investigation of the behaviors and predefined developmental trajectories that infants are capable of displaying after birth, which is not an easy task. In the case of robots, this task is easier, as it is reduced to describe the algorithms and initial knowledge that are preprogrammed into the robot before the experiments begin (Demiris and Meltzoff, 2008). Later, the problem of nurture in robotics could be solved through imitation mechanisms for example. Imitation, which is seen as a major avenue of learning in infants (and humans in general), has been proposed as a promising method for a compromise between nature and nurture (Demiris and Meltzoff, 2008, Gopnik et al., 2001).

Regarding the compromise between nature and nurture, Lungarella et al. (2003) formulates two important questions that should be answered before attempting to emulate cognitive phenomena using the developmental approach: *How much has to be predefined?* and *How much should be acquired?*. In general, when reproducing given phenomena, these questions may have a broad range of answers but on the manner in which we approach those questions will have an impact on the complexity of the solution to be implemented and on the scalability of the resultant cognitive system.

The Piaget view has been a cornerstone of the research dealing with theories of infants' development. However, there is another perspective that has played an important role in developmental theories: the Vygotskian view. From this view, language is not only a communication system, but it is also a cognitive tool. As mentioned in Lungarella et al. (2003) and Mirolli and Parisi (2011), Vygotskian's theories on psychology may be a promising approach to include in robotics to achieve important progress to scale up cognitive systems. Later discussed in this work, if language comes to the scenario of relevant elements to achieve the emulation of complex cognitive development, then social interaction and the acquisition of language become a key element for developmental studies. If language is necessary for the

emergence of complex cognitive behaviors, then *Should the robots of the future be born to talk?*.

1.2 Should the Robots of the Future be Born to Talk?

If we start asking the question *Should a robot communicate with other agents?* then the answer would be: *not necessarily*. There are plenty of tasks where a robot does not need to communicate with other agents, mainly in the industry, where the environment is structured, and the task for each robot can be accurately defined. Even in some tasks where other agents are present, communication may not be crucial. A good example would be the iRobot's Roomba robot: even though it may interact with other agents, in the sense that it modifies its behavior if one stands in its way, it does not necessarily need any communication. For this kind of robots, it is enough to acquire the right information from the environment and react accordingly.

Despite not being a mandatory requirement, complex communication skills will be required for robots in certain scenarios. As the tasks which robots are designed to become more complex, robot-robot or human-robot interaction will emerge as part of the tasks and communication will become necessary for interactions and collaborations to be successful. The communication between robots could be reduced to a structured and simplified system, where a group of robots is designed to be part of that communication system.

When interactions or collaborations of a robot occurs with a human, then the problem becomes considerably more complicated. As social machines, robots should be built to interact efficiently with humans; one could propose a basic communication system between robots and humans. In that case, the problem would be reduced to the right education of the human-user. However, if robots are genuinely intended to become complex social machines that could interact with humans in any scenario, even with humans without any preliminary instruction, then they should be endowed with human-like interaction mechanisms. In this sense, the complexity of the task has placed human-like communication systems as a relevant topic of research for roboticists.

How to deal with the Human-Robot Interaction (HRI) has been widely studied. A survey presenting the history and main advances of HRI was presented in Goodrich and Schultz

(2007). Therein, the authors emphasized that different levels of socialization are required across different tasks. In general, the interaction mechanisms will be defined according to the degree of autonomy, information sharing, and evaluation required by the task at hand. In general, one also has to distinguish two different branches of interactions: the interactions with a computer which is somewhat hard to define as embodied and interactions with a robot, which represents an embodied interaction.

With the emergence of embodiment as a cornerstone of cognitive robotics, and as the interaction between embodied agents represents a social cognitive phenomenon, HRI has been studied as an embodied phenomena. For instance, Mutlu et al. (2016) provides some insights on the relevance of embodiment to HRI. Therein, dialog-based interactions are constantly analyzed alongside other communication modalities, e.g., gaze and pointing. For example. dialogue-based could be used for supervisory control through interactions to solve tasks such as navigation, collaborative exploration, and multi-robot teleoperation. Dialog could be the channel to share information and control at critical points in the collaborative task. Regarding dialog-based interactions, it is constantly assumed that robots are endowed with speech synthesizing and recognition capabilities.

Mutlu et al. (2016) acknowledges the fact that dialogue-based interaction mechanisms have the drawback of being rule-based systems. Those mechanisms have difficulty managing the many uncertainties that stem from noisy speech recognition or linguistic ambiguities. Those errors could be compensated by educating the user on how to speak correctly to the system. However, an ideal social robot must be able to infer user intentions and orders under uncertainty robustly.

Finally, Mutlu et al. (2016) emphasized three objects of research that will contribute to improving the performance of robots in HRI tasks. The first object is to build a better understanding of human cognition in HRI. The second one is to build models for simulating human cognition in robots. The last one is to build models that support human-robot joint activity, including dialog-based and other models that enable robots to reason about the physical and cognitive properties of the environment and the actions of their human counterparts. Furthermore, robots must integrate this knowledge to plan actions toward achieving communicative or collaborative goals.

An example to understand the failure of humans to endow machines with human-like language are chatbots. They use text or a synthesized voice. Chatbots are virtual agents usually endowed with some artificial intelligence that can conduct real-wise conversations (Hill et al., 2015). Despite the significant quantity of information and examples of real human conversations available to them, they usually fail and lack of shared-intentionality (Thompson et al., 2013). However, Google Duplex may be a good example that this failure and this lack of shared-intentionality are coming to an end. We leave that discussion out of the scope of this work as the information on this new technology is somewhat scarce.

In general, speech is one of the most studied communication systems because it allows human-spoken language. However, as mentioned in Kuhl (2004), 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 surrounding speech might be solved if we could understand all the mechanisms underlying early speech acquisition in children. In fact, the weaknesses of current speech synthesis and speech recognition systems may be attributed to the fact that these systems are not designed acknowledging the human embodied and neural processes of speech production and perception (Kröger et al., 2009). These processes may also include the developmental process in which speech and language are acquired.

Cognitive roboticists and artificial intelligence scientist have addressed the study of speech recurrently. However, the lack of an accurate understanding of the mechanisms used by infants to learn the speech 'code', from babbling at 6 months of age to full sentences by the age of 3 years, may be one of the most critical obstacle to prevent the achievement of an advanced artificial equivalent to natural language based on speech, as mentioned in Kuhl (2004). Fundamental to the explanation of how humans communicate is an understanding of the mental processes that support language comprehension and production (Tooley and Bock, 2014), abilities that may be developed during early infancy (Kuhl, 2004).

Building realistic speech-based communication systems requires an accurate understanding of the mechanisms used by infants to learn the speech 'code' (Kuhl, 2004). Infants show preparedness to master speech and acquire language: from the onset of canonical babbling at 6 months of age, infants achieve to produce full sentences by the age of 3 years. Lack of knowledge underlying this developmental process has been a principal obstacle to achieve advanced artificial equivalents to natural language (Kuhl, 2004).

As defended in Iverson (2010), the embodiment paradigm can be extended to language acquisition. In other words, early-vocal development as a prerequisite for spoken language can be studied as a result of embodiment, self-organization and emergence mechanisms produced by human evolution. The findings can be later used to endow robots with similar mechanisms to improve their communication skills, and therefore its capabilities to perform in social and collaborative tasks.

In general, many studies have demonstrated 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 (Kuhl, 2004). If one of the aims of roboticists is to build robots endowed with human-like speech and language capabilities, then we ask the question *Should the robots of the future be born to talk?* Some of them should probably be.

1.3 Motivation

In the previous sections, we have briefly discussed that studies in developmental sciences have strongly suggested that infants are born to talk. In the next chapters, we talk in more detail about the scientific results from biological sciences. Infants are born to talk in the sense of having the necessary simple cognitive skills to build more complex cognitive mechanisms required for language. We also elaborate on the idea that robots may take advantage of the findings in developmental sciences to become human-like machines. The road to achieving this may be hard, and the progress could be even slower, but the failures in achieving high performance in dialogue-based communication between artificial agents and humans are an indicator that the effort is worth.

During the scientific quest of building a robot that learns a language as an infant does, there is also a contribution to understanding the human mind and the cognitive development of infants. In this sense, artificial intelligence technologies and robotics have acquired a significant relevance in the study of the human mind. The emergence of advanced machine learning techniques, fast robot-prototyping, and complex simulators has fueled the appeal of artificial agents for studying the mechanisms of cognitive development with the parallel interest on building more intelligent robots and artificial systems. When developing complex robots, some of them will need to communicate with humans: robots helping in rehabilitation,

health care of the elderly, rescue missions, collaborative tasks in the industry, social robots for touristic assistance, a digital entity that schedules a medical appointment for us, and so forth.

How could robots communicate with humans? Communication can be carried through messages encoded verbally and non-verbally (Wagner et al., 2014). However, for developmental robotics, at this point, it is of particular interest to study the verbal channel: the one based on speech. There are two main reasons for this interest according to Wagner et al. (2014). First, the availability of research on human communication that has been mainly focused on speech. Secondly, the fact that the emergence of perception and motor speech skills in infants have been suggested to be a product of a developmental process. It is during early infancy that humans become aware of the communication value of speech. In this sense, within the studies of language emergence and its links to cognitive development, early works as Liberman and Mattingly (1985) have established perception-motor links according to evidence in the neurophysiological and behavioral levels. The perceptuo-motor link during development is of particular interest for this thesis.

The advantages of building a robot that could learn a language as a human does in order to later communicate with users are one important motivation for this thesis. That robot will be the perfect candidate to satisfy human needs, present and future, in different areas: medical surgery, nursery, rehabilitation, pets, tourism, manufacturing, unfortunately military, among others. Moreover, in building that robot from a developmental perspective, we are going to gain a deeper understanding of language emergence, including all the developmental subprocesses involved within the infants' embodiment in order to pronounce their first word. This thesis focuses on that specific subprocess of early language development: early vocal development in prelinguistic infants. We argue that in studying this developmental process, we are going to obtain relevant cues about all the phylogenetic and ontogenetic mechanisms that cooperate to transform a newborn human into an adult capable of communicating and solve complex cognitive tasks.

Developmental psychologists have found much evidence regarding infants and early language emergence. When infants are born, they have the necessary skills to learn any language. However, during the infants' early development their speech perception systems are specialized to their native language. This 'perception closure' decreases drastically infants' capacity to learn other languages. In studying early-vocal development, we expect to contribute in

the quest to find mechanisms to avoid the perception closure that impede us to perceive all the nuances of any foreign human language (Kuhl, 2004).

Moreover, the investigation of motor and perceptual theories on language that using a developmental perspective combined with a synthetic approach using intelligent artificial systems can be later extended to a broader range of natural and artificial cognitive systems. On the reverse way, studies of natural and artificial cognitive systems may also contribute with new insights about early vocal development, as the proposed extension of intrinsic motivated sensorimotor exploration architectures to study the emergence of language stages in Moulin-Frier et al. (2013). In general, it is of our interest to contribute to the investigation of developmental phenomena using artificial agents. In the long term, this investigation would contribute to the generalized tasks emphasized by Mutlu et al. (2016): build a better understanding of human cognition; build models for simulating human cognition in robots, gaining cognitive capabilities through imitation and interaction with the physical environment, hopefully in an open-ended manner as mentioned by Oudeyer et al. (2007).

Most of the works aimed at studying artificial speech-based communication systems are instead focused on the natural language understanding problem. The lack of focus on early vocal development and, in general, on prelinguistic communication is not surprising. As mentioned in Gros-Louis et al. (2006), just a couple of decades ago it was still assumed that vocal development was the result of maturational programs, which were independent of environmental influence. Therefore, as developmental psychologists were not primarily interested in vocal development during the prelinguistic stage until recent years, developmental roboticists did not have sufficient evidence to implement into artificial systems. As a consequence, works on these aspects are sparse. Despite the difficulties, the scientific literature has been enriched by a series of studies using artificial early vocal development as a mechanism to understand language emergence from an embodied developmental perspective, for example Forestier and Oudeyer (2017), Howard and Messum (2011), Moulin-Frier and Oudeyer (2013b) and Najnin and Banerjee (2017). This thesis aims to enrich the evidence found in those studies.

Embodiment has been argued as one of the central concepts to be considered through this work; embodiment imposes constraints at different levels. Motor, perceptual, social, and learning ability constraints, and their maturation during infant development play a key role in the emergence of intelligent behaviors, including spoken communication (Kuhl, 2004). In this sense, the study of the role of those constraints is of particular interest for this work,

and thus it is enlisted in the motivations of this thesis. The role of motor constraints during early vocal development is especially considered and in general their role during perceptuo-motor (sensorimotor) learning. In general, it is of our interest to study somesthetic senses, as tactile perception, proprioception, and nociception (perception of pain). Other authors have also urge further study of these perceptual modalities that may foster the emergence of intelligence behaviors in living beings during development (Navarro-Guerrero et al., 2017b).

About studying infants' early-language development and their openness to learn, Patricia Kuhl (2010) said in a conference:

> *"Just as the poets and writers described, we're going to be able to see, I think, that wondrous openness, utter and complete openness, of the mind of a child. In investigating the child's brain, we're going to uncover deep truths about what it means to be human, and in the process, we may be able to help keep our own minds open to learning for our entire lives."*

This quote is the perfect synthesis of one of the most important motivations for this work. In general, it represents a good reason for computer scientist and roboticist to continue the expansion of an artificial cognition branch that allows a substantial contribution to the study of human development aimed at unveiling the deepest secrets of infant's brain and development. As a roboticist, we are necessary because theories have almost no impact if they cannot be adequately tested. In this sense, integrated implementations of speech processing in robots, and artificial agents in general, provide valuable environments for the formal and empirical evaluation of cognitive models and theories (Wagner et al., 2014).

Finally, from Hall, Hulit et al. (2011) and Gopnik et al. (2001) there are important conclusions of the relevance and motivations to study the mind of children and the early development of language. These resources offer the picture of language in early stages as a seed, if a child is the pot were that seed grows, interaction must be the water through speech or signs, and other affective interactions. Hall says:

> *"Like a growing plant, language can develop into a twig or a tree, depending upon the nourishment it receives."*

Regular developing infants cannot decide whether or not acquire speech and language. Asking the questions mentioned by Hall is mandatory:

- What language does for the child and us?
- How does language affect our lives when we have it?
- How does it affect our lives when we do not have it?
- How it does or doesn't develop in the child?
- What we can and should do about its development in any case?

As evidence regarding the nature of the phenomena occurring during early human development, articles as The Economist (2018) will continue emerging placing important questions and answer in the hands of people and their governments. *Does growing up poor harm brain development?* This article argues that growing up in a low-income family does affect child development, or at least does not foster strong language and memory skills. More affluent children usually perform better in school and are less likely to end up in jail.

Language allows time travel, mental time travel (Hulit et al., 2011). Language allows us to connect with our past, our present, and our future as individuals, but also as a civilization. It allows us to imagine what others might be thinking, and hopefully, it allows to connect with ourselves.

As Hall, we would like to contribute to work in the scientific and technological basements that will allow us one day *"to get for each child a bridge as broad as the Brooklyn Bridge, or better yet, the Golden Gate"*. A bridge that will allow them time travel, a connection with themselves and with others, it will, at last, contribute positively to the quality of human societies. To build that bridge, we must recognize the processes through language development, the time windows and the things to do to foster the success of each process.

Finally, summarizing the motivations to develop the project contained in this thesis, we want to contribute to the study of processes occurring during the early development of children. We look especially at the processes involving early vocal development, as it represents a sensorimotor learning process, we adopt as a general object of study sensorimotor exploration observed in many behaviors during infancy. Once we have established the embodied nature of sensorimotor learning, given the relevance of constraints to understand implications of embodiment, we believe that it is essential to study how available studies on artificial early vocal development might be affected by the active consideration of motor constraints. Finally,

if interactions are necessary for the acquisition of language and speech, studying feasible mechanisms of interaction during prelinguistic artificial vocal development is also of particular interest for this thesis.

1.4 Objectives

We have established the objects of research that will be carried through this work, and also the reasons we have to argue the relevance of those investigations. This research is a study of the role of constraints and social interaction during exploratory sensorimotor behaviors, especially those related to prelinguistic vocal development.

Our general objective is to contribute to answering some of the paradigms regarding early prelinguistic development. We argue that, to some extent, it would contribute to answering some parts of prelinguistic developmental processes required for the emergence of spoken language in children. Our objective is to use available methodologies in developmental robotics and coherent with developmental psychology studies to understand the developmental progression which allows the emergence of complex behaviors in developmental living beings and machines. In this sense, this work is aimed mainly to study the motor and perceptual systems involved in speech production and perception.

When we find a gap in the available methodology, then we must provide or at least contribute to the generation of a clear methodology. Following a clear methodology will make our results easily reproducible by interested researchers in order to foster new contributions and cooperation. Clear methodologies also foster the debate and discussions that are required for a branch of science to grow. Thompson et al. (2013) mentioned some critical facts that we will consider when evaluating our contributions to the early vocal development from the psychological point of view. They recommend to make as few assumptions as possible, as it guarantees a degree of generality and leaves more details to science instead of philosophy, in current stage of developmental robotics it is hard to make few assumptions, but as an objective we attempt to justify and clarify the implications of assumption made through the experiments within this work.

In general, to be consistent with previous sections, our objective is to study speech emergence according to behavioral and physiological evidence using a developmental approach.

Roboticists using the developmental approach to investigate the early vocal emergence and vocal development should focus on the role of embodiment and social interactions in the course of development must also be investigated Asada (2016). In this sense, the general objective of this work is to contribute to developing the basement of a discipline that will allow building complex social robots. Those robots, through interactions with their environments, must incrementally build new and more powerful mental and behavioral structures through developmental processes. So far we have argued that the success on the quest of building a complex social robot with human-like cognitive skills should be approached in an interdisciplinary way, but it is more important to find a principle shared by different disciplines and its contribution to the gaining of new insights.

The specific objectives, approached through experimentation along this work, are enlisted below.

- Our first objective is to collect and understand the series of studies that led to the findings in Moulin-Frier et al. (2013). Therein, early vocal development was studied as a result of exploration behaviors, in which an agent endowed with an artificial ear-vocal tract can generate a map from articulatory gestures to auditory outcomes. Exploration is not just random, but the agent attempts to reach auditory states that maximize the learning progress, based on intrinsic motivations inspired by behaviors observed in children.

- The second objective is to reproduce the experiments performed in Moulin-Frier et al. (2013).

- Based on the general objectives and the work by Kuhl (2004), we propose our third objective. It is to study the role that constraints imposed by embodiment may have if the information provided by those constraints is actively included in the mental processing path. We especially propose to study the role of motor constraints as, based on biological evidence, we argue in the coming chapters that motor constraints are something children learn during early development.

- Language nor speech are elements that could be learned in isolation as remarked in previous sections. It is necessary to contemplate social mechanisms when they are studied. Imitation has been mentioned as one important mechanism for children and

robots to incrementally acquire knowledge from other humans or robots. Inspired by works similar to Howard and Messum (2011), our fourth objective is to study, based on biological evidence, how to integrate imitation mechanisms to sensorimotor exploration applied to the vocal development and study what the role of imitation scenarios is through the course of artificial development.

We also have the next minor objectives enlisted below.

- An important objective is that sensorimotor exploration algorithms developed through this work must be presented in such a way they can easily be applied to any sensorimotor system, not only to vocal development experiments.
- Finally, as we believe that science should be openly available to enrich scientific results and collaboration, the codes developed through this work must be open to any interested researcher willing to work in similar applications.

1.5 Contributions

Based on the objectives introduced in the previous section, this work provides new results to contribute to the study of early speech development using machines. Among the most important concepts to consider when applying the developmental approach, this work emphasizes the embodiment paradigm and sensorimotor exploratory behaviors. The contribution of this work is extending the study of early prelinguistic vocal 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 and imitation episodes during sensorimotor exploration for any sensorimotor system subjected to constraints.

We divide our scientific contribution into three main parts mentioned in the following. Along with our scientific contributions, we mention scientific publications that were accepted in peer-reviewed conferences and journals. The first part of our scientific contribution is related

to the first, second and third listed objectives. Our work on these objectives is contained in the following publications:

- J.M. Acevedo-Valle, C. Angulo and C. Moulin-Frier (2017) Autonomous Discovery of Motor Constraints in an Intrinsically-Motivated Vocal Learner. IEEE Transactions on Cognitive and Developmental Systems. 2017. DOI 10.1109/TCDS.2017.2699578.
- J.M. Acevedo-Valle, C. Angulo, K. Trejo and C. Moulin-Frier (2016) The Role of Somatosensory Models in Vocal Autonomous Exploration. Innovation Match MX 2015-2016, Guadalajara, Mexico. Revista Internacional de Investigación e Innovación Tecnológica. ISSN: 2007-9753. [riiit.com.mx/apps/site/files/art._4_immx_v1.pdf]
- J.M. Acevedo-Valle, C. Angulo, N. Agell and C. Moulin-Frier (2015) Proprioceptive Feedback and Intrinsic Motivations in Early-Vocal Development. 18th International Conference of the Catalan Association of Artificial Intelligence (CCIA 2015), pp. 9-18, Valencia, Spain. IOS Press. [DOI 10.3233/978-1-61499-578-4-9]

In these publications we provide an extension of the studies in Moulin-Frier et al. (2013). Inspired on somesthetic senses, these new studies provide an architecture in which constraint awareness can be successfully integrated into intrinsically motivated sensorimotor exploration architectures. Taking into account constraints during learning of sensorimotor regularities is suggested to be major contribution to the performance of those exploration architectures according to the results provided in this thesis.

The second part of our scientific work is related to the fourth objective mentioned in the previous section. This part of our work was a partial collaboration with the Adaptive System Group of the Humboldt-University of Berlin. The contribution made through our work in this part of the objectives is contained in the following publications:

- J.M. Acevedo-Valle, V. V. Hafner and C. Angulo (2018) Social reinforcement in artificial prelinguistic development: A study using intrinsically motivated exploration architectures [Submitted].
- J.M. Acevedo-Valle, C. Angulo and Verena V. Hafner (2017). Social Reinforcement in Intrinsically Motivated Sensorimotor Exploration for Embodied Agents with Constraints Awareness. 2017: ICDL-EpiRob, Lisbon, Portugal.

- (Poster) J.M. Acevedo-Valle, c. Ruiz-Camps, Verena V. Hafner and C. Angulo (2017). Deep Neural Networks in Social Reinforced Sensorimotor Exploration. 2nd Workshop on Language Learning. 2017: ICDL-EpiRob, Lisbon, Portugal.

In these publications, we provide an extension of the work developed during the first part of this thesis. Therein, we provided a feasible architecture to integrate, apart from constraints, social interactions as a part of early sensorimotor exploration. It is not the first time social interactions are considered as an element for artificial mental development. However, we argue that it is the first time that a sensorimotor exploration architecture, especially one aimed at study vocal development, considers that three critical elements occur in parallel through the course of early development: intrinsically motivated exploration, constraint awareness, and social reinforcement. In this sense, we do not assume that each of the different modalities develops one after the other, but we consider that they evolve in parallel. One modality reaching a milestone might produce abrupt developmental changes in the others as discussed later. Hence, we argue that observed development stages might also be the product of those abrupt changes.

Finally, the third contribution of this work is the product of a need that emerged during the development of this work. In order to learn sensorimotor maps during sensorimotor exploration, we found relevant to rethink algorithms for incremental learning of Gaussian Mixture Models (GMM). This part of our contribution was published in:

- J.M. Acevedo-Valle, K. Trejo and C. Angulo (2017). Multivariate Regression with Incremental Learning of Gaussian Mixture Models. 2017: 20th International Conference of the Catalan Association of Artificial Intelligence (CCIA 2017). Terres de l'Ebre, Spain.
- (Abstract) J.M. Acevedo-Valle, C. Angulo and K. Trejo (2017) Incremental Learning of Gaussian Mixture Models for Multivariate Systems. Innovation Match MX 2016-2017, México, Mexico.

In these works and for the first time, we combined an incremental learning approach for GMM based on the geometry properties of Gaussian with Gaussian Mixture Regression (GMR) to solve the inference and prediction problem of static input-output maps. Apart from being

necessary for our sensorimotor exploration approach, we argue that this contribution could be useful for a broader range of applications.

Our final contribution is a couple of open source Python packages. First, a simulated vocal tract that facilitates the study of early vocal development using Python libraries, where many tools for developmental robotics and machine learning are available. Secondly, a library with our proposed implementation for the incremental learning of GMMs and GMR.

1.6 Short Academical Stays Abroad

I made two international academical stays during the doctoral studies; they are briefly described below.

Eidgenössische Technische Hochschule Zürich (ETHZ) It took place from November the 1st, 2015 until January the 31st, 2016. I visited the Autonomous and Dexterous Robotics Laboratory and worked under the supervision of Prof. Dr. Jonas Buchli and Dr. Diego Pardo. The stay was co-sponsored by the National Centre of Competence in Research Robotics of Switzerland. As this stage was done during the consolidation period of my doctoral research, the work was not included in this thesis. However, it is important to mention that the project had as objective to design robust rime variant controllers for the stabilization of optimal trajectories in underactuated systems.

Humboldt-Universität zu Berlin (HU Berlin) It took place from the 1st/March/2017 the 30th/June/2017. I visited the Adaptive Systems Lab of the Informatics Institute under the supervision of Prof. Dr. Verena Hafner. This stay was of crucial relevance for the elaboration of this thesis. The main objectives of this stay were studying the role of social reinforcement, somatosensory and proprioceptive systems during the emergence of sensorimotor explorations behaviors.

From this collaboration, as we mentioned beforehand, two papers were written. One paper was presented in the ICDL/EpiRob Conference (Acevedo-Valle et al., 2017a), and the other has been submitted to the IEEE Transactions on Cognitive and Developmental Systems (Acevedo-Valle et al., 2018).

Apart from the two written papers, an important step in the implementation of necessary software for the experimentation within this thesis was completed, especially that related to the implementation of the *divapy* package, explained in Appendix A.

Following the main objectives of the stay, I was exploring new techniques based on deep learning to implement proprioceptive and somatosensory systems into our vocal tract. The basic idea was to use autoencoders to have a simple representation of touch information. Finally, so far our speech perception system only considers the trajectory of formant frequencies, which are a good account for vowel description, but not good for consonant perception. In order to make a more powerful architecture, we started exploring new speech features that allow us to perceive consonants. As we are studying the emergence of speech, it is important to have a speech perception system similar to that of the humans.

1.7 Thesis structure

Besides this introductory chapter, this thesis is structured into five more chapters. A brief description of each of the remaining chapters is provided below.

Chapter 2. State of the art. This chapter is aimed at fulfilling the first objective of this work. The reader will be introduced through a journey of two branches of knowledge converging to the results in Moulin-Frier et al. (2013), where our contributions start. On the one hand, we study the different steps from the artificial intelligence perspective and robotics in order to generate the proper architectures that later were applied to the study of vocal and language development using machines. On the other hand, we visit different theories an experimental results regarding speech production and perception, and the developmental processes that may be involved during the period in which a child learns to perceive and produce speech. Literature regarding motor constraints, somesthesis, and the role of imitation episodes will be considered in Chapters 4-5.

Chapter 3. Incremental Learning and the Regression Problem with Gaussian Mixture Models. This chapter is aimed at introducing our approach to the incremental learning of GMM and GMR. We also present samples examples that illustrate how our approach works. A simple sensorimotor system which includes constraints

is proposed in this chapter. The learning and regression mechanisms, along with the simple sensorimotor system example, are later used to test the cognitive architectures for sensorimotor exploration presented in Chapters 4-5.

Chapter 4. Motor Constraint Awareness in Sensorimotor Exploration. One of the two main contributions of this work, even though it is based on Acevedo-Valle et al. (2015, 2018), it provides further references regarding the role of somesthetic senses and constraint awareness during early sensorimotor development. It also presents new results, obtained with the most recent version of the software developed during this project.

Chapter 5. The Role of Imitation Episodes in Intrinsically Motivated Sensorimotor Exploration. This is the second main contribution of this work, it is based on Acevedo-Valle et al. (2017a) and on the submitted work Acevedo-Valle et al. (2018), it provides further references regarding the role of imitation episodes observed between mothers and children to early vocal development and sensorimotor development. It also makes a brief review of the relevance that imitation mechanisms may have to create more complex robots. It also presents the most recent results of this work, and therein the reader will find the best picture of early vocal development that we achieved to obtain in a simple vocalization scenario.

Chapter 6. Conclusions and Future Work. It is the final chapter of this thesis. Therein we summarize the discussion carried through the thesis regarding the obtained results. We also assess the results and findings with respect the objectives of the thesis. Finally, we provide further lines of research to continue with the study of early vocal development considering this work and similar works that have been carried in parallel by other research centers.

Chapter 2

State of the Art

> *"Seek it with your hands, don't think about it, feel it. Your hands are wiser than your head's ever gonna be."*
>
> — Steven Pressfield, *The Legend of Bagger Vance*

In the previous chapter, the reader was introduced to the basic ideas, objectives, and motivations of this research. Therein, sensorimotor exploration and prelinguistic vocal development were emphasized as central issues within the framework of developmental robotics. In this chapter, we discuss a series of relevant studies for the development of this project. These studies cover a broad range of topics, from cognitive and developmental robotics passing by embodiment, intrinsic motivations, and sensorimotor exploration. Moreover, researchers related to speech and spoken language are covered to a considerable extent as well, including psychological literature about vocal development, speech emergence, speech perception and production, and language.

As emphasized in the introductory chapter, when performing in unstructured situations, robots should be robust, flexible, social, as well as adapt to their environment's changes, just as the living beings do. Learning, locomotion, navigation, orientation, manipulation, imitation, and cooperation were emphasized as critical challenges to achieve complex robots. Moreover, it was mentioned that biological sciences might provide relevant knowledge to face those challenges (Pfeifer et al., 2007).

In the perspective of embodied cognition, agent's behavior is not only the result of a system control structure. The behavior of agents is affected by their ecological niche, morphology and material properties (Pfeifer and Scheier, 1999, Pfeifer et al., 2007). In the case of an infant, embodiment plays a crucial role in the bootstrapping of mental competence empowerment (Ferrell and Kemp, 1996).

The research works and developments mentioned through this chapter provide important clues on embodied cognition. Those clues should be considered in any attempt to build an artificial agent endowed with the mechanisms that allow language emergence in infants. Therefore, those studies are relevant to build artificial agents that attempt to mimic prelinguistic vocal development. In this dissertation, we stand with the perspective of Iverson (2010), about emphasizing that language should be viewed in the context of the body in which the developing language system is embedded. Therefore, language is considered an embodied mechanism of communication.

This chapter is organized as follows. First, in Section 2.1 the concept of embodiment is discussed from biological and artificial perspectives. Secondly, in Section 2.2 the concept and mechanisms for sensorimotor exploration from a developmental perspective are introduced. Section 2.3 is aimed at briefly introducing artificial architectures that have been proposed to mimic sensorimotor learning in infants. We focus on architectures that are related to those that will be introduced in the following chapters. Section 2.4 presents the sensorimotor exploration architecture from Moulin-Frier and Oudeyer (2013b).

The second part of this chapter focuses on speech and language. First, we briefly discuss speech from a biological perspective in Section 2.5. Next, Section 2.6 focuses on the relation of speech and development. Then, we briefly discuss artificial mechanisms for speech perception and production in Section 2.7. Finally, in Sections 2.8-2.9, we follow the evolution of prelinguistic vocal development studies with artificial agents until the studies performed in Moulin-Frier et al. (2013), which are the starting point of this thesis.

2.1 From Embodiment to Rhythmic Behaviors

As mentioned in the Introduction, Liberman and Mattingly (1985) established that given the evolutionary origin of cognition, it needed to be understood as embedded in a meaningful

ecological context and embodied in living perceiving-acting systems. The behavior of those perceiving-acting systems is the result of its environmental complexity, and the interactions, mainly through goal-oriented actions, with their ecological niche (situatedness), their morphology and material properties (embodiment), and other individuals (Pfeifer and Scheier, 1999, Pfeifer et al., 2007). It is emphasized in Ferrell and Kemp (1996) that embodiment is an interesting candidate to understand how the body, the environment, and the mind interact to drive development and simplify learning through constraints and biases. On the one hand, a well-defined set of constraints will assist learning by reducing the space of possibilities in both, inputs and outputs. On the other hand, an input bias will produce a particular input easier or more likely to be executed. The embodiment paradigm changed the way in which cognition is understood.

In recent years embodied cognitive science is returning to focus on agent-environment interaction and embodied sensorimotor mechanisms. Machine learning fits perfectly in the paradigm of embodiment, approaches including artificial neural networks, behavioral-based systems, artificial life and evolutionary computing being commonly used in the literature (Angulo et al., 2009). It is showed in Angulo et al. (2009) how perception can be used as a relevant feature for action planning, therein an architecture that provides an autonomous agent with an 'inner world' based on internal simulations of perception rather than an explicit representational model was proposed with positive results.

So far, robots have been mostly designed for particular tasks. Therefore, they are built to meet the needs to perform within particular environments to achieve specific behaviors. Roboticists have frequently appealed to biological systems as inspiration to build robots that achieve their goal (Ferrell and Kemp, 1996). Despite this appealing to biological inspiration, roboticists failed to reproduce or mimic even simple behaviors of living beings. However, seminal works by Pfeifer and Scheier (1999), Pfeifer et al. (2007) and Ferrell and Kemp (1996) helped to spread the concept of embodiment among roboticists, which changed in many ways the way in which intelligence is understood.

Embodiment plays an essential role in constructing a cognitive foundation. As the interface with the world, embodiment allows to compose and administer relevant queries of the environment efficiently. In living beings, our bodies are demonstrated to be optimized to build low-level knowledge through body-oriented activity (Ferrell and Kemp, 1996). Conceived as an embodied phenomena, embodiment inspired by Pfeifer and Scheier (1999) helps to

argue that findings obtained from the fields of biology and self-organization may strongly benefit the construction of robots. If robots, as embodied agents, are described as dynamical systems, then it is possible to extend the concepts of self-organization and emergence to them. Those concepts are applied at the induction of sensory stimulation level, movement generation level, exploitation of morphological and material properties level, and finally at the interaction between individual modules (Pfeifer et al., 2007).

Pfeifer et al. (2007) argued that despite the interesting implications of the idea, embodiment had not been sufficiently explored at that time. Consequently, robots were –and they are still– energetically inefficient and lack adaptability when confronted with unexpected situations. However, a change in the perspective in which robots are built has fostered the emergence of more intelligent robots. As an example of the significance acquired by embodied cognition, searching *'embodied cognition'* in *GoogleScholar*[1] and filtering the results from 1980 to 2007, the number of results is about 20K; repeating the search for *'embodied cognition robotics'* for the same period the results are almost 18K. On the other hand, searching for the period 2008-2018, for *'embodied cognition'* the number of results is about 28K and for *'embodied cognition robotics'* is about 17K. If we use these figures as an approximated indicator to the relevance that embodiment has gained in the study of cognition and robotics, then we observe that the quantity of works produced in the last 10 years is approximately the same to the quantity produced during the 27 years before.

From a developmental perspective, let us think about humans. Body development acts as a regulator of information complexity that can be acquired by an infant. Thus, affordable knowledge is considerably limited in early age. However, as the body changes along development, then the available information increases in complexity and the infant can deal with more complex knowledge. For that, the infant relays on the accumulated knowledge and skills acquired along the whole developmental process (Ferrell and Kemp, 1996). From the developmental perspective, to guarantee incremental acquisition of abilities on previous competencies, it is important to consider how the system's goals, morphology, environment, and cognitive abilities grow in complexity (Ferrell and Kemp, 1996). As mentioned by Lungarella et al. (2003), adaptivity of developing agents comes from their morphological plasticity, i.e., changes over time of sensory resolution, motor accuracy, the mass of muscles and limbs, and so on. Despite being restricted in many ways, infants are tailored to the idiosyncrasies of

[1] https://scholar.google.com/

their ecological niche, even to the point of displaying a rich set of adaptive biases toward social interaction.

How the embodiment paradigm can be applied in the design of robots is straightforward and well represented by the quote "understanding by building" from Pfeifer and Scheier (1999). According to Pfeifer et al. (2007), bio-inspired robotics should work out embodied principles of biological systems and transfer them to robot design. The most explicit example in this sense is bionics, which seeks to design technology by mimicking the salient features of biological structures. Providing some experiments, Pfeifer et al. (2007) showed that physical constraints shape the dynamics of the interaction of the embodied system with its environment. Embodiment is a central player in the emergence of information regularities, coupled sensorimotor activity and body morphology. Embodiment induces statistical regularities in sensory inputs and within the control architecture and therefore fosters internal information processing. In the studies of language emergence in infants, Kuhl (2004) showed that infants use statistical an probabilistic information to learn their native languages.

Gottlieb et al. (2013) defended that despite the constant exploitation of biological mechanisms in robotics, there is a poor understanding of how intelligent animals explore and obtain information. As mentioned by Pfeifer et al. (2007), the mechanisms for perception are poorly understood. In the field of language emergence in infants, Kuhl (2004) emphasized that the perceptual changes that occur in infants from their birth to their first year of life are essential for language acquisition, however, to a large extent, the mechanisms that produce those changes are still a mystery.

As mentioned by Iverson (2010) and Ejiri (1998), a beautiful picture of embodiment significance is when infants engage in rhythmic behaviors, e.g., body rocking, head banging, head rolling, hand banging (Sallustro and Atwell, 1978). When hand banging emerges, infants feel themselves moving, they see the movement of their arms, and they hear the resultant sound, all occurring in synchrony. A large number of studies suggests that infants are highly sensitive to this type of synchrony and that the presence of such redundant cues facilitates recognition of contingencies. For speech emergence in infants, this sensitiveness to synchrony indicates that when infants begin to babble, they are prepared to recognize the contingent auditory feedback from their sound productions. This feedback allows them to monitor and adjust the state of the vocal tract as their sound production varies (Iverson, 2010).

In conclusion, rhythmic behaviors are necessary to infants in order to create sensorimotor maps. They allow infants to learn the regularities between motor actions and perceptual states. Then, infants may be interested in those perceptuo-motor regularities and, somehow, they feel motivated to gain more knowledge on those regularities, and in fact more control over them. The latter description is a picture of sensorimotor exploration and intrinsic motivations, which are studied in the next section.

2.2 Sensorimotor Exploration and Intrinsic Motivations

Rhythmic behaviors are mechanisms for sensorimotor exploration. Thus, they are required by infants in order to create internal body representations and maintain them through life. However, rhythmic behaviors are not the only mechanism used during sensorimotor exploration. In the literature can be identified at least two other ones: goal-oriented exploration and imitative behaviors (Demiris and Meltzoff, 2008, Gopnik et al., 2001, Oudeyer et al., 2007). Through this chapter, we are mainly focused on reviewing rhythmic behaviors and goal-oriented exploration, whereas imitative behaviors are discussed in Chapter 5.

There exist different sensorimotor relations that an infant acquires during early development, e.g., saccading, gaze fixation, joint attention, hand-eye coordination, visually-guided reaching, and vocalization-auditory consequences. As indicated by Lungarella et al. (2003), during self-exploration and self-learning, spontaneous movement activity play an important role even though the activity lacks of a functional goal, but gives infants the possibility of learning to sense and predict the consequences of their own actions through self-exploration.

In general, it is observed that the self-experience of perceptuo-motor regularities to gain sensorimotor control knowledge is a fundamental building block for different developmental pathways. In other words, as established by Schillaci et al. (2016), sensorimotor control and learning are fundamental prerequisites for cognitive development in humans. In infants, it is not very clear how all the mechanisms for sensorimotor control learning work. On the one hand, infants borrow some goals from observing others (imitative behaviors). On the other hand, evidence strongly suggests that infants are able to generate random acts (e.g., rhythmic behaviors) and then want to refine those acts by themselves in goal-oriented exploration (Demiris and Meltzoff, 2008, Gopnik et al., 2001, Oudeyer et al., 2007).

Similar examples are presented in Demiris and Meltzoff (2008) and Lungarella et al. (2003) to describe the engagement of an infant with sensorimotor exploration. Citing to Demiris and Meltzoff (2008): imagine an infant watching its hand floating across its visual field after performing a random motor act, then imagine that the infant wants to gain control of this scene. This desire causes her to repeat it again and again until the infant has mastered it. Lungarella et al. (2003) differentiates between two different sources of sensory information, one originated from outside the body (called exteroception, e.g., vision, audition or touch), and the second coming from inside the body (e.g., proprioception). In Chapter 4, the latter modality is studied with more detail.

Infants seem to be born with this innate willingness to master perceptual consequences through sensorimotor experiments (Demiris and Meltzoff, 2008, Gopnik et al., 2001). This willingness, according to Oudeyer et al. (2007), suggests the likely existence of a kind of intrinsic motivation system which provides internal rewards during these play experiences. Ferrell and Kemp (1996) remarked that a system engaged in sensorimotor exploration should build input-output representations which later may be used to avoid overwhelming and confusing detail (redundancy) or bored by unchallenging simplicity (motivation for the unknown). In this sense, developmental studies suggest that infants have an innate tendency to autonomously generate goals that foster development. In fact, unreachable, unreasonable and overly simplistic goals are rejected by infants through boredom and frustration. Ferrell and Kemp (1996) also explained that failures to execute successfully a task motivates infants to spend more time and cognitive resources trying to achieve that goal, that modifies their models of reality and improves their skills in order to master the task. This effort fosters the emergence of new tools and resources that later may be used to compose more complex behaviors.

From Demiris and Meltzoff (2008) and Baranes and Oudeyer (2013), it is observed that in the case of robots, forward models and inverse models to master sensorimotor knowledge can be acquired through exploration processes inspired in those observed in infants. In executing a series of arbitrary motor commands, or goal-oriented experiments a robot can associate its motor commands with its sensory consequences, e.g., visual, proprioceptive, touch, auditory, and so forth. There are some important problems of a robotic system learning to coordinate the amount of sensorimotor regularities as a child does. These problems are associated to the many degrees of freedom of a potentially redundant non-linear physical system (Lungarella

et al., 2003). However, the imitation mechanisms and self-exploration of the sensorimotor as described above may be the elements needed to achieve such a robotic system.

From a developmental embodied perspective, acquiring complex motor skills may benefit from the introduction of initial sensor, motor and neural constraints, which over time are gradually released. Intrinsic motivations, interaction with the environment and social interactions may play a role regulating sensorimotor learning as well (Lungarella et al., 2003).

In order to develop in an open-ended manner, robots should certainly be equipped with capacities for autonomous and active development, and in particular with intrinsic motivation systems (Oudeyer et al., 2007). Regarding the trade-off between nature and nurture, from Oudeyer et al. (2007) and Meltzoff et al. (2013) one can borrow some conclusions. It is important to distinguish three mechanisms in which infants acquire sensorimotor knowledge during early development as mentioned at the beginning of this section: motor babbling, goal-directed babbling and imitation learning. Therefore, the success of the approach to be presented in this work will depend on our capacity to integrate those learning capabilities in an artificial developmental agent. First, we will consider motor babbling as an element to initialize internal models of the agents. Next, we will consider goal-babbling as a way to enrich and refine sensorimotor control knowledge. Finally, imitation learning is hypothesized to be responsible of refining sensorimotor knowledge and provide opportunities to obtain new knowledge.

Among the vast number of active learning architectures, this work considers the exploration architectures proposed in Baranes and Oudeyer (2013) and Moulin-Frier and Oudeyer (2013b). This architecture reproduces the formalism of intrinsic motivation inspired by psychological literature as proposed previously in Oudeyer et al. (2007) and also in Gottlieb et al. (2013). From the concepts and implications of sensorimotor exploration requirements, an intrinsic motivation system must mediate learning, promote parameter exploration, drive action selection and regulate social interactions (Lungarella et al., 2003).

2.3 Sensorimotor Exploration Architectures

As depicted in the previous section, there are a series of exploratory behaviors that emerge during infancy that work as processes to learn sensorimotor regularities. Evidence suggests

that the mind and brain development are strongly intertwined with these sensorimotor exploratory behaviors, where internal body representations are formed and maintained. Those representations are used to master sensorimotor control, which is considered by developmental psychologists as a fundamental prerequisite to more complex cognitive and social capabilities (Schillaci et al., 2016).

A complete compilation of different relevant approaches is provided in Schillaci et al. (2016) to implement sensorimotor exploration architectures for robots, along with their equivalent in natural sciences. In general terms, from what Schillaci et al. (2016) compiled and the works we have reviewed through this thesis, that random exploration along with the selection of predefined actions (e.g., predefined motion primitives) is a common approach selected by roboticists. However, it is possible to observe that intrinsically motivated architectures are becoming more popular when addressing the sensorimotor exploration problem, some examples are Baranes and Oudeyer (2013), Moulin-Frier and Oudeyer (2013b), Oudeyer et al. (2007), Pape et al. (2012), Ribes et al. (2016), Saegusa et al. (2009), Shaw et al. (2015) and Schmerling et al. (2015). Another important element, mentioned before, is the natural relevance of goal-directed exploration observed in infants. As mentioned in Schillaci et al. (2016), the introduction of the goal-directed nature of exploration changed the way in which the problem of learning sensorimotor maps was addressed. This change in the paradigm of artificial sensorimotor exploration was promoted by Rolf et al. (2010).

In Rolf et al. (2010), an approach for inverse kinematics learning in redundant systems without prior or expert-knowledge was presented. Inspired by the fact that infants likely babble goals instead of motor commands, the authors demonstrated that goal babbling could be advantageous in learning in the early stages of development, as observed in developmental theories. From Rolf (2013), we emphasize the key idea of *learning by doing*; thus goal babbling can enhance learning control. For instance, robots could learn to reach by trying to reach as infants do.

In Oudeyer et al. (2007), an experiment was conducted with a robot endowed with curiosity-driven learning mechanisms. It was capable of self-organizing its own learning experiences into a sequence of behavioral and cognitive stages. Through these stages, it spontaneously acquired a number of affordances and skills of increasing complexity. As mentioned in many psychological works (Ejiri, 1998, Kuhl, 2004, Oller and Eilers, 1988), strong regularities are observed in the structure of the vocal development process independently of inter-individual

differences. In Oudeyer et al. (2007) active learning architectures based on intrinsic motivations were proposed as mechanisms that mimic the exploration behaviors observed during sensorimotor exploration in biological agents. Among the vast number of active learning architectures, this work considers the exploration architectures studied in Baranes and Oudeyer (2013) and Moulin-Frier and Oudeyer (2013b). These architectures reproduce the formalism of intrinsic motivations combined with goal babbling inspired by psychological literature as described in Oudeyer et al. (2007) and Gottlieb et al. (2013).

Using the goal babbling proposed by Rolf (2013), it is presented in Baranes and Oudeyer (2013) a self-adaptive goal generation architecture, or intrinsically motivated exploration mechanism, to actively learn sensorimotor maps of inverse models in high-dimensional redundant robots. In this architecture, based on a measure of competence progress, the robot actively samples novel parameterized tasks in the task space. The results using a robotic arm, a quadruped robot, and another example suggested that exploration in the task space can be much faster than exploration in the motor space for learning inverse models in redundant robots. Developmental trajectories are generated driving the robot to progressively focus on tasks of increasing complexity selecting goals maximizing competence progress according to a model of interest.

Similar architectures than those from Baranes and Oudeyer (2013) are used in Moulin-Frier and Oudeyer (2013b) and Moulin-Frier and Oudeyer (2014), but offering a detailed comparison between different exploration schemes. Moulin-Frier and Oudeyer (2013b) included some experiments with a simulated vocal-tract, attempting to learn the articulatory configurations-auditory outputs relation of the system which is of deep interest for this work. Moulin-Frier and colleagues compared four schemes of explorations: first, they consider random exploration over the motor space and the sensor space, next, they consider intrinsically motivated learning over both spaces as well. From the results in Baranes and Oudeyer (2013) and Moulin-Frier and Oudeyer (2013b, 2014), it is notorious that the architectures considering intrinsically motivated learning over the goal space (goal babbling) perform better in sensorimotor exploratory tasks when considering high-dimensional non-linear redundant systems. In the next section, we explain this sensorimotor exploration architecture with more detail.

Another interesting contribution regarding the active learning architectures to learn sensorimotor regularities is presented in Ribes et al. (2016). Therein, the authors considered time

constraints and 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 Moulin-Frier and Oudeyer (2013b), models employed in Ribes et al. (2016) were developed on the basis of Gaussian Mixture Models (GMM).

In Tenenbaum et al. (2011) an analysis of the importance of the Bayesian approach to the understanding of how human minds work and develop is presented. It is defended that it provides tools for unifying mathematical language for framing cognition as the solution to inductive problems. Powerful abstractions can be learned surprisingly quick using the Bayesian approach. Moreover, Tenenbaum et al. (2011) argued that structured symbolic representations should not be rigid, static, hard-wired, or brittle. Within a probabilistic knowledge, they can grow dynamically and robustly in response to the uncertainty in the data collected from pure-perception and perceptuo-action. In language, these ideas fit with the claims of Kuhl (2004), who suggests that infants use probabilistic and statistical learning mechanisms.

As mentioned by Sandini et al. (1997), to design an artificial agent from using the developmental approach, the first practical problem is to define the subset of sensor and motor skills. A critical issue is the implementation of a complete system and the definition of constraints and abilities at the system 'birth'. The second problem is the definition of a computational framework for sensorimotor coordination compatible with emergence, self-organization, and adaptability.

2.4 Intrinsically Motivated Sensorimotor Exploration

As concluded in the previous section, some of the most prominent architectures for sensorimotor exploration are based on goal-directed motor babbling, where sensory goals are actively chosen according to a model of interest. This model represents how well the agent is performing in reaching self-generated goals through time. Thus, the agent can choose goals that are likely to improve its sensorimotor control skills according to a competence function (Baranes and Oudeyer, 2013, Moulin-Frier and Oudeyer, 2013b). In other words, exploration

occurs over regions in which agents perceive they are becoming more competent to reach self-generated goals. Thus, allowing them to efficiently and actively explore and generate maps from motor capacities to perceived results in interesting sensory space regions. The intrinsically motivated sensorimotor exploration architecture for embodied artificial agents is shown in Figure 2.1. This architecture is considered as the base for the contributions of this work.

To build the sensorimotor exploration architecture, the following elements are required:

- **Physical Embodiment** consists of a sensorimotor system.
- **Sensorimotor Model** is an internal representation that maps motor commands to sensor results. It must be capable of inferring motor commands from provided sensory goals.
- **Interest Model** is the core of the intrinsic motivation mechanism. It allows an active selection of sensory goals according to the evolution of competence measurement through the exploration in order to maximize learning progress.

In Figure 2.1, the learner starts with no knowledge about any of the two models. First, the models are initialized in a first stage. Once they are initialized, the intrinsically motivated exploration begins. The interest model proposes a sensory goal which is then passed to the sensorimotor model. The sensorimotor model computes the motor command that, according to the current knowledge, would produce that sensory goal. Then, the learner executes the selected motor command with its motor system and produces salient signals that are perceived as the sensory outcome. Afterward, the sensory outcome is compared to the sensory goal to generate the competence value, c of the experiment as an index of performance. The signals generated, described by blue arrows in the diagram, are then used to train the models. After training the models, the exploration starts again choosing a new sensory goal.

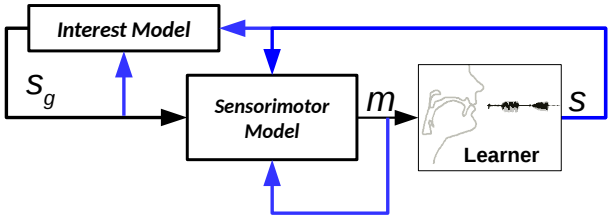

FIGURE 2.1: Intrinsically motivated sensorimotor exploration architecture. Black lines represent the flow of data during each action execution. Blue lines represent signals used to update the models.

2.4.1 The Competence Function

One important concept in intrinsically motivated exploration is how to measure the competence of an agent to produce self-generated sensory goals. Moulin-Frier and Oudeyer (2013b) adopted the competence function:

$$c_1 = e^{-|\mathbf{s}_g - \mathbf{s}|} \tag{2.1}$$

where $\mathbf{s}_g$ is the sensory goal and $\mathbf{s}$ is the actual production of the agent. This function in general assigns higher competence to those experiments that produce lower errors. Along this work we consider Eq. (2.1). In general, we observed that Eq. (2.1) minimizes the error to produce intended sensory goals. However, there are alternatives based on other concepts, for instance in Acevedo-Valle et al. (2017a) we studied the competence function from Baraglia et al. (2015), which can be used as modulator of exploration. The function is written as:

$$c_2 = \alpha e^{\frac{-(|\mathbf{s}_g - \mathbf{s}| - \mu_e)^2}{2\dot{\sigma}^2}} \tag{2.2}$$

where α is a scaling parameter, μ_e is the mean value of $|\mathbf{s}_g - \mathbf{s}|$ and σ its variance. The relevant feature of this competence function is that it fosters exploration in those regions where the error to reach self-generated goals is moderate. In other words, regions where sensory goals are not too easy to be reached by the agent but are not too hard to reach either. However, we did not observe relevant changes in the results, that is the reason we decided to keep Eq. (2.1). It is important to clarify that in Acevedo-Valle et al. (2017a), the three parameters of Eq. (2.2) were considered constants. Thus, studies with a better implementation of this competence function would be of interest.

2.4.2 Algorithmic Architecture

A feasible algorithmic implementation of the architecture in Figure 2.1 is shown in Algorithm 1. As shown in the algorithm, the learner starts without any experience producing intended goals. In line 1, the sensorimotor model $M_{\mathcal{SM}}$ is initialized using random motor experiments comparable with spontaneous random acts or rhythmic behaviors in children. The interest model M_{IM} is initialized in line 2 using the sensory outcomes of the agent in line 1 as initial goals. Next, in line 4 of Algorithm 1, the learner selects a goal $\mathbf{s}_g$ for the next experiment according to M_{IM}. Then the motor command $\mathbf{m}_i$ trying to produce $\mathbf{s}_g$

is computed using the sensorimotor model $M_{\mathcal{SM}}$ in line 5. Next, the motor command is executed with the embodied sensorimotor system and the learner observes $\mathbf{s}$ in line 6. In line 7, the learner evaluates the competence value c. Then a function governing the training of models is called, and the exploration continues back in line 4.

Algorithm 1 Intrinsically motivated sensorimotor exploration with goal babbling.

Set $\{n_e, randomseed\}$

1: Initialize $M_{\mathcal{SM}}$
2: Initialize M_{IM} and $i \leftarrow 1$
3: **while** $i \leq n_e$ **do**
4: $\quad \mathbf{s}_{g,i} \leftarrow sample\,(M_{IM})$
5: $\quad \mathbf{m}_i \leftarrow M_{\mathcal{SM}}\,(\mathbf{s}_{g,i})$
6: $\quad \mathbf{s}_i \leftarrow f\,(\mathbf{m}_i) + \sigma$
7: $\quad c_i \leftarrow e^{-|\mathbf{s}_{g,i} - \mathbf{s}_i|}$
8: $\quad i \leftarrow i + 1$
9: $\quad train\_models()$

2.5 Speech: Perception and Production

As mentioned in the introductory chapter, our main interest is to study the emergence of speech in infants. Understanding all the mechanisms underlying speech emergence in infants would be an important contribution to many areas of robotics and natural sciences. Thus, in the coming sections, we provide a review of a few studies that we consider relevant to roboticists working on artificial vocal development, emergence of language in machines, speech perception and production, and so on.

Recent studies highlight the possibility that language evolved to meet the needs of young man beings. Meeting their perceptual, computational, social and neural abilities, produced a specific communication system that can be acquired by all typically developing humans: speech (Kuhl, 2004). The knowledge about human speech processing is still insufficient. Moore (1994) presented twenty themes considered to be important to achieve a greater understanding of the nature of speech mechanisms and speech pattern processing. Regarding those twenty themes, Anusuya and Katti (2009) argued that the answers to those questions were still not clear a few of years ago, and to our knowledge, the majority of those question is still unanswered. Some of the most interesting question regarding the nature of speech asked by Moore (1994) are:

- How important is the communicative nature of speech?

- Speech technology or speech science?
- How much effort does speech need?
- What is a good architecture for speech processing?
- How important are physiological mechanisms?
- What are the mechanisms for learning?
- What is speech good for?
- How good is speech?

Speech is produced when the vocal fold vibration by the lung air flow provides a source signal of fundamental frequency F_0. The vocal tract acts as a resonator, and according to its shape the harmonics of the fundamental frequency are amplified or faded. The local maxima of the resulting spectrum are the formants frequencies, ordered from the lower to the higher frequencies (Moulin-Frier et al., 2013). They are the most frequent speech feature mentioned in the literature, mainly when describing vowels.

One of the central questions regarding speech is whether the systems for production and comprehension are essentially the same or not. As mentioned in Tooley and Bock (2014), a debate has been created using empirical evidence supporting both ideas. On the one hand, supporters defending that production and comprehension are carried by two separable processing systems, take some arguments as the emergence of comprehension before production in early development. On the other hand, supporters of a non-separable processing system argue that the shift of emergence between production and comprehension is due to the complexity of fine motor control acquisition. Under that condition, evidence is more consistent with substantial similarity across production and comprehension. Thus, evidence supports the idea that there exists a linkage between the systems underlying the production of sounds and the ones underlying their perception. There is available evidence of these linkages between perception and the motor system for monkeys and humans (Galantucci et al., 2006).

Tooley and Bock (2014) hypothesized that the major sources of difference between comprehension an production might be extrinsic to dedicated mechanisms of production and perception of speech. For instance, encoding complex ideas in a speech signal may require a different depth of mental processing, which the perceiver rarely needs. In conclusion, spoken language production and comprehension operate in similar ways and on similar principles. In

this sense, Liberman and Mattingly (1985) established that the skills to coarticulate speech and to perceive coarticulated speech evolved together given that neither skill would be useful without the other.

As a product of evolution, language through speech is an "optimal" or at least "optimized" communication system based on perceptually-shaped articulatory gestures (Galantucci et al., 2006). Spoken language requires parity, which refers to the fact that speakers and listeners can access similar information about words and their combination to express an idea. Any theory of speech must explain how the parity requirement is met (Galantucci et al., 2006, Liberman and Whalen, 2000, Tooley and Bock, 2014). In Liberman and Whalen (2000), the notion of parity was interpreted in three ways. The first is that listeners and talkers must converge on what counts as a linguistic action. The second is that phonetic messages sent and received must be the same. The third one is that production and perception specializations for speech must have co-evolved. In fact, parity is intended to be an abstract constraint on the symmetric co-evolution of the machinery for producing and perceiving speech.

A large body of scientific results supports the existence of neural linkages between the perceptual and motor systems and the involvement of motor competence in perception (Galantucci et al., 2006, Liberman and Whalen, 2000, Schwartz et al., 2012, Tooley and Bock, 2014). For communication purposes, this hypothesis implies that the system that produces a signal of communicative value is connected to the system that perceives the signal. These linkages have been found in perception and the motor system of monkeys and humans. An interesting implication, also related with embodiment, is that knowledge of anatomical constraints affects what people perceive (Galantucci et al., 2006). If perception implies motor competence exploitation, then knowing that a visual pattern can only correspond to a specific human motor action then this information is likely determinant to perceive the action results. One can speculate that specialized motor competence is called upon in the perception of the pattern. Another observation supporting the relevance of embodiment is that perceptual performance may be enhanced in the case of movements produced by the same individual who perceives them because the maximal amount of motor competence is available to support perception.

According to the evidence collected in Schwartz et al. (2012), some principles affect the organization of vowels systems and supports the optimized nature of speech. First, the *Dispersion-Focalization Theory* of vowel systems (Schwartz et al., 1997) establishes that vowel systems, as auditory-optimized structures obeying perceptual dispersion principles for

maximizing their efficacy of communication. Focal vowels correspond to more stable auditory patterns, which seem to drive both infant and adult perception and infant production. Secondly, the principle of *Maximal Use of Available Features* says that systems would combine features orderly. In the case of vowel systems, height features would be combined with tongue/lip configuration features to provide balanced systems. Finally, the *Maximal Use of Available Controls* (MUAC) suggests that in the course of speech development, the young speaker would achieve a sufficient control within a given series, and then transfer the adequate control to another tongue/lip configuration.

Based on observations made in a considerable number of experiments, the most adopted theories of speech affirm that speech perception is organized, at least partially, in terms of motor control signals and their associated vocal tract configurations. In the following, two theories discussed in Galantucci et al. (2006) and Schwartz et al. (2012), respectively, are studied in order to provide some insights on the mechanism underlying speech perception, they also provide some information regarding speech production.

2.5.1 The Motor Theory of Speech Perception

The Motor Theory of Speech Perception (MTSP) was proposed by Liberman et al. (1967) and revised in Liberman and Mattingly (1985). Besides the impact that MTSP had on the study of speech, it also gained a positive reception outside its field, mostly within the research and theorizing in the broad context of cognitive science (Galantucci et al., 2006). The theory proposes that phonetic coarticulated gestures are motor objects of speech perception. Thus, intended gestures instead of actual vocal tract actions were established as the fundamental objects of speech perception (Liberman and Mattingly, 1985). In Galantucci et al. (2006), based on new scientific evidence coming after the publication of the MTSP, some important conclusions regarding the claims of Liberman and Mattingly (1985) are provided.

Liberman et al. (1967) claimed that during the process of speech, perception articulation and sound wave go through separate ways. Furthermore, they claimed that perception goes with articulation instead of sound. However, according to Galantucci et al. (2006), this is not accurate. For instance, this is not true when different second formant transitions can signal the same phoneme or identical stop bursts can signal different phonemes before different vowels. *How distinctiveness and similarity are encoded?* Galantucci et al. (2006)

hypothesized that they are encoded by means of acquired similarity, whereby associating different acoustic signals for the syllables to the same response makes the syllable-initial consonants alike. On the other hand, acquired distinctiveness may explain how similar acoustic signals with different underlying articulations come to sound distinct (Galantucci et al., 2006).

In general, for skilled perceivers, the consequence of the memory representation established by the mimicry is that articulatory movements and their sensory effects mediate between the acoustic stimulus and the event we call perception (Galantucci et al., 2006, Liberman and Mattingly, 1985). In fact, Schwartz et al. (2012) argued that a listener, who knows something about speech production, exploit this knowledge to disentangle the complexity of the acoustic input and access the functional unit more directly related with motor commands. In this sense, the same reasoning is proposed for multimodal perception, e.g., audiovisual interactions in speech perception, which are claimed to be related to the knowledge of the listener about the multimodal coherence of a speech gesture. In this context, it is said that perception involves a procedural knowledge of action.

There is considerable evidence that listeners situate the acoustic signal in a space that captures its gestural causes. Furthermore, the hypothesis that gestures are the objects of speech perception provides a unified account of all of the findings: perceiving speech is perceiving phonetic gestures (Galantucci et al., 2006).

Some speech gestures may be specified by information other than air pressure waves, e.g., labial gestures. When it does, a natural question is whether speech perception is responsive to these additional sources of information. For example, listeners perceive speech in noise presence more accurately when they can see the speaker than when they cannot. For instance, speech imitation responses, which require gesture computation, are faster than non-imitative responses. Thus speech perception must include the gestures that allow the imitation. However, if listeners do perceive gestures, the model matching syllable may have served as a stimulus for an imitative response (Galantucci et al., 2006).

After the discovery of the mirror neuron system in monkeys (Rizzolatti and Craighero, 2004), similar systems were found in humans for finger, hand, and arm movements. This discovery suggested that in primates there is a fundamental mechanism for action recognition. The

mechanism consists in that individuals recognize actions made by others because the neural pattern elucidated in their premotor areas during action observation is similar to that internally generated to produce the action (Galantucci et al., 2006).

The human mirror-neuron system has been proposed to play a fundamental role in speech processing by providing a neurophysiological mechanism that creates parity between the speaker and the listener. In this context, evidence suggests that perception would be mediated by "motor ideas" represented in the brain by sensorimotor neurons enabling to access these "ideas" both, through acting and through observing or hearing somebody else acting (Schwartz et al., 2012). There is evidence that perceiving speech involves neural activity of the motor system. Moreover, studies have demonstrated the activation of speech-related muscles during the perception of speech. Some authors proposed that when simultaneous activation of the perceptual and motor codes occurs, both codes may interact (Galantucci et al., 2006, Schwartz et al., 2012).

2.5.2 The Perception-for-Action-Control Theory

Schwartz et al. (2012) introduced one of the most recent views concerning the relation of speech production and perception. A common debate is to determine whether speech perception involves auditory or multisensory representations and processing, independently on any procedural knowledge about the production of speech units or the contrary if it is based on a recording of the sensory input in terms of articulatory gestures (Liberman and Mattingly, 1985, Schwartz et al., 2012). The proposed Perception-for-Action-Control Theory (PACT) aims at defining a theoretical framework connecting in a principled way, based on behavioral and neurophysiological data, perceptual shaping and motor procedural knowledge in speech multisensory processing in the human brain.

As claimed by the motor theory of speech perception (MTSP), the coarticulation-driven composition of articulatory commands during speech production is non-linearly transformed into a complex composition of acoustic features, so that the acoustic properties of speech sounds are not invariant but context dependant (Schwartz et al., 2012). Given that MTSP does not fix with all the evidence, Schwartz et al. (2012) tried to answer what happens when the relationship between gestures and sounds is many-to-one, and hence the gesture cannot be, in theory, recovered from the sound without additional pieces of information. Evidence

suggests that a gesture is characterized by its functional value, likely evaluated in auditory terms. One possibility is to acknowledge that gestures are not pure motor units, but rather perceptuo-motor units, gestures being shaped by perception. Gestures are not only shaped by perception but also selected in relationship whit their perceptual (acoustic-auditory) value (Schwartz et al., 2012).

PACT considers that speech perception is the set of mechanisms that enable not only to understand, but also to control speech, considered as a communicative process. Thus, first, perception and action are co-structured in the course of speech development, which involves producing and perceiving speech items. Secondly, perception provides action with at least auditory templates, which contributes to defining gestures, providing them objectives, organization schemes, and functional value. In PACT, the communication unit, through which parity may be achieved, is neither a sound, nor a gesture, but a perceptually-shaped gesture, that is a perceptuo-motor unit. It is characterized by both, its articulatory coherence and its perceptual value (Schwartz et al., 2012).

In speech scene analysis, auditory, phonetic and lexical mechanisms are involved. Articulatory coherence is relevant. The fact that articulatory constraints may act on the emergence and stabilization of verbal transitions strongly suggests that they partly rely on motor neural processing. Evidence demonstrates that articulatory based representation plays a crucial role in the endogenously driven emergence and stabilization of auditory speech percepts during a verbal transformation task. The listener combines general auditory scene analysis mechanisms with articulatory principles grouping the acoustical and visual pieces of information coherently and relying on visible labial onsets (Schwartz et al., 2012).

The co-structuring component suggests the possibility to connect perceptual and motor representations for speech communication in a principled way. The sensorimotor maps appear to be dynamic and provide a way by which changes in production can result in changes in perception (as reported for example in Ito et al. (2009)), relating articulatory changes to perceptual changes (Schwartz et al., 2012). Two functions could be associated with perceptuo-motor connection. First, sensorimotor maps that are constructed during the infants' development and later on, all along life, through dynamic adaptations under diverse learning conditions. Secondly, the motor system could be involved in speech perception, perhaps more in adverse conditions, in order to provide a better specification of possible auditory and visual trajectories and enhance speech scene analysis (Schwartz et al., 2012).

Finally, MUAC suggests that a young speaker can transfer adequate control from previous knowledge to another tongue/lip configuration. As a consequence, the developmental articulatory pathway is supposed to be crucial in the achieved shape of vowel system for a given speaker. Thus, vowels are vocalic gestures organized developmentally and shaped by their acoustic/auditory properties. The next section introduces some works regarding the developmental trajectory of speech acquisition.

2.6 Speech and Development

Research on prelinguistic infants demonstrated that they were able to detect most phonetic contrasts at birth (Galantucci et al., 2006, Kuhl, 2004). Moreover, learning to control spoken language has been demonstrated to be constrained at perceptual, computational, social and neural levels determining what (and when) can be learned (Kuhl, 2004). As remarked in Perkell et al. (2001), it is clear that human speech production is one of the most complex motor acts performed by any living being. Producing a linguistic message that can be understood by another human requires exact and rapid coordinated movements of many degrees of freedom in the respiratory, laryngeal and supraglottal articulatory systems. How infants acquire the sophisticated ability to control speech production, which is practically fully developed at the age of 3 years old, and in general how they learn language remains a matter of research (Kuhl, 2004, Lenneberg et al., 1967).

During infancy, significant changes occur in the ways by which human body moves within its environment and the manners an infant interacts with that environment. Although sensorimotor behaviors, as mentioned in Section 2.1, are undoubtedly indices of underlying cognitive change, they also represent advances in the action capabilities of an infant. These changes may impact the development of skills and experiences that play a role in the emergence of communication and language (Iverson, 2010).

In the investigation of the learning process for spoken language, it is essential to define what is innate knowledge and skills, and which knowledge and skills are learned through development. As Kuhl (2004) mentioned, the rules by which infants perceive information, the ways in which they learn words, the social contexts in which language is communicated and the need to remember the learned entities for a long time probably influenced the evolution of language.

In general, it is important to identify constraints on infant learning, from all sources, and determine whether those constraints reflect innate knowledge that is specific to language.

One of the most important events in infants' vocal development for spoken language is the onset of a particular rhythmic behavior: Canonical Babbling (CB). Unlike the sounds that infants vocalize before this stage, such as crying, cooing, and screaming, CB is characterized by the production of well-formed syllables that have adult-like spectral temporal properties. The onset period of CB is stable across infants (7-8 months of age), regardless of their language environments. CB also emerges, with some differences, in infants with hearing loss. During this period, the sounds most favored by infants are also produced most accurately and occur with higher frequency in the languages of the world (Ejiri, 1998).

Based on the facts that rhythmic structure is not unique to vocal behavior and that the appearance of rhythmic vocalizations is developmentally linked to the more general appearance of rhythmically organized motor behaviors, Kent et al. (1991) suggested that the rhythms of vocalizations probably should be considered in terms of a larger picture of developmental rhythms. Early vocalizations produced by the infants regardless of their audibility, may be caused by infants' natural tendency to move their body parts rhythmically, early vocalizations are at first motivated largely by infants' sensorimotor feedback (Ejiri, 1998).

Early babbling, as claimed by Ejiri (1998) may be a direct result of rhythmic mandibular oscillations. However, evidence suggests that, around the onset of CB, infants learn to vocalize based on auditory feedback (Ejiri, 1998, Kuhl, 2004, Perkell et al., 2001). As an exploratory sensorimotor behavior, CB is a milestone in early language development. Through sensory-guided motor experiments, it helps to form and maintain internal body representations for the production of speech. Those representations are considered a fundamental prerequisite to more complex capabilities (Schillaci et al., 2016).

Evidence suggests that different from some motor skills that are determined by maturation, several aspects of early language development are not only determined by general maturation (Iverson, 2010). Based on evidence of developmental studies, Iverson (2010) argued that the acquisition of motor skills provide infants with an opportunity to practice skills relevant to language acquisition before they are used for that purpose. Moreover, the authors showed that emergence of new motor skills changes infants' experience with objects and people in ways that are relevant for both general communicative development and the acquisition of

language. During the first eighteen months of life, infants acquire and refine a whole set of new motor skills that, as mentioned before, significantly change how the body moves in and interacts with the environment.

Practicing sensorimotor skills related to language in the context of a precise action provides infants with immediate and salient visual, auditory and kinesthetic feedback. This feedback is an opportunity to observe perceptuo-motor links and means to begin noticing and attend to the relationship between their motor actions and consequences. When infants subsequently begin to babble, they may very well be better prepared to recognize the contingent auditory feedback from their sound production, feedback that allows them to monitor and adjust the state of the vocal tract as they vary their sound production (Ejiri, 1998, Iverson, 2010).

Developmental progression in action on objects and achievements in early language are closely associated. A complementary perspective is that physical action on objects sets context for attributing meaning to those objects via action. As infants refine their actions, they can attribute increasingly specific meanings to objects interacting with them. They contribute, directly or indirectly, to the development of communication and language, even before infants use those skills for that purpose, for example, the recognitory gesture. This skill is of particular relevance for learning words. Thus, the infants' first words are tightly bounded to action and infants are highly likely to name objects as they act on them (Iverson, 2010).

Infants learn rapidly from exposure to language in ways that are unique to humans, combining pattern detection and computational abilities. Following the same path regardless of culture, infants learn their mother tongue rapidly and effortlessly, by the age of 3 years they can produce full sentences. The idea that speech is a deeply encrypted code is widely accepted. The absence of early exposure to the patterns that are inherent in natural language produces life-long changes in the ability to learn a language. Language acquisition requires the commitment of the brain to patterns that reflect natural language input (Kuhl, 2004).

Some experimental results suggest that social interaction assists language in complex settings. In fact, social deprivation has a severe negative impact on language development, to the extent that standard language skills are never acquired. Speech sounds are strongly preferred in typically developing infants. Social influence is crucial, if simple auditory exposure to language prompts learning, the presence of a live human being would not be essential.

However, according to evidence, infants are apparently not computational automatons, but they might need a social tutor when learning natural language (Kuhl, 2004).

For infants, early social awareness is a predictor of later language skills. Social interaction can be conceived as a regulator of computational learning and thereby protects infants from meaningless calculations. The need for social interaction would ensure that learning focuses on speech that derives from humans in the infants' environment, rather than on signals from other sources (Kuhl, 2004).

Constraints are evident when infants hear or see non-human actions: infants imitate vocalization rather than sine waves analogs to speech and infer and reproduce intended actions displayed by humans but not by machines. These observations imply that social factors may affect language acquisition because language evolved to address a need for social communication. However, the mechanism that controls the interface between language and social cognition remains a mystery (Kuhl, 2004).

2.7 Developmental Robotics and Speech

Automatic Speech Recognition (ASR) is the process of converting a speech signal to a sequence of words using a computational algorithm. It makes possible for a machine to follow human voice commands and 'understand' human languages. ASR systems are widely used for human-machine interfaces, for example, call processing in telephone networks, speech transcription, voice dictation, robotics, and so forth (Anusuya and Katti, 2009).

The progress in ASR systems has been notable with the advances in Deep Neural Networks, consider Google Duplex (Hyken, 2018). ASR systems work well for a particular task if sufficient data is provided for the target domain. However, when ASR systems are migrated from laboratory demonstrations to actual applications, they encounter some serious difficulties (Jiang, 2005). Current speech recognition systems are easily outperformed in the case of non-restricted vocabulary, if the speaker is not well-known by the system and if noise reduces the speech signal quality (Kröger et al., 2009). There are technological barriers to flexible solutions of ASR, the main drawbacks are related to the sensitivity to the environment, the weak representation of grammatical and semantic knowledge and the variation naturally

present in speech. Inter-individual elements burdening the ASR are differences in accent, speaking style, speaker psychology, age, emotions, among others (Benzeghiba et al., 2007).

According to its capabilities, speech recognition can be classified into four different types: isolated words, connected words, continuous speech and spontaneous speech. On the other hand, according to the algorithmic approach, there are three approaches to speech recognition: Acoustic Phonetic Approach, Pattern Recognition Approach, and Artificial Intelligence Approach, the latter could be considered a hybrid of the two earlier (Anusuya and Katti, 2009, Benzeghiba et al., 2007).

The two prevalent techniques for ASR are Hidden Markov Models (HMM) of speech signals and decoding techniques for very-large-scale networks (Benzeghiba et al., 2007, Jiang, 2005). In most ASR systems, a signal is modeled through HMMs, at a first stage ASR front-end analyze short signal frames on which stationarity is assumed. As found by Liberman and Mattingly (1985), speech gestures are coarticulated. The effects of coarticulation have motivated studies on segment based, articulatory, context-dependent modeling techniques (Benzeghiba et al., 2007).

Acoustic modeling for ASR uses very little of the available knowledge about the speech production system. Thus, speech is only modeled as a surface phenomenon, omitting sources of information that may considerably improve available technologies. Articulatory features may be used for language modeling. There are some speech production approaches to ASR based on statistical models. The models are used in a consistent probabilistic framework, where evidence from the acoustic language, the lexicon, and the language model are combined to reach a final decision (King et al., 2007), which is more similar to the principles considered in the PACT theory previously introduced (Schwartz et al., 2012).

Summarizing, human subjects produce one to two orders of magnitude fewer errors than machines in most of the speech recognition tasks. One of the drawbacks to reduce this gap is the limited knowledge about human speech processing. However, as mentioned in previous sections, the theories of human speech perception have evolved rapidly in the last twenty years based on different psychological studies and neurophysiological evidence. Thus, one of the challenges for creating machines with human language capabilities is to integrate this knowledge into intelligent machines. ASR technology has achieved significant results to pupulate the academical and industrial areas. However, significant advances may instead

come from studies in acoustic-phonetics, speech perception, linguistics, and psychoacoustics (Anusuya and Katti, 2009).

We require to endow ASR systems with an efficient way of representing, storing, and retrieving knowledge required for natural conversation. To achieve this system, we could take advantage of the available studies in infants, to create a system that can efficiently acquire the ability to produce and perceive speech. Thus, we would open the door to more realistic human-machine interactions through spoken language in unstructured environments.

2.8 Prelinguistic Vocal Development in Machines

As explained in the previous section, evidence available in psychological and neurophysiological studies can help us to build machines endowed with spoken dialogue-based communication systems. Theories of human speech perception and production have evolved rapidly in recent years. Thus, roboticists can take advantage of this knowledge to build talkative robots. Especially, we argue that a successful way to construct such a robotic system would be to get inspiration from the developmental trajectory observed in infants. Building a robot that would be able to acquire speech in the same way as an infant would be beneficial for building technological solutions and the understanding of the human mind. Thus, it also could contribute to solving problems occurring during human development.

Regarding language and phonological meaning, phonological representation of a target language is not present at birth, but it emerges during speech acquisition (Kröger and Cao, 2015). Thus, the process of speech perception and production acquisition and their stages must be important. Different stages can be observed through the acquisition of speech production and perception by an infant, and also through their communicative value (Kuhl, 2004). Regularities can be observed in the structure of the vocal development process independently of inter-individual differences (Kuhl, 2004, Oller and Eilers, 1988). Evidence suggests that during Canonical Babbling (CB), infants learn to control their ear-vocal tract system based on auditory feedback. In general, infants firstly discover how to control phonation, then focus on vocal variations of unarticulated sounds and finally, in an apparent automatic manner, discovers and focuses on babbling with articulated proto-syllables. Our experiments consider the period around the CB. In this prelinguistic stage, production of

speech utterances may not be associated yet to linguistic meaning, but certainly strong cognitive architectures are build to foster linguistic value emergence.

Important issues to be addressed in early vocal development studies are summarized from Asada (2016) and Mutlu et al. (2016):

- Integration of neuroscientific approaches focusing on neural mechanism inside the learner and interactive ones focusing on social learning issues.
- The relationships between multimodal sensations, not only auditory but also vision and touch should be analyzed.
- Realistic interactions and more experiments with humans.

In infants with regular development, there exists an ordered number of typical stages emerging along the progress from newborns to fully functioning adults (Morse and Cangelosi, 2017). Some works are attempted to explain the emergence of developmental stages during vocal development using artificial intelligence techniques. However, those works do not provide any explanation for the onset of developmental stages. Recently, a model of language development stages from the embodied perspective was introduced in Morse and Cangelosi (2017). However, their efforts are rather directed toward language level development, leaving early vocal development as an open issue. In the following, we describe a series of work that has attempted to study prelinguistic vocal development using different perspectives and focused on specific features of this developmental stage.

From the perspective of developmental robotics applied to prelinguistic vocal development, one of the earliest works is Yoshikawa et al. (2003). Therein, the authors built a robotic human vocal tract and attempted to mimic the way in which humans acquire phonemes through random motor babbling and considering interactions with a caregiver. In general, Yoshikawa et al. (2003) focused on learning model of vowel acquisition despite the different embodiment between a robot and a human.

In Guenther (2006), Guenther et al. (2006), the DIVA model is introduced as a tool to study the neurophysiological mechanisms for speech acquisition and production. Studies focused on prelinguistic and early linguistic language-specific training, especially on the neural pathways for acquiring speech units and considered a simulated vocal tract as the physical embodiment.

In a first stage, the synaptic projections are tuned during a babbling phase in which quasi-random articulatory movements are used to produce auditory and somatosensory feedback. In a second stage, the acquired knowledge is used to build speech sounds. The neural network takes as input a speech sound string and generates as output a time sequence of articulator positions to command the movements of the simulated vocal tract. After babbling, the model can quickly learn to produce new sounds from audio samples provided to it.

Inspired by Guenther (2006), Guenther et al. (2006), a new neurocomputational production-perception model was introduced in Kröger et al. (2009). Similarly to the inspirational works, the new approach comprised self-organizing networks for processing neural states and comprise neural maps for storing phonemic, motor, and sensory states representing speech items. Three main differences with respect to the results from the original works can be highlighted. First, the separation between motor planning and motor execution. Secondly, the new model includes a phonetic map reflecting the self-organization of speech items between sensory, motor, and phonemic representation. Thus, bidirectional mappings are achieved between phonemic, sensory, and motor representations essential in a production-perception model. Thirdly, different to Guenther and colleagues, Kröger and colleagues aimed at modeling both speech production and speech perception.

In Kröger et al. (2009), there is a first stage of random babbling after which the neurocomputational model is capable of reproducing the motor plan state of some prelinguistic speech item from their acoustic state patterns. Hence, the neurocomputational model can perform a language-specific imitation training with training sets comprising language-specific speech items. In Kröger and Cao (2015) vocalic and syllabic speech items are considered for training. Based on a biologically inspired model of speech processing and using interconnected growing self-organizing maps, the phonetic-phonological interface is described here as a numerical computer-implemented model.

A speech acquisition model called *Elija* was developed in Howard and Messum (2011) and extended in Howard and Messum (2014). It can go from babbling to naming objects using infant-like utterances, but the onset of each stage is done by hand. In Howard and Messum (2011), motor patterns are learned by optimizing a reward function instead of combining simpler patterns. In *Elija*, there is a first stage of self-exploration as in the DIVA model. Then a stage of imitation is triggered, in which speech signals are obtained from a caregiver.

Finally, the acquired speech units are used to name objects through audio-visual acquired regularities during the imitation scenario.

In Warlaumont et al. (2013b), a neural network model is introduced to study the role of reinforcement during vocal learning using a vocal tract simulator. Random babbling occurs spontaneously, and if a vocalization meets specific acoustic criteria, it is reinforced, making similar muscle activation increasingly likely to recur. In the results, when reinforcement was contingent on both phonation and proximity to English vowels as opposed to Korean vowels, the model's post-learning productions were more likely to resemble the English vowels and vice versa. In Warlaumont (2013), the authors focused on a spiking neural network model that controls the lip and jaw muscles of a vocal tract simulator and learns to produce canonical babbling. The model was adapted to receive reinforcement when it produced a sound with high auditory salience. Salience reinforced versions of the model increased their rates of canonical babbling over the course of learning more than their yoked controls.

In the line of the research in this dissertation, it is possible to identify three similar works. As mentioned before, we started our experiments based on the results from Moulin-Frier et al. (2013). On the other hand, works following a similar line of research have been recently published (Forestier and Oudeyer, 2017, Najnin and Banerjee, 2017). In Moulin-Frier et al. (2013), intrinsically-motivated learning was used to study the emergence of developmental stages during the first year of an infant life using a computational model. In Forestier and Oudeyer (2017), the authors argued that most of the previous works have the disadvantage of considering agents that are not situated in a physical environment where vocalizations may have a meaning related to objects. We consider that this is true, it may be arguable in the case of Howard and Messum (2011), where objects are considered but not any physical interaction with them. Forestier and Oudeyer (2017) proposed to study intrinsically-motivated sensorimotor exploration applied to language emergence in a scenario of reaching objects, where objects could be reached with a robotic arm, with a tool or asking a caregiver for help. Even though vocal learning was somehow more restricted compared to the original proposal in Moulin-Frier et al. (2013), the results and the experimental setup are impressive.

2.9 The Role of Intrinsic Motivation in Vocal Development

As established in previous sections, infants firstly discover how to control phonation, next they focus on vocal variations of unarticulated sounds and finally, apparently automatically, they discover and focus on babbling with articulated proto-syllables. To achieve this goal, they must learn redundant non-linear high-dimensional mappings of the ear-vocal tract system. Previous works attempted to explain the emergence of developmental stages during vocal development assuming the existence of those stages and hard-coding the onset of each of them during experimentation (Guenther et al., 2006, Howard and Messum, 2011, Kröger et al., 2009, Warlaumont et al., 2013a).

Moulin-Frier et al. (2013) was a first attempt to understand other mechanisms that may explain the structured onset of developmental stages. The authors used an intrinsically motivated exploration architecture to study the onset of those stages. It was argued that intrinsic motivation might play an essential part in the self-organization developmental stages.

The Maeda's vocal tract implemented by Guenther Lab was used in Moulin-Frier et al. (2013) as a sensorimotor model along with the intrinsically motivated sensorimotor exploration explained in Section 2.4. The dynamics of the 10 articulators and the 3 voicing parameters of the Maeda's vocal tract were modeled as overdamped second order systems. Whereas, for the auditory output the two first formant also provided by the Maeda's synthesizer were considered with an extra signal that indicates if speech is produced or not was also considered to build the sensory space.

The embodied architecture depicted previously in Figure 2.1 was built with the following elements:

Physical Embodiment Maeda's vocal tract implemented by Guenther Lab.

Sensorimotor Model Gaussian Mixture Model (GMM) wit incremental learning based on Calinon (2009).

Interest Model GMM built in order to keep track of the progress of the competence (see Eq.(2.1)) with respect to the time for different self-proposed goals.

The results provided in Moulin-Frier et al. (2013) opened a door of a new approach in vocal development to be investigated. A feasible explanation to the trajectory of vocalizations complexity during early babbling was stated, where first an infant produces 'silent' vocalizations, then unarticulated vocalizations and finally more complex coarticulated vocalizations in an ordered transition between stages. The results also indicated that intrinsically motivated learning algorithms can successfully learn sensorimotor coordination skills in vocal spaces. They allow an artificial agent to learn to control its vocal tract progressively. However, more than being a concluding paper, the ideas presented opened a door of a new approach for early-vocal development to be explored.

This thesis expands the results observed in Moulin-Frier et al. (2013). First, in Chapter 4 we include constraint awareness into the architecture. Then, in Chapter 5 we study the role of imitative behaviors during sensorimotor learning in parallel to intrinsically motivated exploration.

Finally, another interesting work to be considered is Najnin and Banerjee (2017). Therein, the author also extended, but in a different direction than ours, the results in Moulin-Frier et al. (2013). In this case, a predictive coding framework was proposed for a developmental agent with perceptuo-motor and learning capabilities. As in the original work, the agent is solely driven by sensory prediction error. A similar developmental transition is observed, which was partially improved by the modifications in the perception systems, given that they considered the Mel-Frequency Cepstral Coefficients instead of the formant frequencies. They also showed that agents learn to vocalize differently in different environments.

Chapter 3

Incremental Learning and the Regression Problem with Gaussian Mixture Models

"I have had my results for a long time: but I do not yet know how I am to arrive at them."
— Karl Friedrich Gauss

In the last part of the previous chapter, recent works aimed at studying speech development from the perspective of artificial intelligence were introduced. The work by Moulin-Frier and colleagues (Moulin-Frier and Oudeyer, 2013b, Moulin-Frier et al., 2013) was remarked as a relevant referent to recent studies in the area (Acevedo-Valle et al., 2017a, Acevedo-Valle et al., 2018, Forestier and Oudeyer, 2017, Najnin and Banerjee, 2017). Moulin-Frier and colleagues also contributed with the toolbox `explauto` for Python, which is aimed at facilitating the implementation of sensorimotor exploration systems (Moulin-Frier et al., 2014). As the developments in this dissertation were, at the start, partially inspired by the results in Moulin-Frier et al. (2013), by default similar approaches were reproduced in our designed systems to replicate the results presented therein. Thus, herein we adopted Gaussian Mixture Models (GMMs) as a modeling approach for sensorimotor systems. Later, through this project, we have developed a learning framework to learn incrementally and solve the prediction and inference problems. As the main advantage, the introduced learning

framework allows learning from data batches without the need of keeping them in memory afterward.

This chapter is aimed at presenting an approach to achieve the solution of the regression problem for multivariate systems. It uses an efficient incremental learning algorithm which is compared to the state-of-the-art approach. Within the machine learning framework, incremental learning of multivariate spaces is of particular interest for online applications, as it is the case for the sensorimotor exploration problem that will be extensively studying in the next chapters. The algorithms introduced in this chapter allows learning high-dimensional redundant non-linear static maps from non-persistent on-line data of input-output systems. Studying the implementation of alternatives to GMMs using incremental learning to solve the regression problem is currently out of the scope of this work, so it is left as a research line for the near future.

Summarizing, inspired by the results in Moulin-Frier and Oudeyer (2013b), Moulin-Frier et al. (2013), a learning architecture is built using Incremental Gaussian Mixture Models in order to solve the regression problem. Hence, two interesting mechanisms are combined: incremental learning of GMMs and Gaussian Mixture Regression (GMR) to solve the inference and prediction problems. Two approaches for the incremental learning of GMMs are considered in order to compare our approach with state-of-the-art ones. On the one hand, an approach based on the codes provided alongside Calinon (2009) and used to obtain the results reported in Acevedo-Valle et al. (2015, 2018). On the other hand, an approach that was implemented during the development of this project and published in Acevedo-Valle et al. (2017b), which was used to obtain the results published in Acevedo-Valle et al. (2017a) and the results showed in Chapters 4-5. Python's source codes for the latter approach are available online for those researchers who are interested in testing this learning mechanism in their work[1]. Through this chapter, simple examples are used to facilitate the comprehension of the approach. In the following chapters, it will be shown that the approach also applies to systems as complex as a vocal tract simulator.

This chapter is organized as follows. Section 3.1 provides a brief introduction to the relevance of Gaussian Mixture Models in the machine learning domain. Section 3.2 defines the learning problem considered in this chapter. Later, Sections 3.3-3.4 introduce the two considered incremental learning algorithms for GMMs. Section 3.5 explains the mechanism to solve

[1] https://github.com/yumilceh/igmm

the regression problem for multivariate systems using GMMs. Section 3.6 presents a simple example that shows how the incremental learning approaches work. Section 3.7 introduces a simple non-linear redundant input-output system used to illustrate how the regression mechanism works to solve the inference problem, considering both learning approaches. Finally, a brief discussion is presented in Section 3.8.

3.1 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are linear combinations of multivariate Gaussian distributions that represent clusters of data. They are frequently appealed in machine learning applications and related areas for problems that require the clustering of data. In such a context, they are commonly employed in tasks where it is necessary to model complex and nonlinear parameters (Bouchachia and Vanaret, 2011). However, recently they have been actively applied to solve the regression problem, and they have also been used to model high dimensional, non-linear redundant maps (Acevedo-Valle et al., 2015, Acevedo-Valle et al., 2017b, Acevedo-Valle et al., 2018, Moulin-Frier et al., 2013, Oudeyer et al., 2007, Ribes et al., 2016).

On the other hand, incremental learning algorithms may play a critical role in many applications. Those algorithms consider the learning scenario for streaming data arriving over time and have been widely applied in machine learning, pattern recognition, data mining, and fuzzy logic (Bouchachia and Vanaret, 2011, Chen et al., 2012, Gepperth and Hammer, 2016). In Gepperth and Hammer (2016), a summary of the challenges for incremental learning is presented. Furthermore, some of the main techniques that have been applied to solve the problem are described. In general, most of the machine learning techniques have been extended to cover the incremental learning paradigm opening the door to new applications, e.g., Support Vector Machines, Decision Trees, Genetic Algorithms, Gaussian Mixture Models, among others.

Incremental learning using GMMs has been previously studied with more emphasis on its applications as a semi-supervised classifier method and density distribution estimator (Bouchachia and Vanaret, 2011, Chen et al., 2012, Engel and Heinen, 2010). However, in this chapter, we focus on its suitability to solve the regression problem using the approach

implemented in Acevedo-Valle et al. (2017b). Moreover, the results are compared with results using the generative approach implemented in Calinon (2009) and Calinon et al. (2007).

Both approaches are suitable, but not limited, to solve the problem of modeling input-output multivariate systems. The learning system must collect data incrementally as it is not available in advance in order to generate a model, so data is collected in batches of input-output data points. Therefore, the model is trained each time a new data batch is available and afterward that data batch is discarded.

3.2 Learning Problem Definition

In this section, the learning problem to be solved is defined according to the requirements of the intrinsically motivated sensorimotor exploration architecture from Moulin-Frier et al. (2013), later extended in our works Acevedo-Valle et al. (2015), Acevedo-Valle et al. (2017a), Acevedo-Valle et al. (2018).

First of all, a GMM is defined by the set of parameters $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$, where π_j, μ_k and Σ_k are respectively the prior probability, the distribution mean, and the covariance matrix of the k-th Gaussian, for $k = 1, 2, ..., K$, being K the number of Gaussian distribution components. A Gaussian distribution is defined as $\mathcal{N}(\mu, \Sigma)$, whereas the probability of a data point $\mathbf{z}$ to belong to that Gaussian distribution is defined as $\mathcal{N}(\mathbf{z}; \mu, \Sigma)$. The probability of $\mathbf{z}$ belonging to the mixture is defined as

$$P(\mathbf{z}) = \sum_{k=1}^{K} \pi_j \mathcal{N}(\mathbf{z}; \mu_k, \Sigma_k), \tag{3.1}$$

with

$$\mathcal{N}(\mathbf{z}; \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_k|}} \exp^{-\frac{1}{2}\left((\mathbf{z}-\mu_k)^T \Sigma_k^{-1} (\mathbf{z}-\mu_k)\right)}. \tag{3.2}$$

Furthermore, we assume a multivariate input-output static system defined as $\mathbf{y} = f(\mathbf{x}) + \varepsilon$. The vector $\mathbf{y}$ is assumed to belong to an m-dimensional output space $\mathcal{Y} \subset \mathbb{R}^m$, which is mapped by the inverse transformation $f^{-1}(\cdot)$ to an n-dimensional input space $\mathcal{X} \subset \mathbb{R}^n$ as the output vector $\mathbf{y}$, and ε is random noise.

Experiments are run with the system to generate data batches with k_{step} samples of the extended vector $\mathbf{z} = [\mathbf{x}, \mathbf{y}]^T$, thus $\mathbf{z} \in \mathcal{X} \times \mathcal{Y}$. Then, a GMM M_0 is computed to represent the distribution of the initial data batch $\mathbf{Z}_0 = \{\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_N\}$ over the extended space

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{Z} \in \mathbb{R}^{m+n}$. Once the mixture is estimated, the data batch $\mathbf{Z}_0$ becomes unavailable. When a new data batch $\mathbf{Z}_1$ is available, M_0 must be retrained. The learning mechanism must be able to update the starting GMM M_0 to represent the distribution that would be described by $\mathbf{Z}_0$ and $\mathbf{Z}_1$ together, generating a new mixture M_1 and making $\mathbf{Z}_1$ unavailable. Finally, this learning process must be repeated each time a new data batch $\mathbf{Z}_i$ becomes available to generate a mixture that models the distribution representing $\{\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3 \ldots\}$. In the following sections, two different approaches to performing incremental learning of GMM are introduced.

3.3 Generative Method for Gaussian Mixture Models

In Calinon (2009), two different approaches for incremental learning of GMMs were introduced, both of them use an online variant of the *Expectation-Maximization* algorithm (EM-algorithm). Both implementations of the algorithms are open source[2].

The first approach is defined as *direct method*. This method was discarded as an option for this work because, as mentioned in Calinon and Billard (2007), the method relies on the assumption that new data to be integrated into the model is close to the model. Due to our hypothesis that sensorimotor systems are redundant high-dimensional non-linear maps, we cannot make this assumption regarding new data.

On the other hand, the second approach, defined as *generative method*, uses a stochastic approach to update the models. This is the one considered in this chapter. Starting with an initial mixture model M_i, given a new data batch $\mathbf{Z}_{i+1}$, the model is updated to become M_{i+1}. To train the model, first a set of random data points $\mathbf{Z}'_{i+1}$ are generated using the data distribution represented by M_i. The number of generated points is

$$k_{gen} = \left\lceil \frac{(1-\alpha)\, k_{step}}{\alpha} \right\rceil,$$

where $\alpha \in [0, 1]$ is the forgetting rate, k_{step} is the number of samples in $\mathbf{Z}_{i+1}$, and $\lceil \cdot \rceil$ indicates the nearest larger integer function.

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

The model parameters $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ are updated using the data batch $\mathbf{Z} = \{\mathbf{Z}_{i+1}, \mathbf{Z}'_{i+1}\}$ according to the *Expectation-Maximization* algorithm. It uses as initial parameters those from M_i. Then the parameters are updated following the following steps:

E-step:

$$p_{k,j}^{u} = \frac{\pi_k^u \mathcal{N}(\mathbf{z}_j; \mu_k^u, \Sigma_k^u)}{\sum_{i=1}^{K} \pi_i^u N(\mathbf{z}_j; \mu_i^u, \Sigma_i^u)},$$

$$E_k^u = \sum_{j=1}^{K} p_{k,j}^u.$$

M-step:

$$\pi_k^{u+1} = \frac{E_k^u}{N},$$

$$\mu_k^{u+1} = \frac{\sum_{j=1}^{N} p_{k,j}^u \mathbf{z}_j}{E_k^u},$$

$$\Sigma_k^{u+1} = \frac{\sum_{j=1}^{N} p_{k,j}^u \left(\mathbf{z}_j - \mu_k^{u+1}\right)\left(\mathbf{z}_j - \mu_k^{u+1}\right)^T}{E_k^u}.$$

Iterations using the *E-step* and the *M-step* are repeated until reaching a stop criteria defined as $\frac{\mathcal{L}u+1}{\mathcal{L}u} < Tol$. Where Tol represents a tolerance, and $\mathcal{L}_*$ is the log-likelihood of the data batch $\mathbf{Z}$ for the given model, defined as:

$$\mathcal{L}(\mathbf{Z}) = \sum_{j=1}^{N} log\left(P\left(\mathbf{z}_j\right)\right) \tag{3.3}$$

where P is defined by Eqs. (3.1)-(3.2).

The incremental learning process using the generative method proposed by Calinon (2009) is summarized in Algorithm 2. In line 1, the three parameters of the model are chosen. In line 2, the initial model is computed iterating the `EM-steps` described above until reaching the given tolerance. From line 3, the model is trained every time a new data batch $\mathbf{Z}_i$ is available. Then, in line 4, the number of samples to be generated with the previous model M_{i-1} is computed according to the forgetting factor α and the number of samples in the

data batch $\mathbf{Z}_i$. In line 5, a set of k_{gen} points is obtained sampling the distribution described by M_{i-1}, the new data batch is concatenated with $\mathbf{Z}_i$ in line 6, and finally, M_{i-1} is updated to M_i iterating the `EM-steps` until reaching the given tolerance.

Algorithm 2 Generative Method to Train Gaussian Mixture Models

1: Set parameters: K, α, Tol.
2: $M_0 \leftarrow$ `EM-steps`$(\mathbf{Z}_0)$
3: **for** $\mathbf{Z}_i$ with i in $i = \{1, 2, 3, 4 \ldots\}$ **do**
4: $\quad k_{gen} = \lceil \frac{(1-\alpha) size(\mathbf{Z}_i)}{\alpha} \rceil$
5: $\quad \mathbf{Z}'_i = sample(M_{i-1}, k_{gen})$
6: $\quad \mathbf{Z} = \{\mathbf{Z}_i, \mathbf{Z}'_i\}$
7: $\quad M_i \leftarrow$ `EM-steps`$(\mathbf{Z})$

3.4 Incremental Gaussian Mixture Models

The learning procedure of the new approach considered for incremental learning of GMMs consists of two main steps: a first step using the *Expectation-Maximization* algorithm (EM-algorithm) to train GMMs and a second step, in which a *growing mechanism* allows to include new knowledge in previously trained GMMs based on general geometric properties of Gaussian distributions.

In Algorithm 3, the incremental learning algorithm used to train a GMM using data batches is summarized. Algorithm 3 is fed with the following parameters: the minimum and maximum number of Gaussian components in the model, K_{min} and K_{max}, respectively, the maximum number of Gaussian components that can be added to the model at each training step, ΔK_{max}, and the forgetting rate, α. In line 2, the GMM is initialized using the first batch of data $\mathbf{Z}_0$, the `getBestGMM` function computes a GMM for each value of K within the allowed interval $[K_{min}, \Delta K_{max}]$. From those models, the one that best fits the data batch according to the *Bayes Information Criterion* (BIC), which is based on the maximum likelihood function, is selected. We will call the selected model, M_0. In `getBestGMM`, the GMMs are obtained using the EM-algorithm implemented in the open-source library *scikit-learn*[3], but also available in other open source tools (i.e. *TensorFlow*, *Open-CV*, and others).

From line 3, the model is trained every time a new data batch $\mathbf{Z}_i$ is available. In line 4, a new GMM, M_{new}, is computed feeding the `getBestGMM` function with $\mathbf{Z}_i$. In lines 5 and

[3] http://scikit-learn.org/stable/

Algorithm 3 Growing Gaussian Mixture Model Process

1: Set parameters: K_{min}, K_{max}, ΔK_{max}, α.
2: $M_0 \leftarrow$ `getBestGMM`($\mathbf{Z}_0$, K_{min}, ΔK_{max})
3: **for** $\mathbf{Z}_l$ with l in $l = \{1, 2, 3, 4 \ldots\}$ **do**
4: $\quad M_{new} \leftarrow$ `getBestGMM`($\mathbf{Z}_l$, 1, ΔK_{max})
5: $\quad M_{new}.gauss[:].\pi \leftarrow \alpha * M_{new}.gauss[:].\pi$
6: $\quad M_{l-1}.gauss[:].\pi \leftarrow (1-\alpha) * M_{l-1}.gauss[:].\pi$
7: $\quad \mathbf{S_{KLD}} \leftarrow$ `getKLDiveregence`(M_{l-1}, M_{new})
8: $\quad$ **while** $M_{l-1}.k + M_{new}.k > K_{max}$ **do**
9: $\quad\quad i, j = argmin(\mathbf{S_{KLD}})$
10: $\quad\quad M_{l-1}.gauss[i] = merge(M_{l-1}.gauss[i], M_{new}.gauss[j])$
11: $\quad\quad delete(M_{new}.gauss[j]), \quad \mathbf{S_{KLD}}[i, j] = \infty$
12: $\quad M_l \leftarrow join(M_{l-1}, M_{new})$

6, the prior of each Gaussian component in M_{l-1} and M_{new} is updated, respectively. The prior's update is done according to the forgetting rate, α.

The most important step for the incremental learning mechanism is the merging of Gaussian components. Choosing which components of M_{new} will be merged to which components of M_{l-1} is the most challenging task of our approach. Therefore, before any components could be merged, a divergence matrix is obtained to evaluate the similarity between Gaussian components in M_{new} and M_{l-1}. Equally to Bouchachia and Vanaret (2011), we consider the *Kullback-Leibler divergence* (KLD) for two Gaussian distributions defined as

$$DKL(g_1, g_2) = \log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_1^{-1}(\mu_2 - \mu_1) - D, \tag{3.4}$$

where $g_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $g_2 = \mathcal{N}(\mu_2, \Sigma_2)$. As the KLD is not symmetric, we use a symmetrized version defined as

$$DKL_s(g_1, g_2) = \frac{1}{2}\left(DKL(g_1, g_2) + DKL(g_2, g_1)\right). \tag{3.5}$$

Finally, lines 8-11 of Algorithm 3 represent the merging process. Therein, the most similar Gaussian components in M_{new} are merged to their most similar counterpart in M_{l-1} and dropped. This process is repeated until the sum of components between both models is not greater than the maximum number of components K_{max}. Based on the geometric properties of Gaussian functions, the merge operation is summarized in Algorithm 4 from Bouchachia and Vanaret (2011). Finally, after the merging process, the remaining components in M_{new} are joint with M_{l-1} to become M_l in line 12.

Regarding Algorithm 4, it represents the steps needed to merge two Gaussian distributions, $g_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $g_2 = \mathcal{N}(\mu_2, \Sigma_2)$. These Gaussian distributions are assumed to be part of a GMM with priors π_1 and π_2, respectively.

Algorithm 4 Merge Gaussian Distributions

merge(g_1, g_2)

1: $f_1 = \frac{\pi_1}{\pi_1+\pi_2}, \quad \pi_2 = \frac{\pi_2}{\pi_1+\pi_2}$
2: $\pi_{new} = f_1 + f_2$
3: $\mu_{new} = f_1\mu_1 + f_2\mu_2, \quad \Sigma_{new} = f_1\Sigma_2 + f_2\Sigma_2 + f_1 f_2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$
4: $g_{new} \leftarrow \pi_{new}, \mu_{new}, \Sigma_{new}$
5: *return* g_{new}

3.5 Solution to the Regression Problem with GMR

Once we have introduced two different mechanisms to incrementally train GMMs, in this section we present the mechanism to solve the regression problem using the computed GMMs. The regression mechanism follows our previous works in Acevedo-Valle et al. (2015, 2018) and it is based on Gaussian Mixture Regression (GMR) from Calinon (2009). It is summarized in Algorithm 5. As defined in Section 3.2, an n-dimensional input space $\mathcal{X} \subset \mathbb{R}^n$ is mapped onto an m-dimensional output space $\mathcal{Y} \subset \mathbb{R}^m$. Thus, the function $\mathbf{y} = f(\mathbf{x}) + \varepsilon$ is assumed, where $\mathbf{y} \in \mathcal{Y}$, $\mathbf{x} \in \mathcal{X}$ and ε is random noise.

Herein, we solve the regression problem as an inference problem in order to determine the input $\mathbf{x}$ that maximizes the probability to produce the output $\mathbf{y}$. Considering the partitioned vector $\mathbf{z} \in \mathcal{Z}$ with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}, \tag{3.6}$$

once a GMM has been computed to model the distribution of a collection of data $\mathbf{Z} \in \mathcal{Z}$, for each j-th Gaussian in that GMM the partitions

$$\mu_j = \begin{pmatrix} \mu_j^x \\ \mu_j^y \end{pmatrix} \quad \text{and} \quad \Sigma_j = \begin{pmatrix} \Sigma_j^x & \Sigma_j^{xy} \\ \Sigma^{yx} & \Sigma_j^y \end{pmatrix} \tag{3.7}$$

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

$$\hat{\mu}_j = \mu_j^x + \Sigma_j^{xy}(\Sigma_j^y)^{-1}(y - \mu_j^y) \quad , \quad \hat{\Sigma}_j = \Sigma_j^x + \Sigma_j^{xy}(\Sigma_j^y)^{-1}\Sigma_j^{yx}. \tag{3.8}$$

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

$$P(\hat{z}_j) = \pi_j \frac{1}{\sqrt{(2\pi)^K |\Sigma_j|}} \exp^{-\frac{1}{2}\left((\hat{\mathbf{z}}_j - \mu_j)^T \Sigma_j^{-1} (\hat{\mathbf{z}}_j - \mu_j)\right)}, \tag{3.9}$$

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

Algorithm 5 Infrence Problem Solution with GMR

infer($\mathbf{y}$, $M = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$, nn)

$\mathbf{d} \leftarrow zeros(K, 1)$ % *distances*
for i in $range(K)$ **do**
  $\mathbf{d}[i] \leftarrow |\mu_i^y - \mathbf{y}|$
$\mathbf{idx} \leftarrow argsort(\mathbf{d})$ % *ascending order*
$\mathbf{X} = zeros(n, nn)$, $\mathbf{P} = zeros(nn, 1)$ % *Recall that* $\mathbf{x} \in \mathbb{R}^n$
for i in $range(nn)$ **do**
  $\mu_i \leftarrow \mu_{idx[i]}$, $\Sigma_i \leftarrow \Sigma_{idx[i]}$ $\pi_i \leftarrow \pi_{idx[i]}$
  $\mathbf{X}[:, i] = \mu^x + \Sigma^{xy}(\Sigma^y)^{-1}(y - \mu^y)$
  $\hat{\mathbf{z}}_i = [\mathbf{X}[:, i], \mathbf{y}]^T$
  $\mathbf{P}[i] = \pi_i \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} e^{-\frac{1}{2}\left((\hat{\mathbf{z}}_i - \mu_i)^T \Sigma_i^{-1} (\hat{z}_i - \mu_i)\right)}$
return $\mathbf{X}[:, argmin(\mathbf{P})]$

It is worth mentioning that, in order to minimize computation time to obtain only $\mathbf{x}$, the regression can be restricted to the k-nearest Gaussian components to $\mathbf{y}$ according to their mean μ_j^y. As observed in Algorithm 5, the regression mechanism considers nn nearest neighbors. Finally, depending on how the partitions are defined in Eqs. (3.6)-(3.7) the mechanism can be used either, for inferring $\mathbf{x}$ from $\mathbf{y}$ or for predicting $\mathbf{y}$ from $\mathbf{x}$.

3.6 Incremental Learning Example

In this section, we present a simple example to illustrate the growth of a GMM using our proposed incremental learning algorithm. We consider data batches randomly generated from 2-dimensional Gaussian distributions. Those data batches arrive at three different times and are summarized in Table 3.1. Figures 3.1-3.2 show the training steps using the generative method to train GMM, and Figures 3.3-3.4 show the results using our incremental method of training based on growing GMM.

TABLE 3.1: Training data for incremental learning. The number of samples considered per Gaussian distribution is 100.

Training Step	Mean	Covariances
$t-2$	$\mu_1 = [0, 0]^T$	$\Sigma_1 = [[0, -0.1]; [1.7, 0.4]]$
	$\mu_2 = [-6, 3]^T$	$\Sigma_2 = 0.7 * \Sigma_1$
$t-1$	$\mu_3 = [-5, 4]^T$	$\Sigma_3 = 0.5 * \Sigma_1$
	$\mu_4 = [1, 1]^T$	$\Sigma_4 = [[0.8, 0.2]; [0.1, -0.2]]$
	$\mu_5 = [4, 4]^T$	$\Sigma_5 = [[0.05, -0.05]; [0.5, 0.4]]$
	$\mu_6 = [-1, 1]^T$	$\Sigma_6 = [[-0.4, 0.5]; [-0.05, -0.05]]$
t	$\mu_7 = [0, 0]^T$	$\Sigma_7 = 0.5 * \Sigma_1$
	$\mu_8 = [-5, 4]^T$	$\Sigma_8 = 0.1 * [[1, 0]; [0, 1]]$
	$\mu_9 = [-2, 1]^T$	$\Sigma_9 = 0.1 * [[1, 0]; [0, 1]]$
	$\mu_{10} = [-6, 0]^T$	$\Sigma_{10} = 0.4 * [[1, 0]; [0, 1]]$
	$\mu_{11} = [4, 5]^T$	$\Sigma_{11} = 0.4 * [[1, 0]; [0, 1]]$

3.6.1 Generative Method for Gaussian Mixure Models

Figure 3.1 shows the training steps $t-2$ and $t-1$, whereas Figure 3.2 shows training step at t. The model parameters considered in the scenario running Algorithm 2 are $K = 5$, $\alpha = 0.05$, and $Tol = 0.001$. In Figure 3.1 (left), it is observed that at time $t-2$, the 5 components of the model M_{t-2} are fitted to the data batch $\mathbf{Z}_{t-2}$. Then, at time $t-1$ a new data batch $\mathbf{Z}_{t-1}$ is available as showed in Figure 3.1 (center). At this point, as indicated in line 5 of Algorithm 2, model M_{t-2} is sampled to generate a data batch $\mathbf{Z}'_{t-1}$ that represents the previous knowledge embedded into the model. With the concatenated data $\mathbf{Z}'_{t-1}, \mathbf{Z}_{t-1}$ the parameters are updated to become the new model M_{t-1}, shown in Figure 3.1 (right).

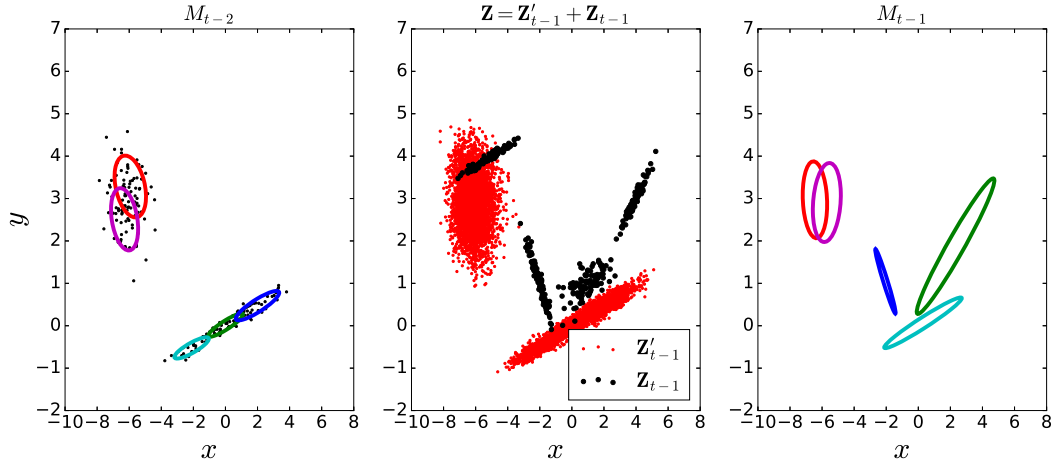

FIGURE 3.1: Incremental learning of a GMM using the generative method. (Left) At $t-2$ EM-algorithm is used to initialize the model. (Center) A data batch generated with the model obtained at $t-2$ is concatenated to new a new incoming data batch. (Right) The model parameters are updated using the EM-algorithm.

A second step of incremental learning is observed in Figure 3.2 after a new data batch $\mathbf{Z}_t$ arrives. First, model M_{t-1} is sampled to generate a data batch $\mathbf{Z}'_t$, the generated data is concatenated with the new incoming data $\mathbf{Z}'_t, \mathbf{Z}_t$, the concatenated data is used to update the parameters of the model to define the new model M_t, shown in Figure 3.2 (right).

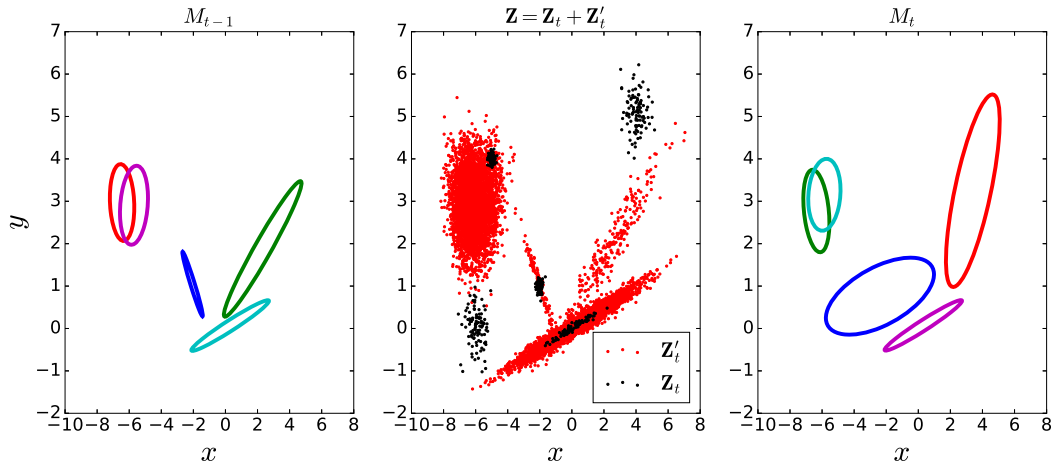

FIGURE 3.2: Incremental learning of a GMM using the generative method. (Left) Initial model. (Center) A data batch generated with the model obtained at $t-1$ is concatenated to new data batch. (Right) The model parameters are updated using the EM-algorithm.

3.6.2 Incremental Gaussian Mixture Models

Figure 3.3 shows training steps at $t-2$ and $t-1$, whereas Figure 3.4 shows training step at t. The parameters considered to obtain that figure are $K_{min} = 2$, $K_{max} = 5$, $\Delta K_{max} = 5$ and $\alpha = 0.05$.

In Figure 3.3 (Left), it is observed that at time $t-2$, the model M_{t-2} obtained with the data batch $\mathbf{Z}_{t-2}$ is a mixture with two components. As it is the first step, the model is the result of the pure EM-algorithm choosing the number of components which maximizes the BIC as indicated in line 2 of Algorithm 3. Then, at time $t-1$ a new data batch $\mathbf{Z}_{t-1}$ is available as showed in Figure 3.3 (center). A GMM M_{new}, is trained and selected according to line 4 of Algorithm 3. The next step is to merge the Gaussian components of M_{t-2} and M_{new} into M_{t-1}, in this process the total number of components should be kept lower or equal to K_{max}. For instance, in the figure, the only components which are merged are those colored in red.

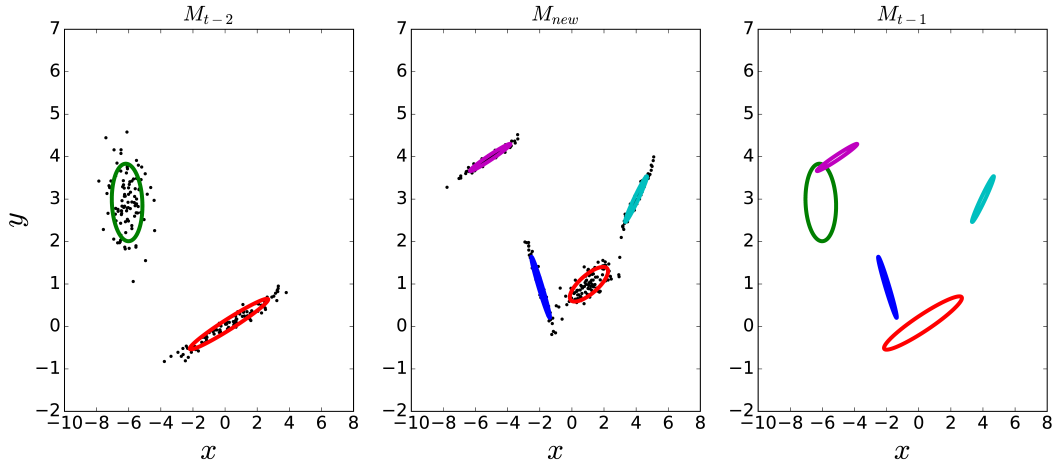

FIGURE 3.3: Incremental learning of a growing GMM. (Left) The EM-algorithm with the BIC criteria to obtain M_{t-2}. (Center) A new data batch arrives and a model M_{new} is fitted to the new data using the EM-algorithm and the BIC criteria. (Right) M_{t-2} and M_{new} are merge to obtain a new model M_{t_1}

A second step of incremental learning is observed in Figure 3.4. After a new data batch $\mathbf{Z}_t$ arrives, a GMM M_{new} is fitted to the new data using the EM-algorithm and the BIC criteria. In this case as the number of components of the model has already reached its maximum ($K_{max} = 5$), all the components in M_{new} are merged to the components in M_{t-1}. Figure 3.4 indicates with distinctive colors which components in M_{new} and M_{t-1} are merged to obtain the new M_t.

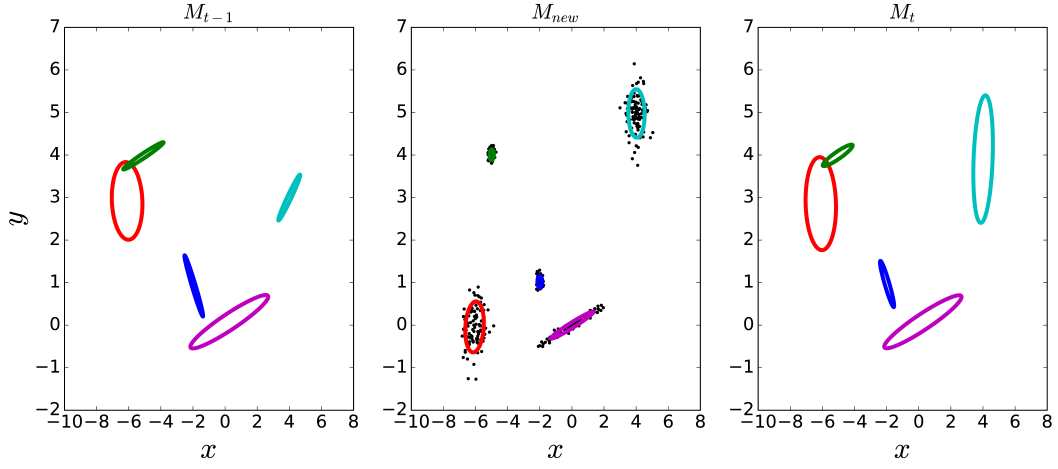

FIGURE 3.4: Incremental learning of a growing GMM. (Left) Initial model M_{t-1}. (Center) A new data batch arrives and a model M_{new} is fitted to the new data using the EM-algorithm and the BIC criteria. (Right) M_{t-1} and M_{new} are merge to obtain a new model M_t.

Figure 3.4 contains relevant information to understand what happens at the incremental learning level mechanism. Consider for instance the green components, the mechanism

chooses to mix these two components based on the KLD_s measure. However, the result in M_t is very similar to the component in M_{t-1}. This result relates to the prior weight π_i of each Gaussian; the prior weights in M_{new} are scaled by the forgetting factor. Thus, when f_1 and f_2 are computed in line 1 of Algorithm 4, the component that is already in the model is considered more relevant, and it is slightly modified by the new data. It means that it is imperative to choose a good value for the forgetting factor α which not necessarily must be constant. An adequate value for α will depend on the system to be modeled, the size of the incoming data batches and the mechanisms used to draw those data batches. For example, in former works (Acevedo-Valle et al., 2015, 2018) we adopted an active learning architecture inspired in Gottlieb et al. (2013) and Moulin-Frier et al. (2013) to draw data batches in order to maximize a measure of learning rate. In general, choosing K_{min}, K_{max} and ΔK_{max} will depend on the complexity of the system to be modeled.

3.7 Regression Problem Example

In the previous section, we have observed the mechanisms in which the studied learning approaches work. From Figure 3.2 and Figure 3.4, it is observed that the same sequence of data produced different models. It is not the objective of this work to evaluate which of the final models is better regarding the final configuration of the model but to determine which modeling approach is better suitable to solve the regression problem.

The aim of this section is to asses the performance of the learning mechanisms presented in this chapter to solve the regression problem. We proposed to solve the problem of finding the inverse model $\mathbf{x} = f^{-1}(\mathbf{y_g})$ for a simple toy example $\mathbf{y} = f(\mathbf{x})$. Here, $\mathbf{y_g}$ represents the desired output of the system, which is represented as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = f(\mathbf{x}) = \begin{pmatrix} x_1 \\ (x_2 - 3)^2 \end{pmatrix} \quad \text{and} \quad c = \begin{cases} 1 & \text{if } \mathbf{y} \in \text{constraints} \\ 0 & \text{elsewhere} \end{cases} \tag{3.10}$$

where y_i are the components of the output space, x_i the components of the input space and c is a signal indicating whether constraints are violated. As it is observed in Figure 3.5, the output-space projection is a parabolic shaped region where the red regions represent constraints (in the Figure 3.5, $y_i = s_i$).

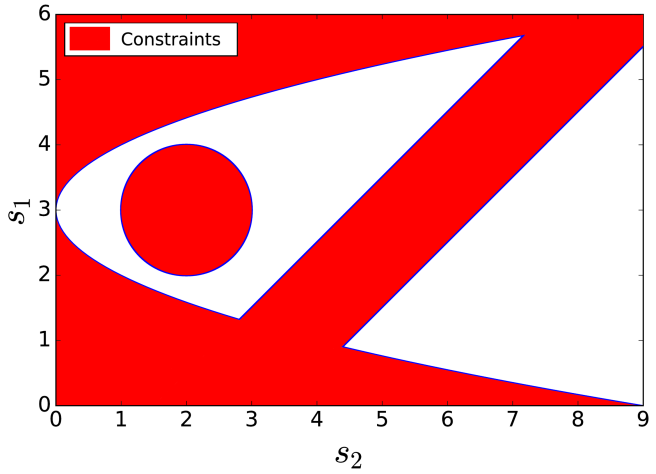

FIGURE 3.5: Constrained Parabolic Shaped Region System.

When constraints are violated, variable $\mathbf{y}$ takes the closest value $(y_1, y_2)^T$ on the valid region. Both input components are constrained to the interval $[0, 6]$, whilst output dimensions are constrained to the white region and its blue borders in Figure 3.5. Thus, given the definition of the system, it is non-linear, constrained, and redundant. Therefore, the system becomes interesting to study in a simple-fashion manner the validity of our approach before applying it to more complex sensorimotor systems which are also constrained, with more dimensions, and more abrupt nonlinearities.

Three GMMs for each incremental learning mechanism are obtained in order to show some examples with the regression mechanism for the parabolically shaped region system. Each model is trained using one of three different datasets $\mathbf{Z}$ obtained randomly using different random seeds. Each dataset contains 600 samples of couples $(\mathbf{x}, \mathbf{y})$, which were generated using 600 random inputs $\mathbf{x}$ from the allowed input space of the toy example and obtaining its corresponding output $\mathbf{y}$. In the following, we present the results for each of the learning approaches.

3.7.1 Generative Method for Gaussian Mixture Models

First of all, the set of parameters chosen to generate the models are $K = 20$, $Tol = 0.001$ and α starting with a value of 0.2 and evolving logarithmically down to 0.01 along each of the considered scenarios. The models are trained with mini-batches $\mathbf{Z}_i$ of data with 20 couples $(\mathbf{x}, \mathbf{y})$, as it is the minimum required to train a GMM with 20 components. Thus, each model

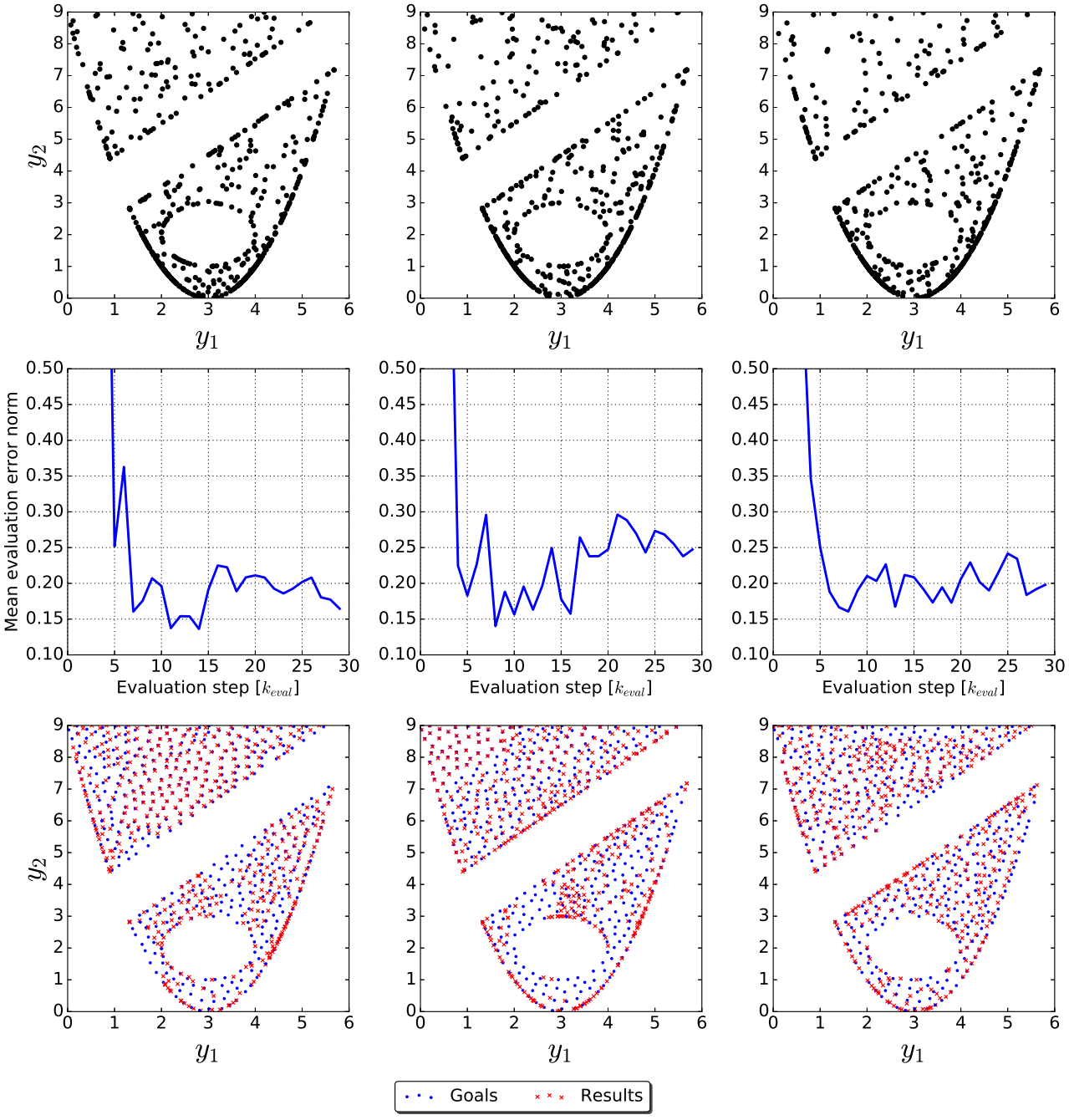

FIGURE 3.6: Inference problem results using the generative method to train GMMs incrementally. The columns correspond to experiments using different random seeds. The samples in the first row are the training samples used during all the training steps. The second row corresponds to the mean evaluation error norm at each training step. The third row corresponds to the evaluation results after the last training step.

is trained with 30 data batches. After each training step, the models are evaluated against a dataset of $N = 441$ samples uniformly distributed over the whole allowed output space $\mathcal{Y}$ of the toy example.

In Figure 3.6, the results of the simulations to learn the parabolically shaped system are shown. Each column corresponds to a simulation with each of the three different random datasets. Whereas the first row corresponds to the accumulated random data points used to train the models, the second row corresponds to the mean evaluation error norm e_{mean} against the evaluation dataset after each training step. Measure e_{mean} is defined as

$$e_{mean} = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{y_{g,i}} - \mathbf{y_i}| \tag{3.11}$$

where $\mathbf{y_{g,i}}$ are the goals in the evaluation dataset and $\mathbf{y_i}$ are the outcomes when attempting to reach those goals inferring the input $\mathbf{x_i}$ with the available GMM. Finally, the third row corresponds to the output projection of the system after a final evaluation. The blue points represent each of the output goals in the evaluation dataset. The small red crosses are actual reached output configurations.

3.7.2 Incremental Gaussian Mixture Models

The set of parameters chosen to generate the models are $K_{min} = 3$, $K_{max} = 20$, and $\Delta K_{max} = 7$. The starting value for α is 0.2 and evolves logarithmically down to 0.01 along each of the considered scenarios. Unlike the results with the generative method, here models are trained with data mini-batches $\mathbf{Z}_i$ with 15 couples $(\mathbf{x}, \mathbf{y})$, as we do not have restrictions in the minimum of data points as in the generative method case. Thus, each model is trained with 40 data batches. After each training step, the models are evaluated against the same dataset of $N = 441$ samples as in the previous case.

In Figure 3.7, the results of the simulations to learn the parabolically shaped system with our approach are shown. As in Figure 3.6, each column corresponds to a simulation with each of the three different random seeds. The first row corresponds to the accumulated random data points used to train the models, which are the same as in Figure 3.6 in order to keep the results comparable between both approaches. The second row corresponds to the mean evaluation error norm e_{mean}, defined in Eq. (3.11), obtained by evaluating the model after each training step. Finally, the third row corresponds to the output projection of the system after the final evaluation. The blue points represent each of the output goals in the evaluation dataset. The small red crosses are the reached output configurations.

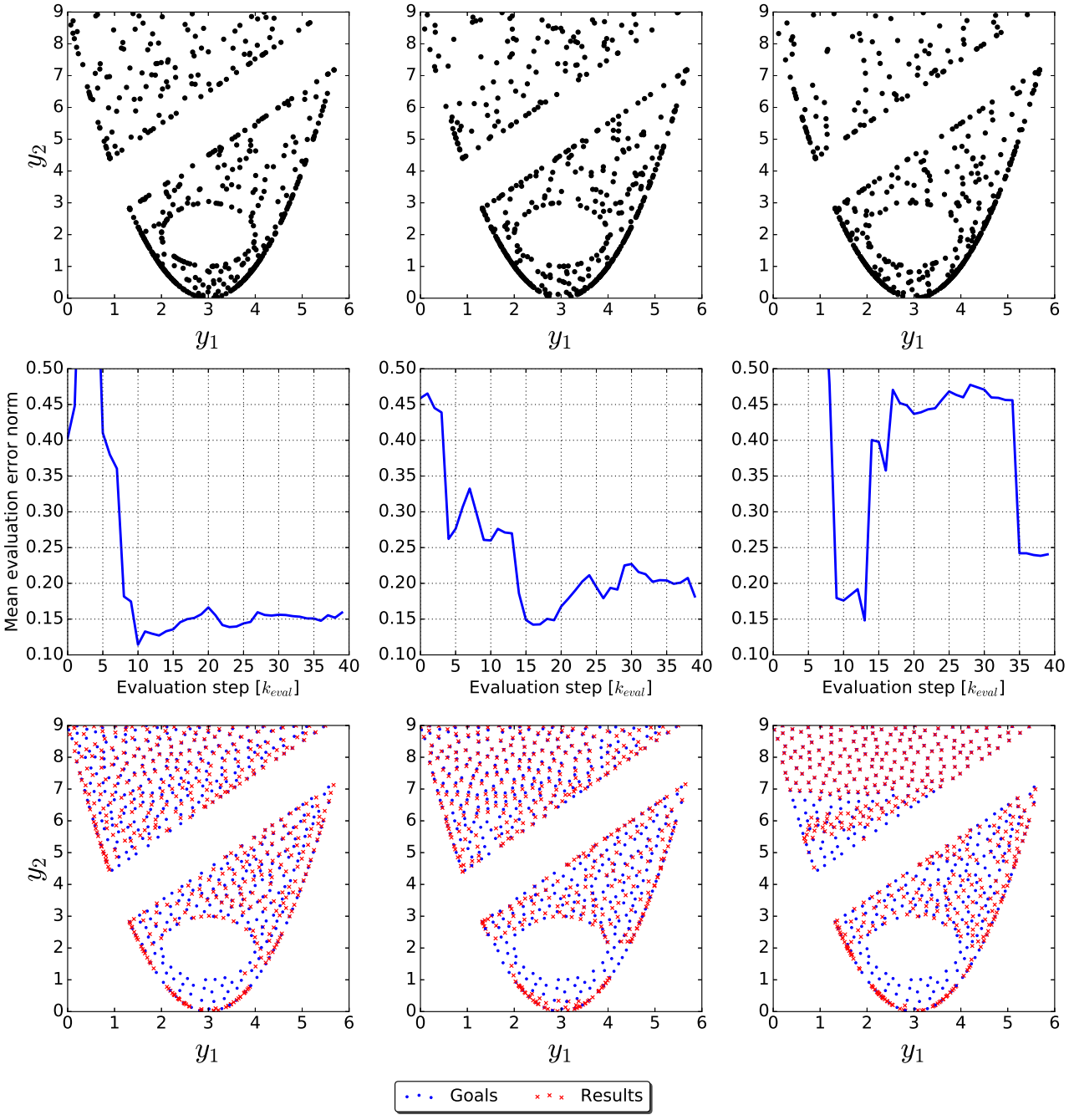

FIGURE 3.7: Inference problem results using the growing GMM approach. The columns correspond to experiments using different random seeds. The first row contains the training samples used during all the training steps. The second row corresponds to the mean evaluation error norm at each training step. The third row corresponds to the evaluation results after the last training step.

3.7.3 A Final Comparison

In the third row of Figures 3.6-3.7, and despite the small number of training samples considered (600 samples), we observe a good performance solving the regression problem with both

modeling approaches. Moreover, the training samples are not uniformly distributed along the output space due to their random source. From the last row, it is also obvious that the learning system struggles to fit the model to the system around the constrained circle. We refer this observation to the presence of many constraints in the neighborhood as well as the non-linear nature of the system. As it is hard to draw any conclusion from Figure 3.6-3.7 leading to claim that one of the learning approaches is better, in this section we propose another test.

To achieve an informed comparison between learning approaches, we ran a final experiment considering a larger number of simulations per each learning scenario. Following the same structure of experiments in Sections 3.7.1-3.7.2, fifty GMM for each incremental learning mechanism are obtained with fifty different randomly generated datasets $\mathbf{Z}$, each dataset contains 600 samples of couples $(\mathbf{x}, \mathbf{y})$. Datasets are generated using 600 random inputs $\mathbf{x}$ from the allowed input space of the toy example and obtaining its corresponding output $\mathbf{y}$. The same parameters for the models and data batches' size than experiments in Sections 3.7.1-3.7.2 are kept.

In order to compare the results, the obtained evaluations after each training step are averaged for each for the learning approaches in order to obtain an average mean evaluation error vector e_{av} for each approach. We define e_{av} as

$$e_{av}(k) = \frac{1}{n_{sim}} \sum_{i=0}^{n_{sim}} \left[e_{mean,k,i} \right], \tag{3.12}$$

where n_{sim} is the number of simulations ran per each learning approach, k is the current evaluation step, and $e_{mean,k,i}$ is the mean evaluation error for the k-th evaluation step when simulating with the i-th random seed.

As it is observed in Figure 3.8, the growing Gaussian approach overperforms the generative method through the whole training steps. It is only at the beginning when the generative method is trained with 20 samples, and the growing Gaussian method is trained with 15 samples that the latter method overperforms our approach. However, the difference in performance is not significant. The best performance of each method is that shown in Figure 3.8. The difference considering those results is around 10%. Finally, it is also notorious that the convergence to the minimum achieved seems slightly smoother for our approach compared to the generative method.

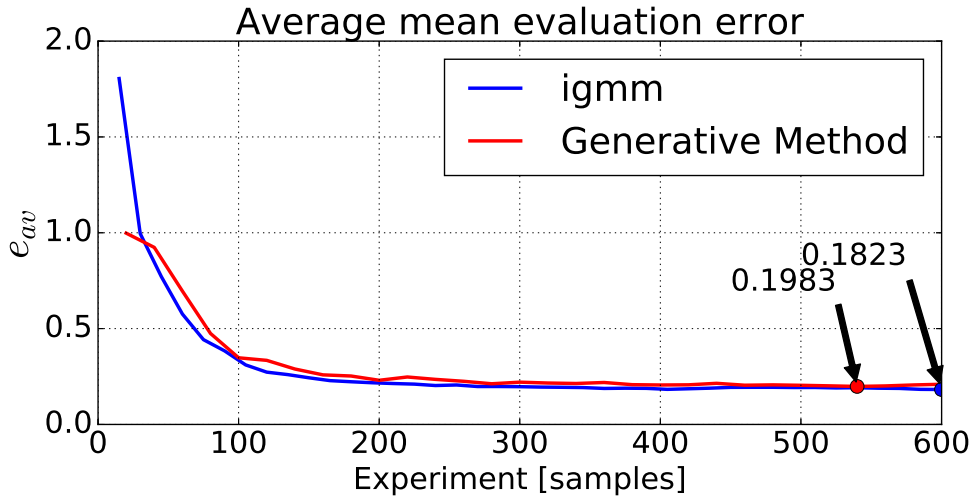

FIGURE 3.8: Comparison between learning approaches for GMMs.

3.8 Discussion

We have introduced a novel approach (*igmm*) to solve the regression problem for multivariate input-output systems using Gaussian Mixture Models (GMMs). Moreover, we compared the proposed approach with the *generative method* from Calinon (2009). The illustrative examples show the suitability of both approaches to efficiently learn a probabilistic model of the system and update the model with new incoming data batches without the need of keeping in memory already learned data. On the other hand, we introduced a simple example to demonstrate with good results the applicability of the approaches to solving the inference of input commands from desired output goals. The learning approaches are suitable to solve the prediction problem, that is they can be used to predict output results from input commands.

When learning approaches are compared, we emphasize that it cannot be argued that any of them is the best. Choosing between each of them will depend on the learning problem at hand. Through the development of our works (Acevedo-Valle et al., 2015, Acevedo-Valle et al., 2017a,b, Acevedo-Valle et al., 2018), we have observed that for the case of more complicated systems, with higher dimensionality, as the vocal tract studied in the following chapters, the approach *igmm* overperforms the results obtained with the generative approach.

Chapter 4

Motor Constraint Awareness in Sensorimotor Exploration

"Touch is not optional for human development... From consumer choice to sexual intercourse, from tool use to chronic pain to the process of healing, the genes, cells, and neural circuits involved in the sense of touch have been crucial to creating our unique human experience."

— David J. Linden, *Touch*

In Chapter 2, the role of intrinsic motivations during sensorimotor exploratory behaviors was widely discussed. To this day, there is a large number of works aimed at studying sensorimotor exploration. It is not the same case regarding works studying how motor and perceptual limitations affect cognitive development during sensorimotor learning using artificial agents, for which the number of studies is somewhat limited. However, in recent years, the number of published works including mechanisms to deal with motor information and constraints has increased.

On the other hand, Chapter 2 provided a discussion on the elements that foster prelinguistic development toward speech emergence. If works studying the role of constraints during early development using artificial agents is sparse, the available literature regarding the role of constraints in artificial vocal development is rather nonexistent.

In this chapter, one of the main contributions of this dissertation is presented. This contribution is aimed at studying the integration of constraint awareness into intrinsically motivated sensorimotor exploration, especially applied to vocal development. The advances presented in this chapter resulted in two publications, Acevedo-Valle et al. (2015, 2018).

This chapter is organized as follows. Section 4.1 is devoted to providing biological perspectives on somesthetic senses, especially haptic perception and nociception. This section also presents studies on how somesthetic senses influences vocal development. Section 4.2 introduces some works that have borrowed somesthetic mechanisms to integrate them into autonomous systems design. Therein, works related to artificial language development and artificial speech mechanisms are emphasized.

Afterwards, next sections are devoted to present the development and results of the proposed approach. Hence, Section 4.3 explains the artificial cognitive architecture that actively takes into account motor constraints during sensorimotor exploration. Section 4.4 provides an experimental setup aimed at testing the proposed cognitive architecture. Finally, Section 4.5 and Section 4.6 present the experimental results and a final discussion, respectively.

4.1 The Role of Somesthetic Modalities in Sensorimotor Development

The sense of touch is often used and interpreted as a unique sensor modality. However, in fact it consists of a broader range of perceptual modalities. These perceptual modalities may be defined as somesthetic senses, including touch, thermoception, nociception, and other bodily sensibilities (Hollins, 2010). In general, this family of perceptual modalities responds to mechanical, thermal and chemical energy enabling perception of vibration, texture, location and movement, temperature and pain. For the purposes of this work, haptic modalities (touch and proprioception) and nociception are of special interest. In the following, the modalities of interest are defined:

Touch relies on different kinds of low-threshold mechanoreceptors. Their location and density vary with respect to the skin tissue where they are located. For example, the density of receptors in the fingerprints is higher than most of the other parts of the

integument (Beckstead, 1996). Touch provides information on light contacts, slip, texture, flutter, vibration, pressure, stimulus shape and stretch (Prescott and Ratté, 2017). It is one of the richest sources of information from the environment that humans, and animals in general, possess.

Proprioception is sometimes defined as a specialized variation of touch. It relies on receptors located in muscles, tendons, and joints. It provides conscious or unconscious awareness of joint position (Norris, 2011). Besides joint position, proprioception is sometimes used to refer to the sense of joint motion, which is also known as kinesthesia (Prescott and Ratté, 2017). Proprioception is crucial for balance and motor control as it endows agents with a sense of their movements. Moreover, it provides body-centered cues that generate a spatial reference additional to that provided by vision (Millar, 2005).

Nociception depends on nociceptors (from Latin word *noci*="hurt"), which are nerve cell endings responsible for producing signals that travel to the brain in the presence of chemical levels, temperature or mechanical factors that might be harmful to the body. In the brain, those signals are perceived as pain, which is a psychological experience. The receptive fields in the brain for pain are large, presumably because the detection of pain is more important than its precise localization (Purves et al., 2001). Pain is such a salient modality that even captures attention. When attention is captured, pain awareness appropriates processing resources that would otherwise contribute to performance on other competing tasks (Hollins, 2010).

Besides the relevance of intrinsic motivations to sensorimotor exploration, Acevedo-Valle et al. (2015, 2018) discussed the relevance that touch, proprioception, and nociception have during early development. Those works emphasized the relevance that somesthetic senses have during the emergence and progress of rhythmic behaviors necessary to foster sensorimotor control learning. Experimental evidence suggests that somesthetic information is relevant for development in different ways. For example, as a perceptual mechanism to explore the world, during early development (from 5 months of age) infants interact mostly with objects through mouthing (Klatzky et al., 2005). Mouthing is defined in Fagan and Iverson (2007) as contact of an object with the mouth, lips, or tongue. Another example of somesthetic relevance is provided by Lewis et al. (2008, chapter 22), therein the effects that tactual information has to infant's responsiveness is stressed. For instance, when touch is provided

infants increase smiling, vocalizing, and gazing. In general, touch can be used by caregivers as a mechanism to present affective information in order to modify infant's behavior.

Somesthetic modalities, as mentioned in Corbetta et al. (2014), are essential for an agent to learn how to drive its movements to reach intended body states. The evidence found by Corbetta et al. (2014) states that during the emergence of reaching, as a product of a deeply embodied process, infants first learn how to direct their movements in space using proprioceptive and haptic feedback. During sensorimotor learning of proprioception, nociception, and tactile modalities, agents must first discover their motor limitations.

Recent works regarding somesthesis, have been addressed to discover new principles of these modalities. However, other studies have been focused on the cognitive level and emphasized similarities and interactions across modalities (Hollins, 2010). This idea makes sense due to the structure of neural pathways, as somesthetic data of many submodalities goes along parallel central pathways. Later, that information is used for a variety of tasks with different complexity and computational purposes, e.g., identification of objects by feel (Beckstead, 1996).

Millar (2005) introduced a roadmap of the interaction between somesthetic modalities with vision during early development and its neural implications. She emphasized the relevance of an intertwined network of haptic models. Furthermore, she provided a study about mechanisms and implications of collaboration between a somesthetic network with visual information through the course of development. In another work (Klatzky et al., 2005), it was observed that visual and haptic information is used by preverbal infants to decide if an object can be reached or grasped, and also to decide whether a surface will support locomotion or not.

Somesthetic Modalities in Prelingusitic Vocal Development

Some insights regarding the relevance of somesthetic modalities have been presented in this chapter. Hereafter, the reader is introduced into some works that provide discussions and strong evidence on the relevance of somesthetic senses to vocal development and language emergence.

During early vocal development, when infants are engaged with rhythmic behaviors, it is not adventurous to say that somesthetic modalities must play a key role according to evidence. It is enough thinking about somesthetic senses as perceptual modalities that are a rich source of information available to infants, even long before they can control phonation. For example, think about mouthing behaviors mentioned beforehand.

In general, experimental results suggest that somesthetic inputs related to movement play an important role in speech maintenance (Galantucci et al., 2006, Nasir and Ostry, 2008, Tremblay et al., 2003). Furthermore, the fact that canonical babbling emerges, with some variations, in deaf infants suggests that somesthetic feedback plays a relevant role during prelinguistic vocal development. For deaf infants, tactile and proprioceptive information help to find regularities between motor actions and sensory results (Iyer and Oller, 2008).

In a different experiment, Ito et al. (2009) used a robotic device able to generate patterns of facial skin deformation related to specific speech productions. Results showed that when the facial skin is stretched whilst subjects are listening to words, the sounds they hear are altered. Therefore, experimental evidence strongly suggests a linkage between somesthetic information and speech perception, at least as a mechanism to reinforce auditory speech perception.

Another interesting argument, *perceiving speech is perceiving gestures*, comes from Chapter 2, where we introduced the Motor Theory of Speech Perception and the Perception for Action Control Theory. As mentioned by Schwartz et al. (2012), and supported by ideas from Galantucci et al. (2006), the listener who has some motor competence, knowing something about speech production, exploits this knowledge to decipher the acoustic input and access the functional unit more directly related with motor commands. Moreover, audiovisual and audio-haptic interactions in speech perception are claimed to be related to the knowledge of the listener on multisensory coherence of a speech gesture. In this context, a hypothesis suggests that perception involves a procedural knowledge of action (Schwartz et al., 2012), and proprioceptive knowledge is an interesting candidate to enclose action information.

On the one hand, Galantucci et al. (2006) reports results suggesting that knowledge of anatomical constraints affects what people perceive. On the other hand, there is the claim that motor competence is exploited during perception mentioned beforehand. Merging both hypotheses, if the perceiver knows that a visual pattern can only correspond to a specific

motor behavior, and if humans are genetically equipped to produce that behavior, one can speculate that specialized motor competence is called upon in the perception of the pattern. Finally, we remark that evidence also suggests that perceptual performance is enhanced if the perceiver produces the same movements which are being perceived, probably because the maximal amount of motor competence is available to support perception.

Assembling the series of ideas mentioned above, there is a reason to believe that perception is particularly attuned to the general anatomical and dynamical constraints on natural movements, as well as to the specific subtleties of individual movements (Galantucci et al., 2006, Liberman and Whalen, 2000).

From the evidence provided in Schwartz et al. (2012), it is clear that auditory, phonetic and lexical mechanisms are involved in speech scene analysis. It is important also the role of articulatory coherence. The fact that articulatory constraints may act on the emergence and stabilization of verbal transitions strongly suggests that they partly rely on motor information. Evidence demonstrates that articulatory based representation play a crucial part in the endogenously driven emergence and stabilization of auditory speech percepts during a verbal transformation task. In general, experiments show that speech scene and analysis process appears to be driven by both perceptual and motor coherence.

There is evidence that perceiving speech involves neural activity of the motor system. Studies demonstrated the activation of speech-related muscles during the perception of speech. For instance, Liberman and Mattingly (1985) proposed that infants mimic the speech they hear and that leads to associations between articulation and its sensory consequences. Perhaps through acquired similarity, whereby associating different acoustic signals for the syllables to the same response makes the syllable-initial consonants alike. Another process, of acquired articulatory distinctiveness, may explain how similar acoustic signals with different underlying articulations come to sound distinct (Galantucci et al., 2006). For individuals perceiving a speech signal, who are capable of producing a similar signal, the consequence of the memory representation established by the mimicry principle is that articulatory movements and their sensory effects mediate between the acoustic stimulus and the event we call perception (Galantucci et al., 2006, Liberman and Mattingly, 1985).

4.2 Somesthetic Modalities in Artificial Cognition

From the artificial cognition perspective, arguments regarding the relevance of somesthetic senses and multimodal perception to development can be traced to relatively old works. For example, Sandini et al. (1997) already mentioned the relevance of considering the simultaneous exploitation of vision, touch and motor schemes in order to solve complex tasks, e.g., grasping in early developmental stages.

In the literature, there are a few works that use somesthetic systems in the design of robots or other artificial agents. However, roboticist agree upon the interest on proprioception, nociception and touch modalities as tools to endow artificial agents with new bioinspired mechanisms that foster learning and developmental performance, and in consequence the emergence of intelligent behaviors (Luo et al., 2016, Navarro-Guerrero et al., 2017b, Schillaci et al., 2016).

Schillaci et al. (2016) mentioned proprioception as an important modality to learn motor constraints based on the evidence provided by Corbetta et al. (2014). In another example, and without mentioning any somesthetic modality explicitly, Rayyes et al. (2017) proposed a scheme to learn inverse static mappings for gravitational compensation forces in robotic arms. Based on the argument that exploratory noise could lead to inadmissible motor configurations, Rayyes and colleagues propose to learn or estimate admissible motor actions. Thus, in order to avoid reaching invalid configurations, they propose to avoid any limit violations during bootstrapping and then to learn the embodiment's constraints. To learn those constraints, they use a directed sampled architecture employing a modified goal babbling scheme, similar to that proposed in Rolf (2013). Within the framework of somesthetic modalities, what Rayyes and colleagues do is similar to the mechanisms that infants use for motor learning using proprioception as mentioned by Corbetta et al. (2014).

Luo et al. (2016) also borrowed ideas from Corbetta et al. (2014), in this case, the authors proposed a system that endows a robot with the tools to develop its reaching ability based on infant's development. They divide the developmental processes into five prespecified stages. Sequentially in those stages the robot learns: (1) sense of joint position, (2) sense of arm position and orientation, (3) learn forward and inverse motor models, (4) in this step the models from (1), (2) and (3) are integrated to model proprioception, and (5) learn the mapping of the sight of an object onto the embodied sense of space to determine grasping

commands. The results show that the robot was capable of developing its reaching ability following a similar path to that described by Corbetta et al. (2014).

Navarro-Guerrero et al. (2017a,b) found that nociceptive mechanisms produce a boost in performance for some tasks, e.g., inverse kinematic learning. Navarro and colleagues also provide an interesting discussion regarding the role that nociception and punishment signals have played in robotics. In Navarro-Guerrero et al. (2017b), nociception is used as an additional state, whereas punishment is used as a negative reinforcement signal. Their experiments showed that nociception improved the results achieved when using TD-learning algorithms. In their discussion, the authors argued that the main reason for the improvement may be the fact that the extended state also expanded the differences between input vectors. Despite just being optimized for position error, this architecture improved the results regarding positioning error, the potential for damage, and positioning speed.

Somasthetic Modalities in Artificial Speech Studies

Summarizing the ideas discussed so far, we conclude that it is not trivial to argue that any attempt to study early vocal development using artificial agents requires, at least at some point, to consider that somesthetic senses may play a crucial role through the developmental trajectory of speech during infancy.

Among the efforts to mimic the acquisition of speech as it occurs in infants, one interesting project is the Diva model developed by Guenther and colleagues (Guenther, 2006, Guenther et al., 2006). The model, which is inspired by neurophysiological evidence, includes the premotor, motor, auditory and somatosensory cortical areas in the cognition level, and a simulated ear-vocal tract system as the embodiment of the agent. Guenther and colleagues integrated the somatosensory modality effectively, based on proprioception and tactual information, into the processes for acquisition and production of speech. However, the somesthetic modalities were not used as an element to integrate motor constraint awareness. Instead, they were used as a part of the central sensorimotor system (extended sensory state) to produce learned speech gestures regardless of perceptuo-motor coherence. The mechanism is comparable to the that from Navarro-Guerrero et al. (2017b), at least in the sense that somesthetic information is used to extend the perceptual state vector.

A somesthetic modality based on haptic information, defined as the somatosensory system, is taken into account in Howard and Messum (2011). There, the authors use tactile information in their architecture for speech acquisition. The integration of the somesthetic modality is done similarly to Guenther (2006), under the assumption that from a motor control perspective an infant learns to correlate certain activation of the muscle related to the vocal tract and the breathing apparatus to somatosensory and auditory sensory consequences.

How somesthetic information affects artificial ear-vocal tract exploration is an open question that was not studied by Moulin-Frier and colleagues (Moulin-Frier and Oudeyer, 2013b, Moulin-Frier et al., 2013). Therefore, we proposed some modifications to the architecture and experiments proposed in Moulin-Frier et al. (2013) to include somesthetic information. This information would endow artificial vocal learners with the awareness of its physical constraints. Hence, agents would avoid executing vocalizations that could lead to undesired vocal tract configurations.

In the following an architecture proposed in our works Acevedo-Valle et al. (2015, 2018) is presented, and results using such an architecture are discussed. The proposed architecture accounts for embodied systems with motor constraints. The embodied agents must be endowed with a system that generates a nociceptive signal indicating if a motor configuration was reached.

4.3 Sensorimotor Exploration with Constraint Awareness

So far, through this chapter, we have reviewed studies regarding the relevance of somesthetic senses to development during infancy. Particular emphasis has been put on the role that these modalities have during prelinguistic vocal development. In this section, a cognitive architecture, which was first introduced in Acevedo-Valle et al. (2015, 2018) is described. The architecture integrates a simple mechanism to endow intrinsically motivated sensorimotor exploration architectures (Baranes and Oudeyer, 2013, Moulin-Frier and Oudeyer, 2013b, Oudeyer et al., 2007) with motor constraint awareness.

The first step toward fulfilling the objectives of this work was to reproduce the architecture presented in (Moulin-Frier and Oudeyer, 2013b). As mentioned beforehand, Moulin-Frier and colleagues provided an architecture based on infant development to reproduce sensorimotor

exploratory behaviors using machines to learn inverse models. Moreover, they wanted to show that this kind of architectures may explain the progress through structured developmental stages during early vocal development (Moulin-Frier et al., 2013). Despite their promising results, their work did not consider any somesthetic modality nor made any reference to the relevance those modalities may have to the emergence of stages through development. In consequence, there was an open question regarding the impact that somesthetic modalities may have over the learning performance of inverse models.

Once the architecture from Moulin-Frier and Oudeyer (2013b) was reproduced, we started building up our architecture over it. Summarizing, we included a new element, a mechanism to deal with constraints. In general, there were two main reasons to include motor awareness in the architecture shown in Figure 2.1. On the one hand, there is a need for including multimodal perception, especially somesthetic senses, in any system that attempts to mimic early human development as discussed previously. On the other hand, it is important to notice that, for the specific application to early vocal development, the vocal tract used in Moulin-Frier and Oudeyer (2013b) does not consider physical constraints. Thus when an articulatory trajectory is executed, the result may lack physical sense. As this work uses the same synthesizer, considering our implementation described in the Appendix A, it is important to consider the lack of motor constraints in this synthesizer. This concept might be extended to other embodied cognition applications, but the nuances of those applications may change the interpretation of this argument.

The Maeda-based synthesizer implementation by Guenther Lab, used widely in the literature to study speech (e.g., Acevedo-Valle et al. (2015, 2018), Forestier and Oudeyer (2017), Guenther (2006), Guenther et al. (2006), Moulin-Frier and Oudeyer (2013b)), allows executing motor commands that lead to collisions or articulatory superpositions. Those circumstances produce non-phonatory articulatory gestures as a result of blocking the air flow. Furthermore, superpositions between articulators lack physical sense (see Figure 4.1).

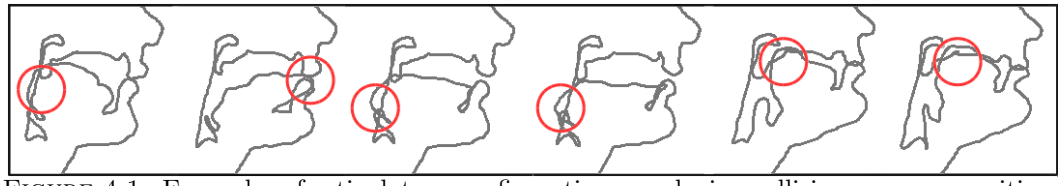

FIGURE 4.1: Examples of articulatory configurations producing collisions or superpositions in the Maeda's synthesizer (from Acevedo-Valle et al. (2015)).

In order to endow artificial agents with motor constraint awareness, Acevedo-Valle et al. (2015) established the foundations of a simplified architecture inspired by somesthetic senses to deal with physical constraints, under some assumptions regarding somesthesis. Later, in Acevedo-Valle et al. (2018) the architecture and its implementation was revisited, and initial results were extended. Finally, in Acevedo-Valle et al. (2017a), a more in-depth review and extension of the architecture was made. Therein, a new consensus of nomenclature was adopted based on the series of developmental psychology studies mentioned through this work. However, to avoid any confusion with specific somesthetic modalities, herein, the proposed system will be referred as to *somesthetic system*; alongside to its description, parallelism with individual somesthetic modalities, i.e., haptic and nociceptive, will be discussed.

In Acevedo-Valle et al. (2015), Acevedo-Valle et al. (2017a), Acevedo-Valle et al. (2018), we have argued and provided results that support the relevance of somesthetic mechanism for early development. In the following, we detail the mechanisms we have proposed to deal with motor constraints during sensorimotor exploration and how they achieve a reasonable degree of consistency with evidence from developmental psychology and neurophysiology. The mechanism was inspired mainly by the concepts of nociception, pain, and reflective behaviors.

In general, the architecture considers a nociceptive signal, when the signal is triggered means that a motor configuration might be 'harmful' or physically unreachable. This signal activates a pain-like mechanism that, as in humans and other living beings, attracts attention (Hollins, 2010), the agent is aware that something is wrong, so a cognitive structure is called to keep track of the motor command that caused that unpleasant configuration. Such a structure is a self-generated map of 'harmful' motor configurations and, later, it is used to avoid execution of unpleasant motor commands as it happens in humans. When a human feels pain then, usually, avoid to repeat the task that generated that pain.

Summarizing, we integrate a somesthetic system to the intrinsic motivations mechanism depicted in Figure 2.1. The somesthetic part consists of two main elements. First, there is a nociceptive signal indicating motor incoherence, harmful configurations, or configurations lacking physical sense. This signal is interpreted by the agent as 'pain', even though no information is provided regarding the 'pain' source, which is partially consistent with the description by Hollins (2010), who indicates that detecting pain as quickly as possible is more important than knowing exactly where it was generated. Secondly, there is a cognitive

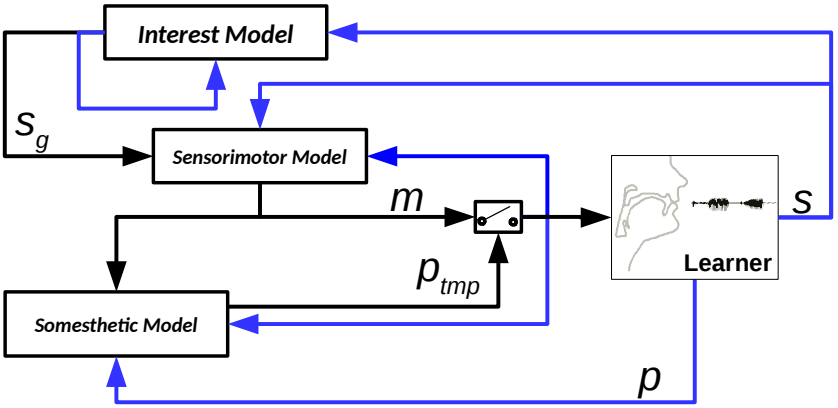

FIGURE 4.2: Exploration architecture considering constraint awareness. Black lines represent the flow of data during each action execution. Blue lines represent signals used to update the models. The simple switch indicates that the prediction made by the somesthetic model is used to accept or reject a proposed interesting goal s_g.

structure that generates a map of motor commands respect to the somesthetic signal that indicates the existence of undesired motor configurations. The generated map, known as the somesthetic model, can be used later by the agent to predict whether motor commands will lead to undesired motor configurations or not. Therefore, those predictions are considered to decide if a motor command is going to be executed or not.

In the case of the vocal tract, the architecture considers an emulated nociceptive signal. This signal is obtained from tactile information: when tactile information is incoherent, the emulated nociceptive signal is triggered. The implemented tactile system collects information from the vocal tract shape about undesired contacts and collisions inside the vocal tract, information more similar to the proprioceptive modality. Afterward, the information is encoded to a single nociceptive signal, when the signal is active, then it is interpreted as pain by the agent, emulating the role of nociceptors and pain center.

Finally, the proposed architecture is represented by the diagram shown in Figure 4.2. Different from the elements in Figure 2.1, we add a somesthetic model that maps motor commands to internal signals triggered or not by an action of the agent, e.g., a nociceptive signal in the case of the vocal tract indicating undesired articulatory configurations as the ones shown in Figure 4.1. The learner starts with no knowledge about any of the three models. First, the models are initialized in a first stage, either randomly or using any other criteria. Once they are initialized, the intrinsically motivated exploration begins. The interest model proposes a sensory goal, that goal is then passed to the sensorimotor model, the sensorimotor model then computes the motor command that, according to the current knowledge, would produce

that sensory goal. Then, the motor command goes to the somesthetic model, which determines if the motor command may produce 'pain' according to the current knowledge. If the motor command is not likely to trigger the nociceptive signal, it is accepted. Otherwise, the motor command is rejected and the interest model is forced to choose a new sensory goal that is subjected to the same information trajectory, and the processes must be repeated until a goal is accepted. Once a sensory goal is accepted, so is the motor command obtained to produce that goal, and the command is executed by the agent (indicated as the learner in the diagram). When the agent executes the motor command, the salient signals are observed obtaining a sensory outcome. Afterward, the sensory outcome is compared to the sensory goal to generate the competence c value of the experiment as an index of performance. The signals generated, described by blue arrows in the diagram, are then used to train the models. After training models, the process of choosing a new sensory goal starts again.

In the next section, we present the proposed approach to deal with constraints during intrinsically motivated sensorimotor exploration as an algorithmic architecture. The details to implement such an architecture will also be introduced.

4.4 Architecture Implementation

The cognitive architecture described in the previous section, shown in Figure 4.2, is rewritten as an algorithm in this section. Furthermore, the elements necessary to build the experimental setup used to implement the generated algorithm are described in detail. In general, the proposed implementation is valid to study any sensorimotor exploration scenario. Therefore, it is presented in such a way the reader with some experience in programming could apply the architecture to any system. Later, we describe the specific implementations to solve the sensorimotor exploration problem for the simple parabolic system presented in Chapter 3. Finally, the implementation details about applying the architecture to study early vocal development in machines are presented.

As shown in Figure 4.2, four elements are interacting inside the architecture. These elements, together, integrate the artificial agent and are described below.

Embodiment. In Chapter 2, we have widely discussed the concept of embodiment. In the case of this work, we consider an embodiment that consists of three elements. First,

a motor system that allows the agent to modify its environment. The environment is affected in such a way there is a salient signal observable by any other agent endowed with the adequate sensory system, including the agent itself. Thus, the second element is a set of physical elements that allows the agent to sense the salient consequences of its motor actions. Finally, the embodiment has a second sensory system that allows sensing, partially or completely, its physical state. When a 'harmful' state is sensed, then a nociceptive signal is triggered and perceived by the agent as 'pain'.

Sensorimotor Model: Mapping salient signals to motor actions. Through this work we consider Gaussian Mixture Models (GMMs) to model sensorimotor systems. As mentioned in Chapter 3, we adopted GMMs as a modelling approach based on the results obtained in Moulin-Frier and Oudeyer (2013a,b), Moulin-Frier et al. (2013). We do not argue that GMMs are the best option to get the most accurate models, but we do argue that they are a fast method to do experiments in order to make proofs of concept keeping a good accuracy and achieving an actual incremental learning approach. In Chapter 3, it was also indicated that two incremental training approaches for GMMs were implemented. On the one hand, one was based on the tools provided along with Calinon (2009). On the other hand, the second approach was presented in Acevedo-Valle et al. (2017b). It is important to mention that different from Acevedo-Valle et al. (2015, 2018), which considered the approach based on Calinon's work was used, in this work the presented results were obtained using the tools provided by Acevedo-Valle et al. (2017b).

Using the definitions from Section 3.5, we have an approach that solves the inference problem for incomplete data of the extended space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. For the sensorimotor model mapping the space of salient signals to the motor space, an m-dimensional motor command space $\mathcal{X} = \mathcal{M}$ is considered, with $\mathbf{m} \in \mathcal{M}$. On the other hand, a s-dimensional sensor space $\mathcal{Y} = \mathcal{S}$ of perceived salient signals, with $\mathbf{s} \in \mathcal{S}$ is defined.

A map f is assumed to exist such that $\mathbf{s} = f(\mathbf{m})$. The agent can observe $\mathbf{s} + \sigma_{\mathbf{s}}$ for any executed action $\mathbf{m}$. Thus, it is possible to find a GMM, $M_{\mathcal{SM}}$, representing the extended space $SM = \mathcal{S} \times \mathcal{M}$. This model allows us to compute the probability distribution $P(\mathbf{m}|\mathbf{s})$ applying Gaussian Mixture Regression (GMR), and later it is

possible to determine which motor command $\mathbf{m}$ is the most likely command to produce a desired sensory goal $\mathbf{s}_g$, thus solving the inverse regression problem $\mathbf{m} = f^{-1}(\mathbf{s}_g)$.

Somesthetic Model For the somesthetic model, we consider the same m-dimensional motor command space $\mathcal{M}$, with $\mathbf{m} \in \mathcal{M}$, and a new binary 'pain' output space $\mathcal{P} = \{0, 1\}$, with $p \in \mathcal{P}$. If the somesthetic system detects that a harmful body configuration has been reached, then a nociceptive signal is triggered and perceived by the agent as pain, then $p = 1$, otherwise $p = 0$. A map g is assumed to exist such that $p = g(\mathbf{m})$ and the agent can observe p for each vocal experiment. Thus, it is possible to find a GMM $M_{\mathcal{SS}}$, with $\mathcal{X} = \mathcal{M}$ and $\mathcal{Y} = \mathcal{P}$, that allows computation of the probability distribution $P(p|\mathbf{m})$ applying GMR, and determine when a motor command $\mathbf{m}$ is likely to lead to a 'painful' configuration, thus solving the prediction problem for g. Strictly speaking, considering what was mentioned in Acevedo-Valle et al. (2018), the somesthetic model is also a kind of sensorimotor model. However, it will be named along this work just as the somesthetic model to keep a simple nomenclature.

Interest Model for Auditory Goals The interest model for auditory goals is an element that endows the learner with the ability to select goals that maximize the expected competence progress in order to improve the quality of its sensorimotor model, resulting in a better control over it. Through this work we use the competence measure used in Moulin-Frier et al. (2013), an later adopted in Acevedo-Valle et al. (2015, 2018), written as

$$c = e^{-|\mathbf{s}_g - \mathbf{s}|}, \tag{4.1}$$

where $\mathbf{s}_g$ is the auditory goal and $\mathbf{s}$ is the actual auditory production after executing a motor command $\mathbf{m} \sim P(\mathbf{m}|\mathbf{s}_g)$.

In Acevedo-Valle et al. (2015), Acevedo-Valle et al. (2017b), we followed the description of the interest model proposed in Moulin-Frier et al. (2013). Hence, to construct the interest model, the auditory goal space was augmented with two extra dimensions: the competence $c \in C$ and time tag $t \in T$. The number of vocalizations N_{IM} considered to build the interest model was fixed. Then, a GMM, M_{IM} with K_{IM} components was computed from the $(s + 2)$-dimensional dataset considering the last N_{IM} sensory results of the agent's life.

Those Gaussian components in M_{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 $P(\mathcal{S})$ over the auditory space. To build $P(\mathcal{S})$, the components in M_{IM} are weighted according to their time-competence covariance magnitudes. Thus, $P(\mathcal{S})$ will prioritize goals in regions where competence is expected to increase. Finally, a sample $\mathbf{s}_g$ is drawn from $P(\mathcal{S})$ for the next vocalization experiment.

However, in this work, we introduce new results using the state-of-the-art interest models provided in the `explauto` toolbox, which in recent works have demonstrated to produce better results (Acevedo-Valle et al., 2017a).

4.4.1 Algorithm for Sensorimotor Exploration with Constraint Awareness

Algorithm 6 corresponds to the cognitive architecture in Figure 4.2. The algorithm for self-exploration with goal babbling and motor constraint awareness starts with the learner having no sensorimotor control experience.

Algorithm 6 Sensorimotor exploration with goal babbling and motor constraint awareness.

Set $\{n_e, randomseed\}$

1: Initialize $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$
2: Initialize M_{IM} and $i \leftarrow 1$
3: **while** $i \leq n_e$ **do**
4: $\quad p_{tmp} \leftarrow 1$
5: $\quad$ **while** p_{tmp} **do**
6: $\quad\quad \mathbf{s}_{g,i} \leftarrow sample\,(M_{IM})$
7: $\quad\quad \mathbf{m}_i \leftarrow M_{\mathcal{SM}}\,(\mathbf{s}_{g,i})$
8: $\quad\quad p_{tmp} \leftarrow M_{\mathcal{SS}}\,(\mathbf{m}_i)$
9: $\quad \mathbf{s}_i \leftarrow f\,(\mathbf{m}_i) + \sigma$ and $p_i \leftarrow g\,(\mathbf{m}_i)$
10: $\quad c_i \leftarrow (1 - p_i * \gamma) e^{-|\mathbf{s}_{g,i} - \mathbf{s}_i|}$
11: $\quad i \leftarrow i + 1$
12: $\quad train\_models()$

First, models $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$ are initialized in line 1 using arbitrary motor commands with small values around the neutral motor system position. In line 2, model M_{IM} is initialized using the sensory results obtained in the first line as sensory goals.

Then, in line 6 of Algorithm 6 the agent selects a sensory goal $\mathbf{s}_{g,i}$ for the next sensorimotor experiment according to the interest model M_{IM}. With $\mathbf{s}_{g,i}$, in line 7, the sensorimotor model is used to obtain the motor command $\mathbf{m}_i$ that according to the current knowledge of the agent would produce $\mathbf{s}_i = \mathbf{s}_{g,i}$.

Unlike similar architectures, in this algorithm $\mathcal{M}_{\mathcal{SS}}$ provides a nociceptive or 'pain' prediction p_{tmp} for $g(\mathbf{m}_i)$ in line 8. That prediction indicates if the selected motor command is likely to trigger the nociceptive signal, thus causing the agent a 'painful' experience. If the pain prediction indicates that the signal p will be triggered when executing $\mathbf{m}_i$, then the agent rejects the goal, the simple switch in Figure 2.1 is open, and the motor command is not executed. Afterward, the interest model proposes a new goal and the prediction process is repeated until a 'safe' goal is obtained. On the other hand, if the 'pain' prediction suggests that there is no risk when executing the motor action $\mathbf{m}_i$, then the simple switch in Figure 2.1 is closed, and the agent accepts and executes $\mathbf{m}_i$.

Next, the motor command $\mathbf{m}_i$ is executed by the motor system. Afterward, the agent observes $\mathbf{s}_i$ and p_i in line 9. In line 10, the learner evaluates the competence value c_i, which receives a penalization according to the parameter γ if the agent perceives 'pain' ($p_i = 1$). Finally, in line 12, the training function for models is called, and each model ($M_{\mathcal{SM}}$, $M_{\mathcal{SS}}$ and M_{IM}) is updated according to its parameters that will be explained in the next section.

4.4.2 Parabolic Shaped Region System Embodiment

The parabolic shaped region system was introduced in Section 3.7, the only element that is added to this system is the mechanism to produce the nociceptive signal, which consists in making the pain signal $p = 1$ if the constrained region of the system is reached and $p = 0$, otherwise. If after executing a motor action the sensory result lies in the constrained region, then it is relocated to the closest point in the allowed region. An obvious consequence of relocation is the increment of sensorimotor redundancy.

4.4.3 Ear and Vocal Tract Embodiment

In this work, we use the Maeda's synthesizer as a vocal tract (motor) system. The implementation, as mentioned in Chapter 2, is based on the implementation made for the DIVA model in Guenther et al. (2006). The proposed vocalization architecture is shown in Figure 4.3. It is similar to that in Acevedo-Valle et al. (2017b), Acevedo-Valle et al. (2018). However, there were some modifications in the dynamical parameters of the motor system and the size of the perception windows.

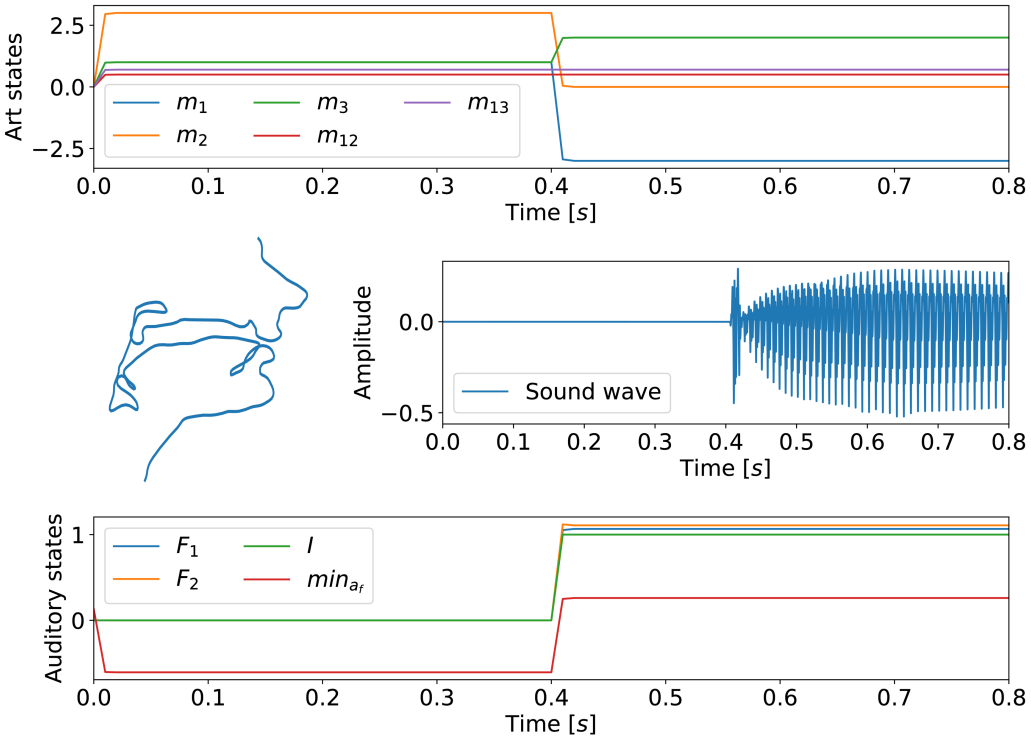

FIGURE 4.3: Ear-vocal tract embodiment. Vocalization experiment example. The upper plot shows the articulatory trajectories. From 0 to 400 ms, the commands m_1, m_2 and m_3 are set, respectively, to 1, 3 and 1, whereas the glottal pressure (m_{12}) and voicing (m_{13}) are set to 0.5 and 0.7, respectively. From 400 to 800 ms, the commands m_1, m_2 and m_3 are set, respectively, to -3, 0 and 2, whereas m_{12} and m_{13} keep their value. The remaining motor commands are set to zero. The middle plot represents the speech sound wave signal. The bottom plot shows the auditory trajectories. There are two perception time windows, one from 0 to 400 ms and the second from 400 to 800 ms. The auditory output s are determined from the average of each trajectories along each one of the perception windows. Auditory output, includes the two first formant frequencies, F_1 and F_2, and an intonation parameter I. Finally, the nociceptive feedback p is determined from the average value of the somatosensory signal $\min(a_f)$.

In the Maeda's synthesizer, the shape of the vocal tract is determined by the position of ten articulators, whereas three phonation parameters control voicing. The articulators and voicing parameters are modeled as dynamical systems. The dynamic behaviors of the articulators and voicing parameters are considered to follow overdamped second order systems' behaviors defined as

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2\,(x - m) = 0, \tag{4.2}$$

with $\zeta = 1.01$ and $\omega_0 = \frac{2\pi}{0.01}$ representing the damping factor and the natural frequency, respectively. The duration of each vocal experiment is 800 ms, whereas m and x represent the motor command for the articulator and the current articulator position, respectively. The structure of a vocalization experiment is shown in Figure 4.3. As two motor commands

are executed sequentially during 400 ms for each of the thirteen articulators, the result is a motor command vector of 26 dimensions.

For the ear (sensory) system, three auditory channels are observed along a couple of time windows: the first two formant frequencies F_1 and F_2, the intonation signal I indicating whether there is sound ($I = 1$) or not ($I = 0$). In fact, the auditory result of the vocalization is computed as a 6-dimensional sensory outcome vector composed by the average of each of the three signals along each of the two perception windows following the execution of the coarticulated motor commands, composed of two articulations of 400 ms each.

Finally, the somesthetic system consists of a signal $\min(a_f)$ which is the minimum of the cross-section area of the vocal tract (shown in Figure 4.4). The minimal value of the area function $\min(a_f)$ is zero when the vocal tract is closed at any point and negative when tissues are overlapping, which lacks physical sense. Thus, when the average of $\min(a_f)$ is negative during either one of the two auditory perception windows, then a nociceptive signal is triggered, causing the agent to perceive pain. This signal is used to build the somesthetic model mapping motor commands $\mathbf{m}$ to pain perception p. Later, the model can be used to predict if a motor action might trigger the nociceptive signal before it is executed.

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 an undesired collision with $p = 1$, and $p = 0$ otherwise.

Through this dissertation, somesthetic information extraction is based on the contact that exists in different sections of an artificial vocal tract surface. It is used to predict violations of motor constraints. When a violation is expected, then the motor command is not executed, and the searching area is moved.

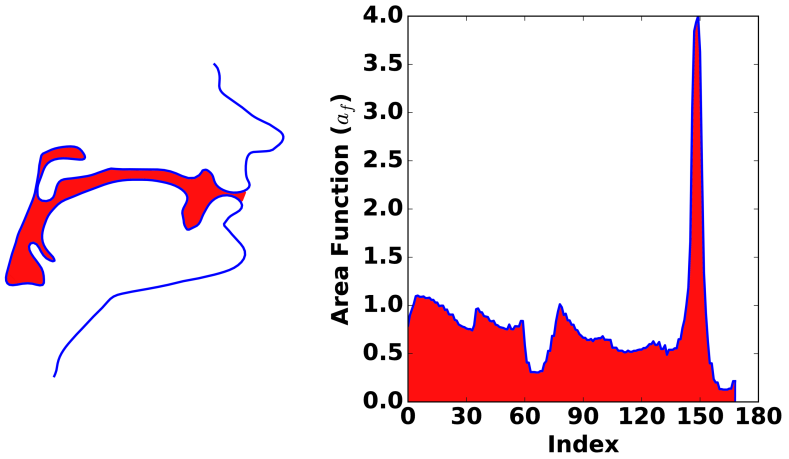

FIGURE 4.4: The area function a_f describes the cross-section of the vocal tract.

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. Finally, to be consistent with the coarticulated nature of speech, only two perception windows are used Kuhl (2004).

4.5 Sensorimotor Exploration Results

As mentioned in previous chapters, the first step in this work was to reproduce the architecture proposed in Moulin-Frier et al. (2013). Making the assumption that in line 4 of Algorithm 6 $p_{tmp} \leftarrow 0$, then the algorithm becomes the one that corresponds to the simple intrinsically motivated sensorimotor exploration represented in Figure 2.1 and used in Moulin-Frier et al. (2013). In the following, we present different experimental results considering the simple and the constraints aware architectures for sensorimotor exploration. The architectures are applied to both, the simple parabolic shaped region system and the ear-vocal tract system.

In this section, simulation results are shown to assess the performance of the constraints aware exploration architecture presented in this chapter with respect to the simple intrinsically motivated exploration architecture. First, the architectures are applied to the toy example, a constrained parabolic shaped region, described in Section 3.7. Secondly, the architecture is applied to the ear-vocal tract system described in Section 4.4.3.

TABLE 4.1: Parameters for Algorithm 6. Parabolic Shaped Region (PSR). Ear-Vocal tract (E-VT).

Parameter	Name	PSR	E-VT
n_e	number of experiments	$10K$	$100K$
randomseed	random seed		
$M_{\mathcal{SM}}$	sensorimotor model	iGMM	iGMM
K_{min}	minimum number of Gaussian components	3	3
ΔK_{max}	maximum increment of Gaussian components	5	10
K_{max}	maximum number of Gassian components	20	30
$\alpha_{\mathcal{SM}}$	forgetting rate	0.2 to 0.05	0.2
$train_{SM}$	training step	100	400
$M_{\mathcal{SS}}$	somesthetic model	wNN ($k = 3$)	wNN ($k = 3$)
M_{IM}	interest model	*discretized progress*	*tree*

4.5.1 Simulation Parameters

To determine the value for all the parameters that are involved in the architecture, the previous results from Acevedo-Valle et al. (2015), Acevedo-Valle et al. (2017a), Acevedo-Valle et al. (2018) are considered. Table 4.1 summarizes all the parameters that must be defined in order to run Algorithm 6.

Parabolic Shaped Region System

Regarding the parameters chosen for the simulation in the case of the parabolically shaped region system, they are chosen based on Acevedo-Valle et al. (2017a), where we first introduced this toy example. The minimum and the maximum number of Gaussian components in the sensorimotor model $M_{\mathcal{SM}}$, which is an Incremental Gaussian Mixture Model (iGMM) introduced in the previous chapter and first introduced to the architecture in Acevedo-Valle et al. (2017a), are $K_{min} = 3$ and $K_{max} = 20$, respectively. The model is trained every 100 experiments. The maximum number of Gaussian components that can be added to the model at each training step is $\Delta K_{max} = 5$. The forgetting rate $\alpha_{\mathcal{SM}}$ for the sensorimotor model is set to 0.2 at the beginning but decreases logarithmically up to 0.05 after 10K experiments. The somesthetic model $M_{\mathcal{SS}}$ is a weighted k-Nearest Neighbor (wNN) model, with $k = 3$. Finally, the interest model M_{IM} is the *discretized progress* model from the `explauto` library.

Ear-Vocal Tract System

Regarding the criteria to choose the simulation parameters in the case of the ear-vocal tract system, the selection was based on the results of Moulin-Frier et al. (2013) and Acevedo-Valle et al. (2017a,b), Acevedo-Valle et al. (2018), the sensorimotor model being an iGMM. Using as a reference $K_{max} = 28$, which was used as a fixed number of components in Moulin-Frier et al. (2013) and Acevedo-Valle et al. (2017b), Acevedo-Valle et al. (2018), when the value of K_{max} is increased then the inference error decreases slightly but the computation time grows considerably, hence a good trade-off was found at $K = 30$. On the other hand, if K_{max} is chosen smaller than 28, then the inference error increases considerably without a significant positive impact on the computational time.

Moreover, if the training step $train_{\mathcal{SM}} = 400$ (from Moulin-Frier et al. (2013) and Acevedo-Valle et al. (2017b), Acevedo-Valle et al. (2018)) is increased, then the training computational time is reduced, but the inference error does not increase considerably. On the other hand, if $train_{\mathcal{SM}}$ is decreased the computational time increases as the training function for iGMMs is called more times without reducing the inference error.

Finally, the remaining parameters for the sensorimotor model were handcrafted to get the best average results. Their impact on the computational time is not considerable. The forgetting rate for the sensorimotor model is set to 0.2. The somesthetic model $M_{\mathcal{SS}}$ is a weighted k-Nearest Neighbor model, with $k = 3$. Finally, the interest model M_{IM} is the *tree* model from the `explauto` library.

4.5.2 Parabolic Shaped Region System

In order to minimize randomness in the results, a large number of simulations were run considering 50 random seeds. Moreover, two scenarios of interest were considered, the first scenario corresponds to the intrinsically motivated sensorimotor exploration from Moulin-Frier et al. (2013), whereas the second corresponds to the constraints aware intrinsically motivated sensorimotor exploration from Acevedo-Valle et al. (2015, 2018). Therefore, fifty exploration simulations were run per scenario using the selected random seeds. Each simulation consists of 100 experiments to initialize $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$. The sensory results obtained from the first step are then used as sensory goals to initialize M_{IM} and 16K exploratory

experiments. For each simulation, several evaluation steps are performed, first after line 1 in Algorithm 6, second after line 2 in Algorithm 6. Thereafter evaluation is performed every 500 samples during exploration, and finally, at the end of the simulation. A set of 441 points described in Figure 3.7 in Chapter 3 is considered to perform the evaluation, the set is evenly distributed along the unconstrained sub-region of the parabolically shaped region.

Figure 4.5 shows the average results during the exploration of the two groups of simulations that were run. The plots were obtained using the average of the 50 simulations considered for each group of simulations. Figure 4.5 (upper) shows the average absolute error $|\mathbf{s}_g - \mathbf{s}|$ during exploration considering a moving average window of 100 samples, defined as:

$$ma_{|\mathbf{s}_g-\mathbf{s}|)}(k) = \left[\frac{1}{n_{rs}}\sum_{i=0}^{n_{rs}}\left[\frac{1}{ws}\sum_{j=k}^{k+ws-1}|\mathbf{s}_{g,i,j} - \mathbf{s}_{i,j}|\right]\right], \tag{4.3}$$

where k stands for the k-th experiment during exploration, n_{rs} is the number of random seeds considered, ws is the window size to compute the moving average, and $|\mathbf{s}_{g,i,j} - \mathbf{s}_{i,j}|$ is the sensory error of the j-th experiment when simulating with the i-th random seed.

Figure 4.5 (lower) shows the average undesired motor configuration ratio along the exploration $ucr_{av,expl}$ defined as:

$$ucr_{av,expl}(k) = \frac{1}{n_{rs}}\sum_{i=0}^{n_{rs}}\left[\frac{1}{k}\sum_{j=0}^{k}p_{i,j}\right], \tag{4.4}$$

where $p_{i,j}$ is the nociceptive signal value of the j-th experiment when simulating with the i-th random seed.

Regarding the evaluation performed over the sensorimotor model every 500 sample, Figure 4.6 was obtained using the averaged results of the 50 simulations for each group. First, Figure 4.6 (upper) shows the average mean evaluation error e_{av} defined as:

$$e_{av} = \frac{1}{n_{rs}}\sum_{i=0}^{n_{rs}}\left[\frac{1}{n_{es}}\sum_{j=0}^{n_{es}}|\mathbf{s}_{g,i,j} - \mathbf{s}_{i,j}|\right], \tag{4.5}$$

where n_{es} is the number of evaluation samples in the dataset and $|\mathbf{s}_{g,i,j} - \mathbf{s}_{i,j}|$ is the evaluation error for the j-th evaluation sample when simulating with the i-th random seed.

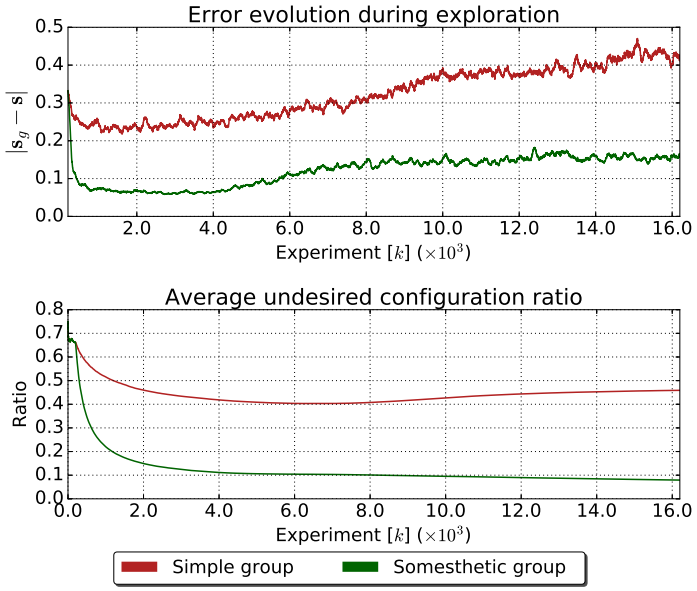

FIGURE 4.5: Results along the exploration running Algorithm 6 using the parabolic shaped region system.

Secondly, Figure 4.6 (lower) shows the average ratio of undesired motor configurations for the evaluation dataset ucr_{av}, which is defined as:

$$ucr_{av} = \frac{1}{n_{rs}} \sum_{i=0}^{n_{rs}} \left[\frac{1}{n_{es}} \sum_{j=0}^{n_{es}} p_{i,j} \right], \tag{4.6}$$

where $p_{i,j}$ is the nociceptive signal value for the j-th evaluation sample when simulating with the i-th random seed.

In Figure 4.6, the size of the round markers in the plot is proportional to the standard deviation between simulations. Finally, Table 4.2 shows some values of interest for further analysis. It displays the values for the undesired motor configuration ratio, the minimum average mean evaluation error achieved per each group of simulations for the evaluation dataset $\min e_{av}$, and the average ratio of undesired motor configurations when achieving $\min e_{av}$. It also displays the standard deviation for those averages.

In Figures 4.5-4.6 and Table 4.2, it is observed that the best results are obtained when the somesthetic mechanism is considered. In general terms, the ratio of undesired motor configurations during exploration and evaluation are lower, and the exploratory and evaluation errors are also noticeably smaller.

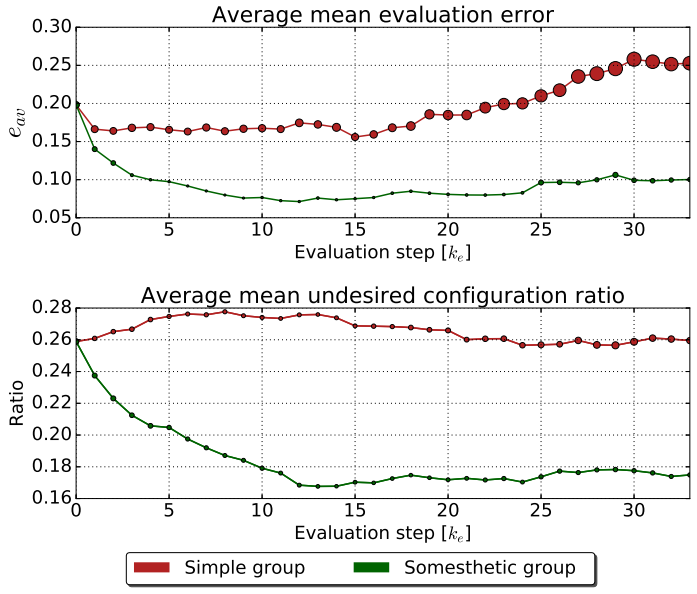

FIGURE 4.6: Evaluation evolution against the dataset evenly distributed along the reachable space of the parabolic shaped region system, running Algorithm 6

TABLE 4.2: Exploration results for the parabolic shaped region system using Algorithm 6.

	Simple group		Somesthetic group	
	value	std	value	std
$ucr_{av,expl}$	0.4589	–	0.0792	–
$\min e_{av}$	0.1560	0.1530	0.0713	0.0321
ucr_{av} for $\min e_{av}$	0.2688	0.0297	0.1684	0.0289

NOTE: The table shows, in order of appearance, the average ratio of undesired motor configurations, the minimum average mean evaluation error, and the average ratio of undesired motor configurations during evaluation for $\min e_{av}$

During exploration, looking at Figure 4.5, it is observed that in general both groups of simulations, simple and somesthetic, achieve a significant decrement of 'painful' configurations. However, the rate of decrement is considerably larger for the group endowed with the somesthetic mechanism. Regarding the behavior of the exploratory error, it is observed that at some point for both cases it starts to increase; for the group without somesthesis, after the sample 2K, and for the group with somesthesis between the samples 4K and 6K. For the latter case, the exploratory error increases until it stays rather steady around a certain value. From the intrinsically motivated learning perspective, we argue that this increment is because the agents have already explored the regions that achieve a high rate of progress in competence values. Thus, they start to exploit regions where the progress is hard, and it might be hard due to their closeness to constrained regions. This idea is backed by the slight increase of

undesired configurations in the simple agents, which indicates that by the time the agent is exploring regions that are leading to some 'painful' configurations. On the other hand, the agents endowed with the somesthetic mechanisms which already possess knowledge about constrained regions, also are going to explore regions that are harder to reach; however, they continue decreasing the rate of 'painful' configurations but at a considerably smaller rate.

Regarding the evolution of the average mean evaluation error e_{av} along the exploration, in Figure 4.6, it is observed that at the beginning (after the initialization step) the evaluation errors and standard deviations (according to the size of the markers) are significant. As the agents explore their sensorimotor system, the evaluation error, and standard deviation decrease notoriously for the agents endowed with the somesthetic mechanism. However, for the agents not endowed with this mechanism, the error decrease slightly, then it keeps slowly changing, until reaching a minimum at around the evaluation step 15. Finally the evaluation error and its standard deviation increase considerably. In general, when the somesthetic mechanism is considered, it is observed a steady and fast improvement. The difference in the behavior between both groups could be directly attributed to the evolution of the ratio of undesired motor configurations during exploration, as indicated in Figure 4.5. As the agents without somesthesis explore constrained regions, the redundancy of the sensorimotor knowledge increases, making harder for the sensorimotor model to represent that knowledge and retrieve accurate motor command inference.

Furthermore, looking at Figure 4.6, standard deviation markers indicate that those agents endowed with the somesthetic mechanism achieved more robust results than the others. Finally, in the case of the agents without somesthesis, it is observed that the ratio of undesired configurations along the exploration for the evaluation dataset does not increase as the error does when the agent explores conflicting regions, *How is this result compatible with the behaviour of the exploratory error and undesired motor configuration ratio during exploration?* If the agents focus more during exploration on regions close to constraints, then the sensorimotor knowledge they have on the permitted region degrades as they start to forget given the learning rate $\alpha_{\mathcal{SM}}$. Hence, an overall increase of the error is provoked when evaluating against a data set evenly distributed along the permitted region. However, as the agents continue exploring close to the constraints, the knowledge they have in those regions is good enough, and they do not produce more collisions. On the other hand, in the case of agents endowed with somesthesis, after reaching the best evaluation results they show both, slight

transient behaviors in the evaluation error and the ratio of 'painful' configurations during evaluation. This behavior might be due to the fact that, as these agents explore more uniformly the sensorimotor regions, if they focus for some intervals in exploring regions far from constraints, then the knowledge they have close to constraints degrades. Later, when they focus exploration on these conflicting regions, they produce more undesired motor configurations, and the error also increases.

4.5.3 Ear-Vocal Tract System

Experimentation for the ear-vocal tract system is divided into two groups subdivided into two subgroups each. For one subgroup the proposed somesthetic mechanism is not considered, whereas for the other group the mechanism is considered. The difference between the two simulations in each subgroup lies in the initialization criteria for M_{IM}. On the one hand, for one simulation all sensory results obtained when initializing $M_{\mathcal{SM}}$ are considered as sensory goals when initializing M_{IM}. On the other hand, for the other simulation, only the goals that did not trigger the nociceptive signal p during the initialization of $M_{\mathcal{SM}}$ are considered to initialize the interest model. Even though in more recent works (Acevedo-Valle et al., 2017a), we found that it is better concerning competence performance to initialize M_{IM} with all the initial sensory results, we also claim that it is important to show the impact that initialization might have on developmental approaches. In Acevedo-Valle et al. (2015, 2018), it was considered a similar criterion to the one that does not include 'painful' configurations. Whereas, Rayyes et al. (2017) proposes an initialization without 'painful' configurations. Thus, simulations in this section may also help to conclude if any criterion is beneficial in some way.

Six different random seeds were considered to generate random initialization sets of motor commands from uniform distributions. In total, twenty-four independent simulations were run using Algorithm 6, six per each of the four groups described above. All simulations consisted of 100K vocalizing experiments plus a number of initialization vocalizing experiments varying from 1K to 2K. The limits for initializing motor commands related to the vocal tract articulators were $[-1, 1]$, whereas for motor commands related to the phonation parameters were $[0, 0.7]$.

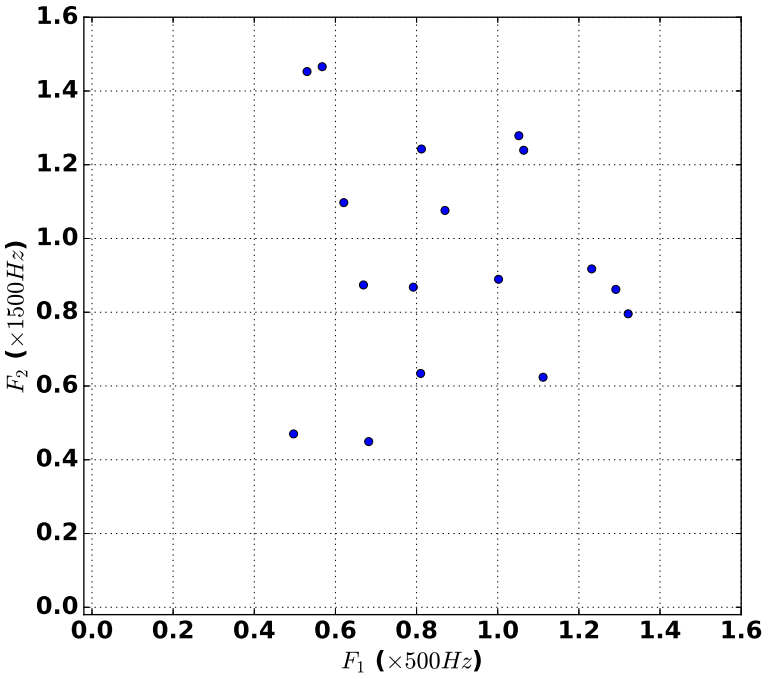

FIGURE 4.7: Fixed vocalizations obtained with `divapy`. Single auditory results for vocalizations with no 'painful' configurations used to generate 323 co-articulated gestures.

Summarizing, $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$ are initialized together as indicated in line 1 of Algorithm 6 with the different initial motor command sets. Later, using the auditory results of the first stage with the criteria for each simulation subgroup described above, the interest model M_{IM} is initialized as indicated in line 2 of Algorithm 6. During the initialization of M_{IM}, $M_{\mathcal{SM}}$ is used to infer the motor actions that will likely produce the initial auditory goals. These commands are executed without considering the nociceptive prediction p_{tmp}. Afterward, the intrinsically motivated sensorimotor exploration is run for 100K experiments.

Finally, to evaluate the exploration respect to some fixed points in the sensory space $\mathcal{S}$, 17 single vocalizations that do not produce undesired motor configurations are chosen using `divapy` shown in Figure 4.7. Next, these vocalizations are recombined to generate coarticulated gestures that are in the format of the embodiment described in Figure 4.3 resulting in 323 samples. This set of samples will be called evaluation dataset $\mathbf{S}_{eval}$ in the following. Evaluation on $\mathbf{S}_{eval}$ is performed every 2.5K samples during each simulation.

In Figure 4.8 (upper), it is shown the average mean evaluation error e_{av}, defined in Eq. (4.5), for each one of the four groups of simulations. As in the case of the toy example results, the markers size is proportional to the standard deviation between the simulations. In Figure 4.8 (center), it is shown the moving average of the mean error through simulations defined in

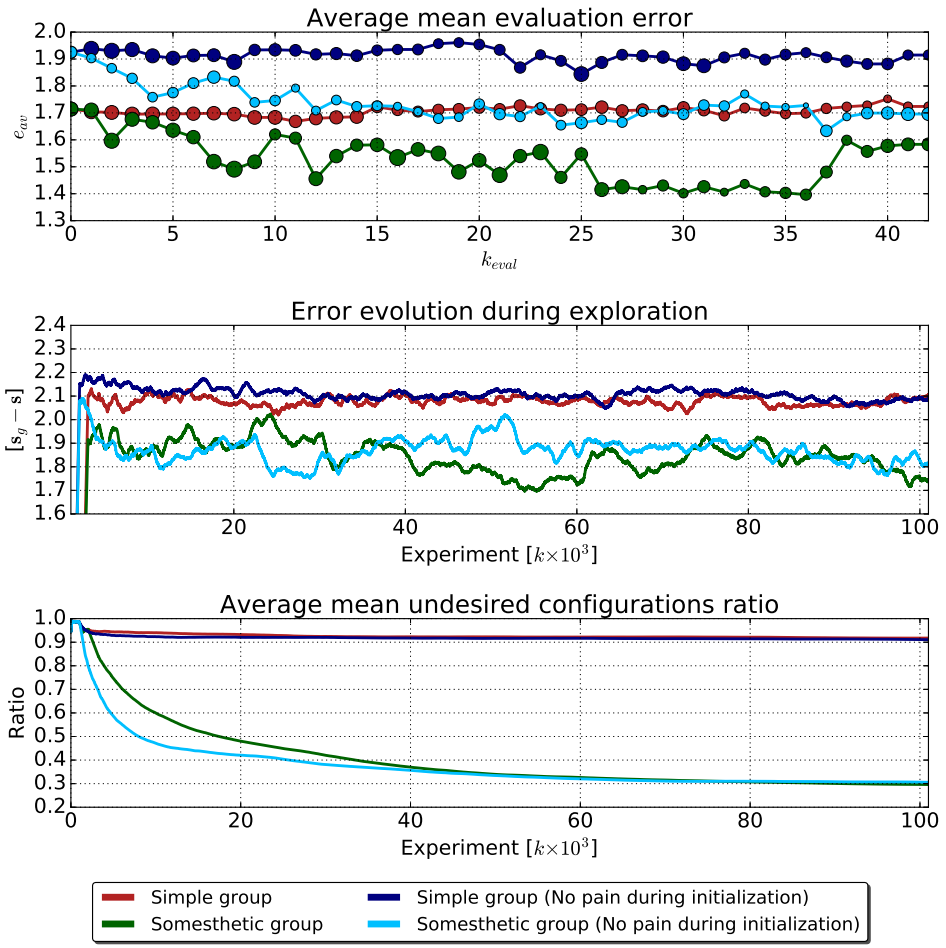

FIGURE 4.8: Results for simulation with the ear-vocal tract system using Algorithm 6. (Upper) Average mean evaluation error performed every 2.5K samples during exploration against $\mathbf{S}_{eval}$. (Center) Average sensory error during exploration using a moving average window of 1000 samples. (Lower) Average undesired motor configuration ratio evolution along the exploration.

Eq. (4.3) considering a window size ws of 1000 samples. Finally, in Figure 4.8 (lower), it is shown the undesired motor configuration ratio (a.k.a 'painful' configurations) defined in Eq. (4.4).

On Competence and 'Painful' Configurations

Results suggest that those agents that are endowed with the somesthetic mechanism perform better than those which are not endowed with it. From Figure 4.8 (upper), it is observed that the simulation groups can be characterized by the fact of using or not 'painful' configurations during the initialization of the interest model M_{IM}. Those simulations that did not consider 'painful' configurations produce higher average mean evaluation error e_{av} at the beginning of the exploration. However, regarding progress on the ability of the agents to

reproduce the evaluation dataset $\mathbf{S}_{eval}$, it is observed that those agents endowed with the somesthetic mechanism improve along the exploration, whereas the agents without somesthetic mechanism do not improve, at least not clearly. In the case of the somesthetic group with full initialization, it is observed that the average mean evaluation error increases at the end of the simulation. However, as the agents are exploring, and they are unaware of those sensory units, it is normal that if the agents start to explore goals that are farther from $\mathbf{S}_{eval}$, then its skills to produce those goals will shrink. In general terms, the best behavior is observed with the somesthetic group, even though the somesthetic group that did not consider 'painful' configurations during initialization of M_{IM} shows a tendency to improve its performance, it barely achieves a similar performance to that of the simple group that considered all the initial configurations during initialization. We argued that these results are because agents that considered 'painful' configurations during initialization, retain more knowledge regarding sensorimotor regularities, and later they might refine this knowledge guided by the intrinsic motivations and shaped by the somesthetic mechanisms if present.

In Figure 4.8 (center), it is observed that there are other interesting differences between the simulation groups. In general, we obtained similar results to that from Acevedo-Valle et al. (2018) after initialization[1]. Therein, the criteria to initialize the interest model was similar to the one that does not include the 'painful' vocalizations[2]. So looking at those agents in Figure 4.8 (center), it is observed that the somesthetic group starts reducing its exploratory error after initialization, whereas in average the simple group has a trend toward higher error after initialization. Thinking about the competence, defined as $c = e^{-|\mathbf{s}_g - \mathbf{s}|}$, this means that for the simple group the competence decreases after initialization. On the other hand, for the somesthetic group the competence increases after initialization, corroborating what was observed in Acevedo-Valle et al. (2018). These error tendencies coincide with a significant decrement of the 'painful' configuration ratio in Figure 4.8 (lower). This fact again suggests that the somesthetic mechanisms are relevant to improve the progress in achieving better competence performance.

Regarding the behavior of the subgroup that considers all the initial vocalizations to initialize

[1]The values of the error during initialization are small, that is the reason that the lines at the beginning come from below of the plot. That part of the lines is left out of the plot in order to obtain a better perspective of the exploration stage.

[2]In the case of the ear-vocal tract motor configurations, and sensorimotor experiments are also referred as to vocalizations.

M_{IM}, it is observed a similar behavior to that of the simple groups without 'painful' initialization, but in general, they show lower errors than the other subgroup. They show a slight tendency of the error to grow after initialization, however as the exploration progresses, at least in the case of the somesthetic group, the error tendency is to decrease.

In Figure 4.8 (lower), it is observed that in general the 'painful' vocalization ratio stays high along the exploration. This observation is related to the lower values of competence during the exploration (higher errors) compared to the somesthetic groups. In the case of the somesthetic groups of simulations, it is observed that the one that does not consider 'painful' vocalizations during initialization of M_{IM}, shows a more abrupt reduction of this family of vocalization after the autonomous exploration begins. It is due to the fact that the interest model is generating goals similar to that of initialization. On the other hand, the group of simulations considering all the initial experiments to initialize M_{IM} shows a slower decrease of undesired vocalizations. Despite the rate of decrement of this configurations by the experiment 60K, the group reaches the performance in this respect of its initialization counterpart. Therefore, regarding undesired configurations, regardless of the initialization criteria, the somesthetic mechanisms is capable of reaching a minimum of undesired configuration ratio in both cases. Further analysis is performed below.

Some observers might ask the reason of low error values (high competence) at the beginning of the simulations. We argue that it is an expected result as the error computation begins when M_{IM} is initialized with sensory goals drawn from the initial productions of the agents. In other words, $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$ models are initialized around a set of initial vocalizations. The initial auditory productions are then selected as sensory goals considering two different criteria. Afterward, 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 error values increase. Afterward, as the intrinsically motivated exploration evolves, there is a general tendency of the evaluation and exploratory errors to decrease. However, it is more notorious in the somesthetic groups. In general, running different experiments through the development in this work, we observed that the initial increment in the exploratory error after initialization and its tendency to decrease through the exploration depends strongly on the parameters used for the sensorimotor model $M_{\mathcal{SM}}$. If the forgetting rate of the sensorimotor model $\alpha_{\mathcal{SM}}$ is 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 rate is high, then the agent will adapt its model to the new data very fast, but it will forget faster its previous knowledge since it is not reinforced.

Figure 4.9 was obtained using the average results obtained for each type of vocalizations' proportions classified into three types: (a) Silent, if no phonation occurs in any of the two perception windows; (b) Unarticulated, if phonation occurs in one of the perception windows; and (c) Coarticulated, if phonation occurs in both perception windows. Plots in Figure 4.9 show the evolution of the proportional contribution of each vocalization type to the total of vocalizations through the exploration for each simulation group. It could be considered that all the silent vocalizations are a waste of energy during the exploration. Knowing which regions of the motor space are leading to 'painful' configurations might be a relevant knowledge for the agent to avoid this kind of silent vocalizations. Whereas the agents which are not endowed with the somesthetic mechanism keep exploring conflicting regions, the agents belonging to the somesthetic groups avoid exploitation of those regions due to their ability to predict nociceptive results from a given motor command.

Therefore, we argue that one of the reasons why somesthetic groups achieve better performance compared to their simple counterpart is the proportion of silent vocalizations. In a significant proportion, those vocalizations are produced by 'painful' configurations for which the vocal tract is blocked and, in consequence, no sound is produced. The agents in the simple groups produce much more undesired 'painful' configurations as observed in Figure 4.8 (lower) and, therefore, produces more non-phonatory vocalizations as corroborated in Figure 4.9. Reversely, 'painful' configuration may also occur frequently in unarticulated vocalization, in which the airflow is blocked in one of the perception windows, or the phonation motor parameters are not positive. Thus, in Figure 4.9 it is possible to observe that an agent without somesthetic mechanism achieve a certain command of unarticulated vocalizations, even though silent vocalizations prevail. A last point to comment on Figure 4.9 is the fact that in proportions, the group that best performs is the somesthetic group without 'painful' initialization, as corroborated with the final ratios shown in Table 4.3. This result agrees with the fact that this group of agents achieved a faster decrease of undesired configurations that are more likely to lead to silent vocalizations as already commented when Figure 4.8 was analyzed.

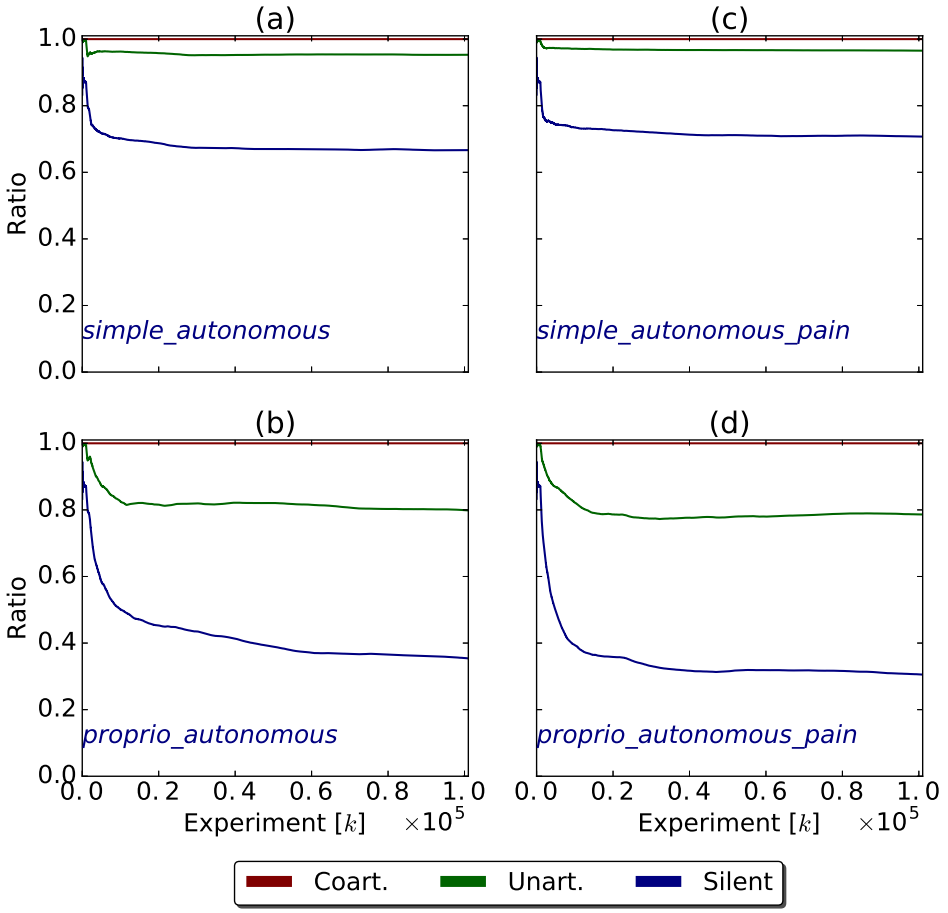

FIGURE 4.9: Proportions of vocalization classes. (a) Simple group. (b) Somesthetic group. (c) Simple group (No pain during initialization). (d) Somesthetic group (No pain during initialization).

On Explored Regions

Results presented in Table 4.3 allow to check numerically what was commented previously regarding Figures 4.8-4.9. Different numerical descriptors were obtained for each of the four groups of simulations. First, it is shown the minimum value achieved for the average mean evaluation error e_{av}, also shown in Figure 4.8 (upper). Then, the average exploratory error during exploration for all the simulations for each group. The table also shows the average 'painful' vocalization ratio produced during exploration for each group of simulations. Afterward, it is shown the average proportion of unarticulated and coarticulated vocalizations computed for each group of simulations. In general, the numerical descriptors corroborate that better results are obtained for those groups of agents endowed with the nociceptive mechanism. On the one hand, the somesthetic group with full initialization achieved better performance, but on the other hand, the somesthetic group without 'painful' initialization

TABLE 4.3: Exploration results for the era-vocal tract system using Algorithm 6.

Group	simple	somesthetic	simple (NP)	somesthetic (NP)
min (e_{av})	1.6684	1.3965	1.8442	1.6334
Average $\lvert\mathbf{s}_g - \mathbf{s}\rvert$ exploration	2.0450	1.8159	2.0908	1.8548
$ucr_{av,expl}$	0.9176	0.2960	0.9108	0.3057
Unart. vocalization ratio	0.2892	0.4493	0.2582	0.4810
Coart. vocalazation ratio	0.0474	0.2040	0.0346	0.2138
Convex hull volume	0.9597	1.0709	0.9517	1.0030

Note: Experiments with different vocalization initial sets for simple and somesthetic agents. The minimum value obtained for the average mean evaluation error, the average 'painful' articulation ratio during the exploration, the average ratio of unarticulated and coarticulated vocalizations along the simulations are shown and finally, the volume of the convex-hull encapsulating the frequency component of the explored auditory data. (NP) Indicates that those groups are the ones initialized without 'painful' configuration.

for M_{IM} achieved better results regarding the proportions of unarticulated and coarticulated vocalizations.

Furthermore, Table 4.3 also contains the volume of the convex hulls described by the explored data for each simulation group. The sensory data obtained during each simulation and related to the formant frequencies are considered (F_{11}, F_{21}, F_{12}, and F_{22}) to obtain such a volume. That is, the intonation dimensions for both perception windows (I_1 and I_2) are dropped. Then, the Python's `scipy`[3] library is used to compute the convex hull that encapsulates the data for each simulation. The volume of the convex hull for the simulations of each group is then averaged, and it is the descriptor shown in the table.

Regarding the trade-off between exploration and exploitation, in this section, we obtained a slightly different result to that from Acevedo-Valle et al. (2018). We attribute the difference in the results to the improvements that were made to the modeling approaches and the better tuning of the simulation parameters. Acevedo-Valle et al. (2018), based on larger volumes of the convex hulls for the agents without the somesthetic mechanism, argued that these agents show better performance compared to agents endowed with the somesthetic mechanism. On the other hand, the somesthetic groups showed better performance with respect to exploitation, as agents avoid exploring uninteresting regions with a high number of 'painful' configurations. However, experimentation in this dissertation shows that agents with somesthetic mechanism achieved better performance in terms of the volume of the explored

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.ConvexHull.html

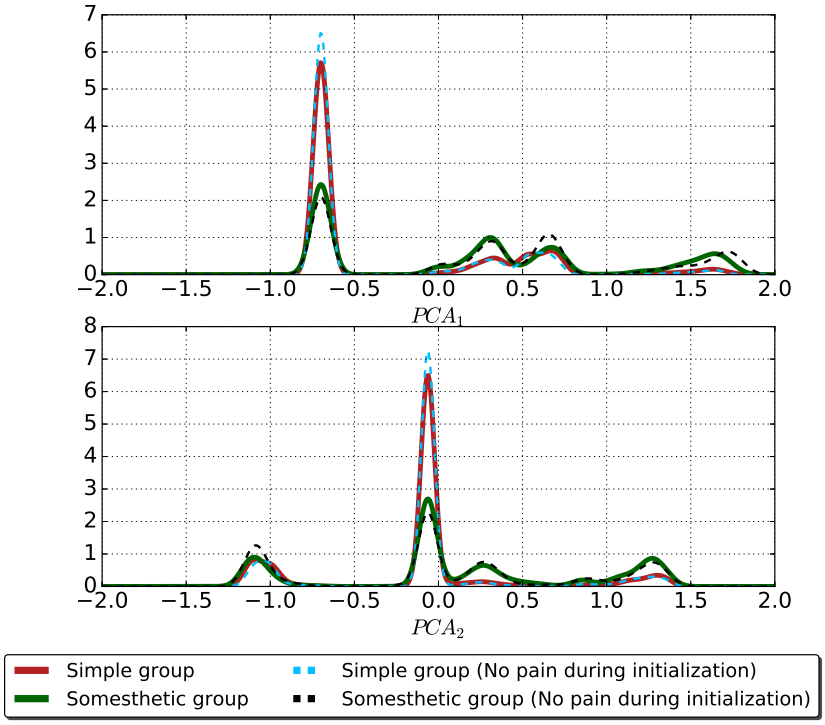

FIGURE 4.10: Data density distributions computed using Gaussian KDE for all the data obtained during simulations and considering the first two PCA components.

region, and in general agents explored larger region than in Acevedo-Valle et al. (2018). Regarding these results, it is essential to remark that in Acevedo-Valle et al. (2018), the limits for initializing motor commands related to the vocal tract articulators were $[-0.1, 0.1]$. Thus, the sensorimotor knowledge acquired by the agents during the initialization stages may also have shaped their capability to explore and discover new regions later during intrinsically motivated exploration.

Looking at the convex hull volumes shown in Table 4.3, first we look at the ratio between the simulations within each subgroup of simulations. For both cases, the subgroup with full initialization (1.0709/0.9597) and the group in which M_{IM} is initialized without 'painful' vocalizations (1.0030/0.9517), convex hull volumes ratios show that agents endowed with the somesthetic mechanism perform better than their counterpart without the mechanism as they achieve a larger explored region. On the other hand, if we compare both initialization criteria, it is observed that in both cases, the simple group (0.9597/0.9517) and the somesthetic group (1.0709/1.0030), those agents that considered the whole initial set of vocalizations to initialize M_{IM} performed better in terms of exploration as they explore larger areas.

In general, as in Acevedo-Valle et al. (2018), we observe that the agents endowed with the somesthetic mechanism performed better than those that are not. However, looking at the

somesthetic groups, which are the most interesting for us at this point? The somesthetic group initialized with all the initial productions explores a 6.77% wider region than those with 'unpainful' initialization for M_{IM}, their performance is 2.14% regarding the exploratory error, 14.5% for the best mean evaluation error achieved against $\mathbf{S}_{eval}$, and their ratio of undesired configurations produced during exploration was 3% lower. The only feature in which the somesthetic group initialized without 'painful' configurations is better with respect to the other group is in the proportion of unarticulated (6.6% more) and coarticulated vocalizations (4.6% more).

Finally, as in Acevedo-Valle et al. (2018), to analyze the vocalization distributions obtained using the different exploration algorithms and initialization criteria, it is performed a sample density analysis over the formant frequency dimensions ($F_{1,1}$, $F_{2,1}$, $F_{1,2}$, and $F_{2,2}$). First of all, in order to make the results easier to visualize we perform a Principal Component Analysis (PCA) procedure. To do this, first we concatenate all the data obtained from all the simulations, in this way the results are comparable. Once the PCA was performed over the whole data, it was observed that the data can be represented with high confidence with the two first principal components which keep a total of 98.07% (56.36% + 41.71%) of the original information. The two principal components kept are:

$$\begin{aligned} \vec{pca}_1 &= \begin{bmatrix} 0.5125 & 0.6111 & 0.3975 & 0.4538 \end{bmatrix}, \\ \vec{pca}_2 &= \begin{bmatrix} -0.3752 & -0.4725 & 0.5174 & 0.6069 \end{bmatrix}, \text{ and} \\ \mu &= \begin{bmatrix} 0.3302 & 0.4004 & 0.3085 & 0.3572 \end{bmatrix}, \end{aligned}$$

where μ is the estimated mean for all the sensory data used to perform the PCA procedure. Once PCA was performed, the sensory data for each group was transformed from 4-D data into 2-D data, thence Kernel-Distribution Estimation (KDE) was performed using Gaussian-Kernels according to Scott (2015).

In Figure 4.10, it is observed the density distributions obtained for each of the four groups considered for simulations. In general, the results in the figure show that the agents explored similar regions, but with different intensity. The two high peaks correspond to the region where the silent vocalizations are located. As one can check, the transformation under the principal components for the null phonatory results is

$$\mathbf{s}[0,1,2,4] = [0,\ 0,\ 0,\ 0] \xrightarrow{\text{PCA transform}} [-0.699,\ -0.063],$$

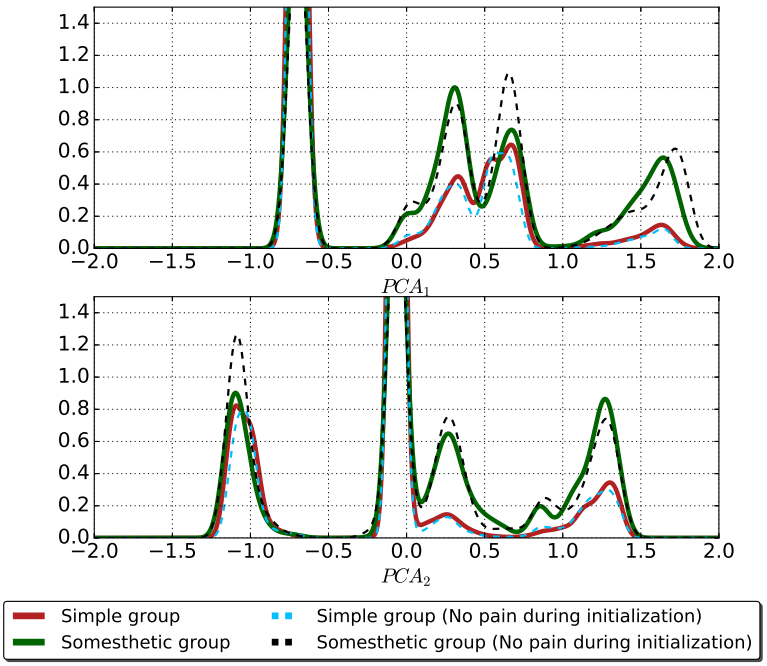

FIGURE 4.11: Data density distributions computed using Gaussian KDE for all the data obtained during simulations and considering the first two PCA components [zoomed].

thus, as expected, the agents in the simple group explored more intensively in the silent region as corroborated by Figure 4.9. On the other hand, agents endowed with the somesthetic mechanism explored more uniformly along the sensory space. The original figure was zoomed to obtain Figure 4.11 to have a better perspective. In the latter figure, looking especially to the somesthetic groups, it is observed that initialization had an impact on the intensity in which the agents lean to explore with different intensity on different regions, showing dominance in different peaks of the exploration.

In Figure 4.12, data distributions along the principal components are shown only considering the 'unpainful' vocalizations along the explorations. As each group of simulations produced a different number of 'unpainful' vocalizations, the results are scaled to the number of 'unpainful' vocalization per group divided by the number of 'unpainful' vocalizations for the group that produced the most vocalizations of that kind. With the scaling rule, data distributions for that simulation without somesthetic mechanisms are diminished because results showed beforehand to show that they produce mostly 'painful vocalizations'. Moreover, again somesthetic agents for both initialization criteria show similar results regarding the explored regions but, again, leaning dominance in different data density peaks.

In Figure 4.13, data distributions along the principal components are shown only considering the phonatory vocalizations along the explorations. The criterion to scale the estimated

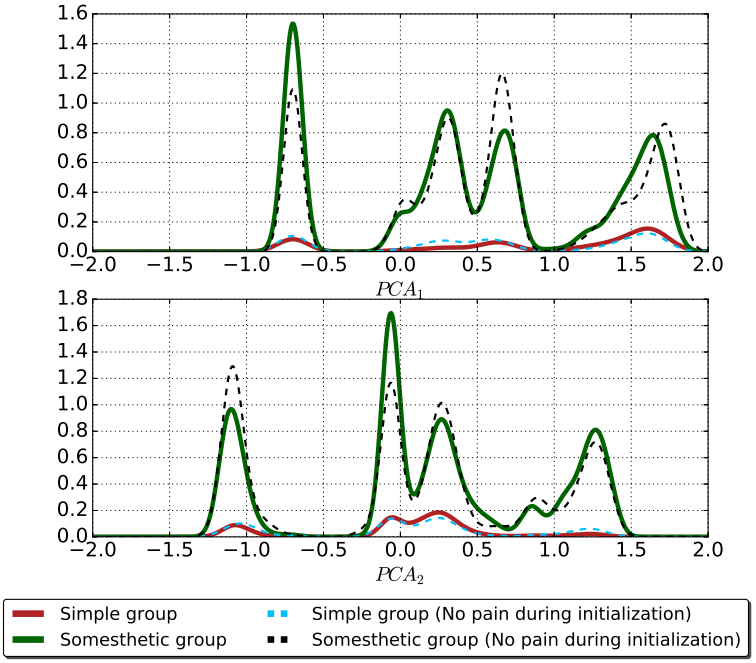

FIGURE 4.12: Data density distributions computed using Gaussian KDE for the 'unpainful' data obtained during simulations and considering the first two PCA components.

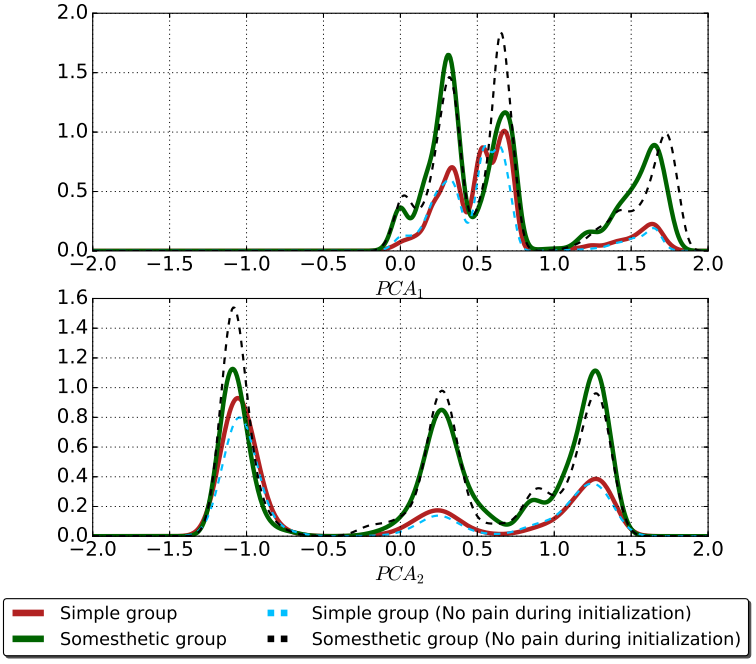

FIGURE 4.13: Density distribution computed using Gaussian KDE for the phonatory data obtained during simulations and considering the first two PCA components.

densities was the same to scale the distributions in Figure 4.12 in order to make the results comparable. In this case, the scaling factor did not diminish data distributions for those

simulations without somesthetic mechanisms as even when they produce mainly 'painful' vocalizations, those vocalizations may still being unarticulated phonatory vocalizations. Furthermore, the behavior of data distributions in somesthetic agents for both initialization criteria show similar results regarding the explored regions but with different data density peaks. However, here it emerges an important difference when we compare Figure 4.13 with Figure 4.12: as for Figure 4.13 the silent peaks of the simulation diminished, it is possible to observe that the exploration over regions of the space described by the phonatory vocalizations discovered under the PCA transformation is similar for both somesthetic groups. However, in fact, the exploration in these regions that produce sounds is more intense for those agents that their interest model was initialized without 'painful' vocalizations. This latter result is consistent with the difference in the proportion of the different types of vocalizations shown in Figure 4.9.

There is one important question left: if the agents without 'painful' initialization explored more intensively the phonatory regions, "why do the agents that considered all initial vocalizations to initialize the interest model perform better with respect to the evaluation data set $\mathbf{S}_{eval}$?" As an attempt to answer this question, it was generated Figure 4.14, therein the Gaussian-KDE for $\mathbf{S}_{eval}$ is shown. Next, we use the distributions in Figure 4.14 to filter the distributions in Figure 4.13. After filtering these distributions, Figure 4.15 was obtained, and the integral of the filtered distributions was computed. Neglecting the integration for the agents without somesthesis, for the agent with full initialization it was obtained $PCA_1 = 0.3230$ and $PCA_2 = 0.3535$, whereas for the agent without 'painful' initialization it was obtained $PCA_1 = 0.3221$ and $PCA_2 = 0.3714$. Recalling that the first principal direction contributes with the 56.3% and the second with 41.71%, then the difference of performance against $\mathbf{S}_{eval}$ may be the slight difference between the integral under the distribution curve of each simulation group, using different initialization criteria, along the regions described by the evaluation dataset.

4.6 Discussion

In this chapter, an application of active learning techniques applied to the study of sensorimotor behaviors in embodied agents considering motor constraints has been introduced. The architecture, based on the literature review from previous chapters, has been presented

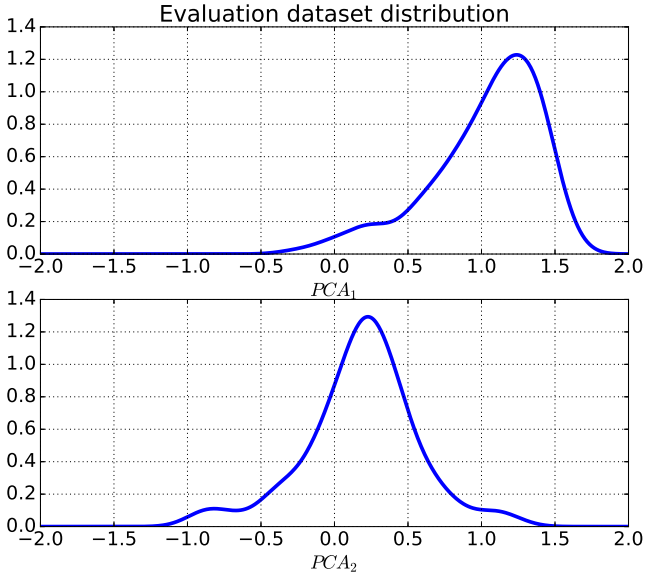

FIGURE 4.14: Data density distributions computed using Gaussian KDE for the evaluation data set $\mathbf{S}_{eval}$.

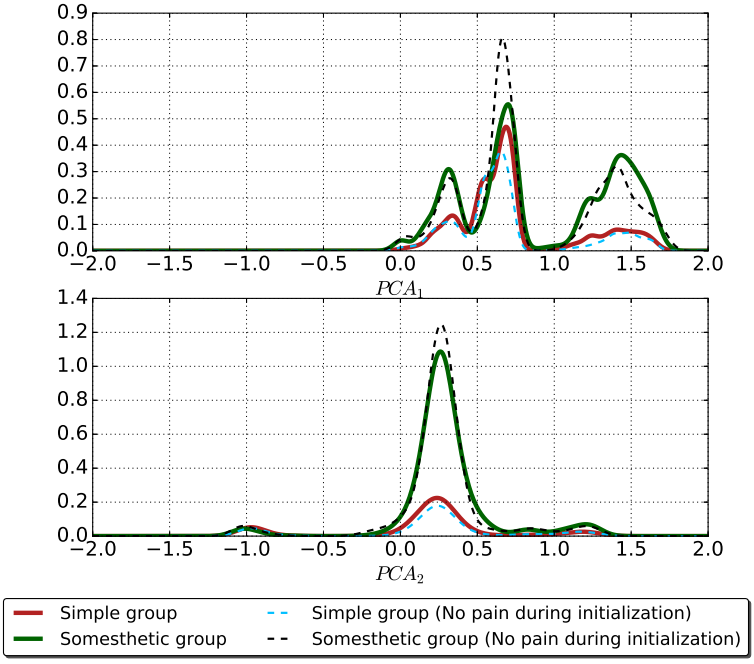

FIGURE 4.15: Density distribution computed using Gaussian KDE for the phonatory data obtained during simulations and considering the first two PCA components and filtered using the Gaussian-KGE for the evaluation data set $\mathbf{S}_{eval}$.

as an intrinsically motivated sensorimotor exploration architecture with motor constraint awareness. Constraint awareness is achieved by providing a somesthetic mechanism which endows an artificial agent with the capacity to autonomously generate a somesthetic model. This model is later used to predict the somesthetic consequences of motor actions and to

avoid their execution if they are expected to generate motor configurations that may lead to the perception of 'pain'.

There has been particular interest in the ear-vocal tract exploration problem. In general, consistent with the results in Acevedo-Valle et al. (2015, 2018), it was observed that the somesthetic mechanism improved the quality of learning according to the sensory exploration error. However, unlike Acevedo-Valle et al. (2018), we did not observe a trade-off between exploration and exploitation in the sensory space when the somesthetic mechanism is considered. We argue that discrepancy with previous results is due to the improvements that were performed through the evolution of this project to the modeling methods as well as a better tuning of the simulation parameters.

First, for a simple toy example it was shown that, in all the considered dimensions, agents endowed with the somesthetic mechanism performed better. Secondly, regarding the ear-vocal tract system, the discussion is wider as the results show many important facts. As mentioned beforehand, vocal-auditory spaces are high dimensional redundant spaces; thus an auditory output may be produced by different articulatory motor configurations, mainly non-phonatory vocalizations. A large proportion of these articulatory configurations may lead to undesired motor configurations, especially those producing silent and unarticulated vocalizations. Hence, we argue that sensorimotor redundancy is reduced when a somesthetic mechanism, based on ideas of proprioception, nociception, and pain, is included in the system allowing artificial agents to spend more energy exploring and exploiting 'unpainful' phonatory vocalizing regions. 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 motor constraints are met, and phonation occurs. In this sense, we showed how sensorimotor exploration, and in general sensorimotor knowledge, can be shaped by constraints and therefore by constraint awareness.

Regarding the emergence of vocal developmental stages, as the experiment was initially proposed in Moulin-Frier et al. (2013) to illustrate how intrinsically motivated exploration may explain the sequence of vocal developmental stages, we observed a clear developmental trajectory in the somesthetic groups as shown by Figure 4.9. Agents start producing mostly silent vocalizations, then they start increasing their articulated vocalizations production, and at a slower rate, they also increase the proportion of coarticulated vocal gestures. The

main difference is that whereas agents endowed with the somesthetic mechanisms achieve a dominant proportion of unarticulated gestures and a significant proportion of coarticulated gestures, those agents without somesthetic mechanisms predominantly produce silent vocalizations, a significant proportion of unarticulated gestures and a rather small proportion of coarticulated gestures. Again, regarding the developmental trajectory, agents endowed with somesthesis perform better.

Regarding the initialization criteria, we considered two options, either to initialize the interest model with 'painful' configurations or without. The improvements to the architecture and other differences in the initialization criteria, allow us to corroborate what was argued in Acevedo-Valle et al. (2018), therein it is that initializing models in the no phonation region of the auditory space, as is the case of infants, would lead to a better picture of the developmental stages for the first year of infants life. Our initialization for $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$ includes a high rate of non-phonatory vocalizations, even though through the agents' life the stages where unarticulated and coarticulated vocalizations emerge are visible. When M_{IM} is initialized without 'painful' experiments, the development was observed to occur faster, with the trade-off of a slightly lower performance regarding exploration error. However, the later agents showed a better performance against a social data set, that is going to be seen in the next chapter correspond to vowel units similar to the German vowels, which gives an important advantage to the 'painful' initialization of the interest models.

On the advance toward vocal exploration, we have shown 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 the study to understand the processes that occur in prelinguistic children during early vocal development, we should consider studying in greater depth the first period of vocalization development. In this sense, it is important to notice that vocal development in infants with regular development occurs in a socially guided environment as mention in many works, e.g., Kuhl (2004). 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.

It is essential to include the social factor in the developmental learning of an artificial agent to understand better the role of somesthetic mechanisms in the presence of other mechanisms that may shape development as social interaction. In the next chapter, the role of social interaction in prelinguistic vocal development in children is discussed. Moreover, an

architecture aimed at studying social impact in constraints aware intrinsically motivated sensorimotor exploration is presented.

Chapter 5

The Role of Imitation Episodes in Intrinsically Motivated Sensorimotor Exploration

"In rising children, we need to continuously keep in mind how we can best create the most favorable environment for their imitative behavior. Everything done in the past regarding imitation must become more and more conscious and more and more consciously connected with the future."
— Rudolf Steiner,

Chapter 4 introduced a cognitive architecture combining ideas from prelinguistic development, embodiment, and somesthetic senses. That architecture is aimed at studying early sensorimotor development taking into account motor constraints, which are implicit to any embodied agent. Simulation results demonstrated that, in terms of developmental learning, the ability of an agent to acquire a simple somesthetic sensorimotor model, based on a nociceptive signal, enhances the developmental performance regarding explored volume, evaluation error and ratio of undesired motor configurations.

At the end of Chapter 4, it was also mentioned that intrinsic motivation mechanisms and somesthetic senses are not the only elements shaping sensorimotor development, but social

interaction may also play a key role during the emergence of complex sensorimotor behaviors. Therefore, in this chapter, we introduce the second major contribution of this dissertation, aimed at extending the study of early development using constraints aware intrinsically motivated exploration algorithms. Herein, a social reinforced sensorimotor exploration architecture is presented to study how imitation behaviors shape sensorimotor development. Similarly to Chapter 4, particular interest is directed toward studying how social reinforcement fosters the emergence of complex vocal behaviors.

Compared to Chapter 4, the cognitive architecture in this chapter includes social elements that provide means to evaluate the behavior of the exploration architecture in the presence of social reinforcement through imitation episodes. Hence, the new architecture considers a social instructor which reinforces learning along the intrinsically motivated exploration. Reinforcement is made using socially relevant sensor units, e.g., sensor units for speech stand for speech utterances. The proposed architecture is built according to developmental studies on prelinguistic social development, in particular, the influence of *imitation/expansion* maternal responsiveness described in Gros-Louis et al. (2014).

The methodology to assess the potential of the proposed architecture is similar to that in Chapter 4. Firstly, we test it executing simulations with the parabolically shaped region system, the toy example presented in Section 4.4.2, that allows a simple evaluation of the proposed architecture. Next, we run simulations using our ear-vocal tract simulator. To provide a more realistic scenario when experimenting with the ear-vocal tract system, the speech utterances considered for the instructor are German vowel-like coarticulated speech gestures. The work presented in this chapter was partially published in Acevedo-Valle et al. (2017a) and the most recent results have been submitted to a peer-reviewed journal and are currently under the first round of revisions (Acevedo-Valle et al., 2018).

Besides studying the role of motor constraints, preliminary results are provided in Acevedo-Valle et al. (2017a) pointing to some evidence that social feedback mechanisms, even considering a simple imitation scenario, drives development more efficiently during intrinsically motivated explorations. Furthermore, it was corroborated that constraint awareness and social reinforcement benefit the efficacy of intrinsically motivated exploration architectures, improving both, the prediction of action consequences and the volume of the explored sensorimotor regions.

This chapter is organized as follows. Section 5.1 is aimed at introducing the relevance of social interaction for sensorimotor control development, specially prelinguistic vocal development, in embodied agents from the perspective of human development. Then, Section 5.2 discusses former results regarding the role of social mechanisms in studies of sensorimotor development and prelinguistic vocal development using artificial agents. Section 5.3 explains how the social reinforcement mechanism is integrated to the intrinsically motivated sensorimotor exploration with constraint awareness. The experimental setup and results are presented in Section 5.4 and Section 5.5, respectively. Finally, the discussion is completed in Section 5.6.

5.1 Social Reinforcement in Sensorimotor Development

Since humans are genuinely social beings, autonomous sensorimotor exploration is just one aspect of developmental learning. Predominately, skills acquired by exploration are reinforced and extended by social mechanisms, e.g., learning by demonstration or imitation learning. In this section, we firstly introduce relevant studies from developmental sciences that show the impact that social reinforcement has on early sensorimotor behaviors. Then, we focus on studies that have established how social interactions, and more important, how imitation may shape early vocal development in prelinguistic infants.

In general, infants exhibit an explicit specialization toward human interaction. They prefer to smell humans, to observe human faces, and to hear human speech. On the other hand, they also display a great talent to imitate facial and manual gestures (Lungarella et al., 2003). Therefore, at early age infants are capable of two different types of imitation. Initially, imitation scenarios are those in which the goal and the current state can be represented in the same modality, such as vocal imitation. Later, the more complex scenario of imitating opaque acts can be considered, in which the agent can neither see nor hear itself performing the imitation, such as facial imitation (Demiris and Meltzoff, 2008).

Infants gradually improve their abilities before reaching a stage where, more than imitating actions, they also imitate underlying intentions and goals of demonstrators (Demiris and Meltzoff, 2008). As mentioned in Tomasello and Carpenter (2007), different from other species, at a very early age human infants are motivated merely to share interest and attention with others. At around nine months of age, infants already engage in joint attentional behaviors by directing others' attention toward objects of interest. By this age, infants know

what others see, and they also start attempting to share attention with others to achieve joint attention, which is not only two agents experiencing the same thing but, simultaneously, knowing they are doing this. Therefore, joint attention is a crucial element in the social development of humans. As also stated in Tomasello and Carpenter (2007), infants are concerned with sharing psychological states with other peers, forming shared intentions and attention with them, and learning from demonstrations produced for their benefit.

As emphasized in some works, e.g., Breazeal and Scassellati (2002) and Lungarella et al. (2003), interaction with adults and peers is a cornerstone for cognitive development in infants. Forms of social support, mimicry, and imitation may play not only a crucial role in the development of early social cognition but, in a broader sense, in the whole cognitive development. As mentioned in Lungarella et al. (2003), evidence points out that, during early developmental stages, adults provide support to help infants bootstrap cognitive, social and motor skills. This support gradually decreases as infants become more confident in their abilities, this mechanism is usually known as *scaffolding*. This mechanism reduces distractions and bias exploratory behaviors toward relevant sensorimotor behaviors. In this sense, the caregiver is also responsible for increasing or decreasing the complexity of tasks to guarantee that the infant learns the most during the sensitive periods in which is more responsive to the caregiver's input. However, as highlighted in Oudeyer et al. (2007), *scaffolding* is just a help, but at the end of the day, infants decide by themselves what they do, what they are interested in, and what their learning situations are.

The perspective of Vygotsky is an important element to understand the relevance of social interactions. Vygotsky argued that linguistically-mediated social interactions cause a radical transformation of elementary cognitive abilities into high-level psychological functions (Mirolli and Parisi, 2011). If the Vygotskian perspective is correct, then what makes human cognition different is not individual brainpower *per se*, but it is rather the ability of humans to learn through other persons and their artifacts, as well as to collaborate with others in collective activities (Tomasello and Carpenter, 2007).

In general, adults tend to teach infants by demonstrating what they should do, and then infants respond to imitating and internalizing what is learned (Tomasello and Carpenter, 2007). Imitation invokes and coordinates the perceptual, representational, memory, and motor neural systems of the infant (Demiris and Meltzoff, 2008). An interesting example discussed in the field of developmental robotics is the onset of *pointing* behavior (Hafner and

Schillaci, 2011, Kaplan and Hafner, 2006). One hypothesis suggests that pointing behavior in young infants initially emerges from the attempt of grasping an object that is out of reach. When the caregiver hands over the requested object to the infant, the pointing gesture is rewarded through social reinforcement (Ramenzoni and Liszkowski, 2016). The pointing behavior plays a relevant role in shared intentionality, as at twelve months of age, infants point for others simply to share interest and attention with them or to point unknown objects to inform others they do not know the objects, even when there is not a benefit for themselves (Tomasello and Carpenter, 2007).

Social Reinforement in Prelinguistic Vocal Development

So far, we have observed that social interactions, especially imitation behaviors, play a key role through the course of early development. In the following, we focus on the study of the relevance of social interactions in prelinguistic vocal development, especially the imitation scenarios observed during this developmental stage.

Many studies have been recently completed by developmental psychologists aimed at understanding the effects of social interaction over early vocal development, especially during prelinguistic interactions. Furthermore, evidence suggests that interactive relationships could be developed on a prelinguistic vocalization framework (Franklin et al., 2014). For instance, it was found in Franklin et al. (2014) that at six months of age, infants are aware of their vocalizations' social value affecting parental engagement. Thus, according to available evidence, in this work, we assume that prelinguistic vocalizations are salient signals to parents, who immediately respond. Moreover, recent studies also suggest that those parental responses may play an important role in vocal development and language acquisition (Goldstein et al., 2009). For instance, the often invoked analogy between human speech and songbirds development was studied in Goldstein et al. (2003). In songbird, imitation is usually considered the mechanism for vocal development, as social contingency provides them opportunities for vocal learning. In this sense, vocal development is socially shaped.

Testing the ability of infants to use social feedback to facilitate developmental transitions, Goldstein et al. (2003) also observed that contingent interactions foster changes in vocal behavior. Their major conclusion is that, simultaneously, babbling regulates and is regulated by social interaction. They also found that different social contingencies may foster changes

in babbling due to social reinforcement as touching (as mentioned in the previous chapter), smiling, and shaking. Later, the results obtained in Goldstein et al. (2003) were extended in Gros-Louis et al. (2006). During naturally occurring interactions, it was found that mothers' vocalizations provide better predictions for infant's vocal utterances compared to other social modalities. In general, adults are sensitive to differences in prelinguistic vocalizations, responding differently to distinctive sounds (e.g., track cries, quasi-voiced vocalizations, voiced 'syllabic', and 'vocalic'). Adults can classify vocalizations of infants in the range between 7 and 11 months of age, even of unfamiliar infants. Adults see infants as 'real talking' when they produce prelinguistic syllabic sounds and respond to this kind of vocalizations with higher frequency. The fact that adults perceive different infant vocal types suggests that maternal responsiveness plays a role in vocal development.

Evidence suggesting that pre-speech vocalizations have a range of pragmatic functions were provided in Oller et al. (2013). However, pragmatic functions were not related by any means to vocalization development. Later, based on experimental results, it was suggested in Gros-Louis et al. (2014) that maternal responses to infants' directed vocalizations contribute to the emergence of vocal usage and the shaping of vocal development. In general, evidence has shown that mothers respond differently according to infants' vocalization directionality (mother-directed, object-directed, and undirected) and acoustic characteristics. Mother's sensitive responding to mother-directed vocalizations was correlated with the increase in developmentally advanced consonant-vowel vocalizations and other language measures (Gros-Louis et al., 2014).

Regarding maternal responsiveness, seven categories of maternal verbal response are distinguished in Gros-Louis et al. (2014): acknowledgments, attributions, directives, naming, play vocalizations, questions and *imitation/expansions*. During the imitation scenario mothers model the word that the sound produced by the infant approximated and expand on it. It was found that imitation in early months of life is a good predictor for an increment in infant mother-directed vocalizations in future months. Infants who received proportionally more responses to their mother-directed vocalizations showed a significant increase in developmentally advanced vocalizations.

Apart from its association with more syllabic vocalizations, the early mother-infant mutual coordinated engagement was also evidenced in Hsu and Fogel (2001). Experiments in

Gros-Louis et al. (2014) provided substantial support to conclude that maternal response contributes to achieving phonologically advanced consonant-vowel sounds and mother-directed vocalizations. It can also be pointed out that prelinguistic communicative behaviors influence caregivers at the moment and over time, showing that the behaviors of infants and caregivers are deeply intertwined.

5.2 Social Reinforcement in Artificial Sensorimotor Development

As in the case of humans, where social development is deeply intertwined with cognitive development, there is a consensus that algorithms for robot learning that build models only with predefined supplied data are unlikely to be successful, at least in unstructured tasks. In general, it has been observed that unguided learning without any support has difficulty for overcoming the complexity involved even in simple tasks (Demiris and Meltzoff, 2008). In this section, we develop on some works aimed at studying how to include social mechanisms within the learning framework in developmental robotics, emphasizing the early vocal development scenario.

First of all, it is important to mention that the architecture presented in this chapter should not be confused with the *reinforcement learning* approach, as defined in Sutton and Barto (1998). Therein, reinforcement learning is defined as a learning process in which an artificial agent learns to map actions to effects as to maximize a numerical reward function (Sutton and Barto, 1998). In the following sections, it will be seen that our approach is socially reinforced in the sense that an external observer rewards specific behaviors in a learner performing an intrinsically motivated exploration. However, the learner is not attempting in any sense to maximize this reward signal nor has an explicit goal, which is an important feature of reinforcement learning.

It has been pointed since many years ago that social interaction is a powerful tool that roboticist may use for transferring important skills, tasks, and information to robots (Breazeal and Scassellati, 2002, Lungarella et al., 2003). As defended in Breazeal and Scassellati (2002), if robots are endowed with social mechanisms, first, it would be easier for humans to transfer knowledge to them, and thus interact with them, helping to achieve one of the primary

objectives of robotics: robots that may be fully integrated into daily life human activities. Secondly, social interactions, in imitation scenarios, provide an interesting mechanism to bias explorations in such a way that search spaces for learning are constrained, which has a direct impact on the central interest of this work: sensorimotor exploration.

From the perspective of social learning applied to artificial agents or robots, one may formulate the question: *Which approaches might be considered as social ones for learning?* Through this work, learning approaches are considered as social approaches if there is at least two embodied agents interacting. Furthermore, the behavior of at least one of the agents must affect the learning trajectory of the other. In this sense, the most common approach for social robot learning in recent years has been the well-known *programming by demonstration* or *learning by imitation* (Calinon, 2009).

Finally, Lungarella et al. (2003) summarizes other important results that show the impact caused by imitative learning. For instance, imitation of the movements of a robotic arm by a human teacher could naturally lead to eye-arm coordination as well as to adequate control of the arm. Nowadays, we know that a human can teach a robot to perform certain types of movements by simply performing them in front of the robot reducing the amount of trial-and-error made by the robot. In other words, for embodied systems to behave and interact in unstructured environments, appropriate coordination of perception and action is necessary. Roboticists, especially those working with a developmental approach, have accepted that action and perception are tightly intertwined. Moreover, the hypothesis that regularities between action and perception can be found and learned as part of a gradual developmental process has also spread.

One of the most interesting approaches for *programming robots by demonstration* is explained in Calinon et al. (2007) and Calinon (2009). In Calinon et al. (2007), the authors presented an approach that generically solves the problems of extracting relevant features for a given task, evaluating how the task should be reproduced, and, finally, finding the optimal controller to generalize the acquired knowledge to several contexts. Calinon and Billard (2007) presented an approach to teach human gestures, specifically gestures with the upper limbs and head, to a robot. The robot incrementally learns those gestures interacting with the human. Therein, two modalities of interaction were considered, on the one hand, the instructor was endowed with a set of motion sensors that recorded its movements. On the other hand, the instructor could also perform movement corrections directly on the robot embodiment, i.e., kinesthetic

teaching. As mentioned in Chapter 3, the incremental learning approach used in Calinon and Billard (2007) and Calinon et al. (2007) was based on incremental learning of Gaussian Mixture Models (GMMs) and the regression problem was solved with Gaussian Mixture Regression (GMR).

From all the social modalities in which infants acquire knowledge from caregivers, developmental psychologists, and therefore roboticist working on developmental robotics, have focused their attention toward the imitation mechanism (Breazeal and Scassellati, 2002). In Demiris and Meltzoff (2008), an exciting study was provided, therein the authors compared available developmental robotic approaches to imitation mechanisms against developmental psychology theories. In Lungarella et al. (2003), the authors examine studies about social interaction and acquisition of social behaviors in robotic systems for a wide range of learning situations and techniques. Interesting areas of research regarding embodied social human-robot interaction, in which the developmental approach may contribute, includes shared and joint attention, low-level imitation, language development, and social regulation (Lungarella et al., 2003).

Social Reinforcement in Artificial Vocal Development

From the developmental robotics perspective, as mentioned in Asada (2016), the developmental process of early vocal sensorimotor learning is assumed to be deeply intertwined with caregiver's feedback. Considering constructivist approaches[1] to emulate early vocal development, there are two distinct ways to assemble self learning and social learning. On the one hand, some approaches consider both learning steps as separated processes. On the other hand, some approaches mix both learning formats from the beginning or do not consider self-learning, but primitives are given beforehand. For instance, some of the most cited works in the field, e.g., Howard and Messum (2011) and Kröger et al. (2009), emphasize that it is somewhat likely that self-learning and social learning occur in parallel, even though in practice they considered that both stages occur in series.

In general, Asada (2016) discussed approaches to whole dynamics of the interaction between an infant and a caregiver. He identified a series of factors that must be considered when

[1] A constructivist model of knowledge attempts to answer the primary question of epistemology, "How do we come to know what we know?" This constructivist model can be summarized in a single statement: Knowledge is constructed in the mind of the learner (Bodner, 1986).

studying interactions through the course of vocal development using developmental robots. Despite the difficulties from a neuroscientific perspective to explicitly handling the role of interactions in development, the essential issue for cognitive developmental robotics is to find a fundamental principle that can be shared by natural and artificial systems. Such a principle should contribute to the acquisition of new insights into early vocal development to reveal how infants learn to vocalize their caregiver's native language, and more generally to understand human cognitive development.

In the following, we visit a series of works that have led the research on the role of social interactions during early artificial vocal development to its current state of the art. Despite a few works within the developmental robotics area that approached the relevance of social mechanisms in early development (Lungarella et al., 2003), Yoshikawa et al. (2003) is one of the earliest works that boosted the integration of cognitive and developmental robotics with infant developmental theories in order to study the role of social interactions during early vocal development.

Yoshikawa et al. (2003) built a real robot aimed at reproducing the human vocal tract and the process in which humans acquire phonemes based on psychological evidence of maternal imitation. As we studied in the previous section, maternal imitation is assumed to reinforce infant vocalizations. For vowel acquisition, the authors considered interactions with a caregiver (human) that repeats the robot's vocalizations. They assumed that the learner does not have any built-in knowledge about the relations between phonemes and its sensorimotor system. Thus, the learner must obtain information through the interactions with its caregiver. In an interaction episode, the caregiver repeats the learner's vocalization using its mature phonemes, thus facilitating the acquisition of vowels by the learner despite its immaturity. Given the different body structures, the learner needs to abstract the observed behavior to some extent since it cannot duplicate it as it is. Yoshikawa et al. (2003) adopted a mechanism for producing random articulations, and the caregiver always utters the vowel that matches the vocalization of the robot if the vocalization can be regarded as a vowel. The proposed learning mechanism extracts clusters considering the statistical consistency in the data and only works if the caregiver tends to be engaged in the repetitive utterances.

The DIVA model, introduced in Guenther (2006), Guenther et al. (2006), is a model that emulates the way in which infants acquire new words using imitation mechanisms consistent with the neurophysiological evidence. After an early babbling stage, when a new speech

sound is presented to the model, it uses its sensory error map and motor cortex's model to generate a time sequence of articulator positions for the Maeda's simulated vocal tract to produce the given speech signal. The DIVA model uses self-organizing networks for processing neural states and build neural maps in order to store phonemic, motor, and sensory states representing speech items. Each time a new acoustic sample of speech is presented to the model, a new cell is recruited into the speech sound map to represent that speech item. Even though the model is inspired by the mechanism by which infants acquire new speech units during imitation, it does not study the environment or context in which those imitations occur. Inspired by Guenther (2006), Guenther et al. (2006), it is presented in Kröger et al. (2009) a neurocomputational production-perception model using a similar architecture. For the imitation part, Kröger et al. (2009) considers imitation training sets and presents them to the model in a similar way as the DIVA model does. Kröger et al. (2009) acknowledged that their approach is only partially consistent as babbling and imitation training items are applied in parallel during the imitation training stage after a short babbling training stage.

Another outstanding series of works that model vocal development considering the social influence of an external caregiver was developed through Howard and Messum (2007, 2011, 2014). In these works, the authors introduced the *Elija* model as mentioned in Chapter 2. Regarding social mechanisms, *Elija* considers an imitation mechanism through the course of vocal learning, based on the mother-child interactions observed in developmental studies. In Howard and Messum (2014), the authors consider the study of an autonomous motor control learning stage (babbling) and later a social learning stage, where *Elija* makes usage of caregivers responses. The experiment included real humans speaking different languages that acted as caregivers of the computational model. The social mechanism of *Elija* considers that infants sound attracts caregivers attention. At this point, both parts are assumed to be aware that the caregiver must regard infant's utterances are equivalent in some way (shared intentionality). The mechanism allows *Elija* to map its motor patterns to caregiver's responses, and the learned speech utterances are used to acquire the name of some objects during a simple interaction scenario (Howard and Messum, 2011).

As in Yoshikawa et al. (2003), in Miura et al. (2012) is studied the vowel acquisition problem when considering imitation between dissimilar embodiments (correspondence problem), as the real scenario between infants and adults. The proposed artificial learner is endowed with the primitives to produce a series of vowels with different levels of clarity for the caregiver's

perception and with a mechanism to learn a set of classifiers of different heard sounds. Each of the classifiers should learn the position and accuracy of a vowel given the caregiver's articulation region. Using an auto-mirroring bias approach the artificial agent learns that the caregiver's response corresponds to one of its own vocalized vowels. In the scenario considered, the learner selects speech utterances according to its expectation to be imitated by the caregiver. The experimental results showed that when a learner is endowed with these mechanisms, it is capable of selecting utterances that resembled clearer vowels which are easier for the caregiver to imitate.

Warlaumont et al. (2013b) presented a model that consists of a bioinspired layer of neurons that controls a speech synthesizer via neuromotor connections. The authors consider the case when neuromotor connections, trained as self-organizing maps, are updated only when reinforcement occurs as a dependence of learning on reinforcement is consistent with neurophysiological works. In general, studies suggest that learning in motor cortex is modulated by a neurotransmitter strongly associated with social reinforcement. Thus, the primary function of reinforcement is to gate the learning of neuromotor connections. This model was agnostic regarding the source of reinforcement but proposes a series of ideas where some auditory features could be integrated as intrinsic reinforcement. Trying to study social reinforcement as a learning modulator using auditory salience as a source of reinforcement, Warlaumont (2013) focuses on a spiking neural network adapted to receive reinforcement when it produced a sound with high auditory salience. Thus, reinforcement was given if the estimated auditory salience of the learner's vocalization was above a threshold. The results indicate that salience based reinforcement can be a high-quality source of feedback in the emergence of canonical babbling. The considered threshold was fixed to be a constant. In contrast to the fixed-threshold salience-reinforced simulations, the authors acknowledge that a human listener is presumably adaptive to the model's performance over time when deciding whether to provide reinforcement or not. Another important idea present in this work, consistent with the works mentioned along this chapter, is that adaptive learner goals and adaptive reinforcement are likely the norm in human infancy and will typically lead to better performance by developmental robots and computational models. For example, human mothers' responses to infant behaviors do appear to change depending on infant vocal repertoire size.

Moulin-Frier et al. (2013) also regarded the role of the social environment as important. In fact, the authors considered a simple scenario to extend intrinsically motivated autonomous

sensorimotor exploration, combining it with emulation behaviors. They considered an emulation scenario where the learner observes the caregiver' outcomes and later employs its sensorimotor knowledge to reproduce those outcomes. In the case of interactions during early vocal development, the learner cannot observe the vocal tract of the caregiver; thus it tries to reproduce the auditory outcome observed by using its means. Moulin-Frier and Oudeyer (2013b) considered a simplified scenario where the caregiver has a finite set of auditory outcomes, and every time the learner chooses to learn by social guidance, it chooses an auditory outcome randomly among the caregiver's repertory. The intrinsic motivation system is used to select when to use the emulation mechanism, as the learner monitors the progress resulting from using the autonomous strategy and the emulation strategy. Then, the learner is capable of choosing the emulation mechanisms if it is more likely to generate learning progress.

Kröger and Cao (2015) presented a deeper analysis of the information generated during learner–caregiver communication and the interaction of both agents with the environment, for example, the pointing gesture. In this architecture, the auditory productions of a caregiver are used to stimulate motor plans considering the knowledge acquired during babbling by the learner. Moreover, during imitation learning and especially with respect to learner-caregiver interaction, strong associations are built upon the one hand for a sensorimotor realization of a syllable, represented by a model neuron as in the DIVA model and Kröger et al. (2009), babbled or imitated by the child and on the other hand for the potential meaning of that syllable. Imitation starts with word productions of the caregiver and leads to a reproduction of this word by the learner. In the case of a reproduction of the word with sufficient quality, the caretaker gives positive feedback. Thus, the model only uses those auditory stimuli accepted by the caregiver for imitation learning. The temporal overlap of babbling and imitation can also be called "guided babbling". Here, babbling patterns become strongly related to the auditory input of the learner and thus babbling is biased here by the target language.

In Forestier and Oudeyer (2017), it is argued that former works using intrinsically motivated sensorimotor exploration, as Moulin-Frier et al. (2013) and Acevedo-Valle et al. (2018), have the disadvantage of not being situated into a physical environment where vocalizations have a meaning related to objects. Thus, Forestier and Oudeyer (2017) studied a scenario where an agent can use dynamical motor primitives to control a simulated robot motor arm and

generate articulatory speech gestures with the Maeda's synthesizer. This work assumes the existence of a caregiver capable of producing speech gestures concatenating three vowels from a social set which contains a total of five vowels. The scenario of interaction considered the existence of three movable objects and a tool that could be handled by the robot arm to move the objects. Therefore there are three mechanisms that the learner can use to move the objects: using the robotic arm directly, handling the tool with the motor arm and moving the objects with it (increasing the reachable space), or asking the caregiver to move the objects (increasing more the reachable space). At the beginning of the simulations, each object is named with a speech gesture by the caregiver. The leaner uses model babbling, as it has to explore two sensorimotor systems. If the agent touches with the arm or the tool one of the objects, the caregiver produces its name, and the word becomes part of a learner's set of imitation gestures. When the learner is babbling with its vocal tract, half of the experiments it makes use sensory goals from the units in the set of imitation gestures. Whether a learner's vocalization is close enough to one of the names of the objects, then the caregiver handle the object to the learner. In general, Forestier and Oudeyer (2017) found a developmental approach that may stand for the linkage between early development of tool use and speech. For the learner in his architecture having the caregiver handling objects through vocalizations is not an explicit goal, but the social interaction emerges from the same drive to refine sensorimotor models. Therefore, the episodes in which the caregiver can understand a learner's production as toy names and make it react and help can be interpreted as an emergent form of social tool use.

In the same line of Moulin-Frier et al. (2013) and Acevedo-Valle et al. (2017a), Acevedo-Valle et al. (2018) and Forestier and Oudeyer (2017), in Najnin and Banerjee (2017) social interaction within intrinsically motivated sensorimotor exploration is included to study early vocal development. Adult vocalizations are provided to the learner, considering two adults, a male, and a female. The authors consider an experiment per adult, allowing the agent to produce 10K imitating vocalizations. Contrary to Moulin-Frier et al. (2013), switching between self-exploration phase and imitation phase in Najnin and Banerjee (2017) is manual. After learning through self-exploration, the agent tries to imitate the speech sound using the learned model. As a consequence, the learner does not exhibit the property of self-organized transition from autonomous learning to socially guided learning as observed in Moulin-Frier et al. (2013).

5.3 Sensorimotor Exploration with *reformulation/imitation* Episodes

So far, through this chapter we have shown that social feedback is an important opportunity for developmental change, producing a two-directed shaping behavior effect, in both the learner and the instructor. Especially, it has been highlighted the fact that social feedback to infant's vocalizations is an underlying mechanism for developmental change. Therefore, to achieve a complete picture of early sensorimotor control learning and exploration, it is important to identify the potential that social interactions of a learner with a skilled instructor under different social circumstances may produce to the progressive emergence of complex behaviors in the learner. In the following, the extension proposed in Acevedo-Valle et al. (2017a) to the architecture presented in Chapter 4 is presented.

The presented architecture is inspired by the social episodes observed between infants and their mothers as reported in Gros-Louis et al. (2014). Specifically, we focused on the study of *imitation/expansions* maternal response as social mechanism shaping sensorimotor exploration. To perform such studies in artificial agents, an instructor – an expert in sensorimotor control (e.g., vocalizing in the case of speech)– is considered. Every time a learner produces a sensory unit similar to a sensory unit relevant to development according to the instructor, then the latter reformulates and produces that sensory unit, so that immediately the learner attempts to imitate the instructor's utterance closing an episode of *reformulation/imitation*.

Starting with the architecture introduced in Chapter 4 for motor constraints aware intrinsically motivated sensorimotor exploration, in this section, we describe an extension initially presented in Acevedo-Valle et al. (2017a) that considers social reinforcement. The goal is to observe the effect that *reformulation/imitation* may produce on the course of developmental learning using artificial developmental agents. We argue that such kind of studies allow roboticists and developmental psychologists to analyze social interactions as elements regulating, i.e., shaping, early sensorimotor exploration. A social instructor was included in the exploration architecture represented in Figure 5.1 to achieve that goal. Such an instructor was assumed to be skilled in sensorimotor control relevant to communication and able to communicate with a learner engaged in sensorimotor exploration. The exploration architecture presented in this chapter, as an extension of the one presented in Chapter 4, is an active learning architecture that mimics the exploration behaviors observed during sensorimotor

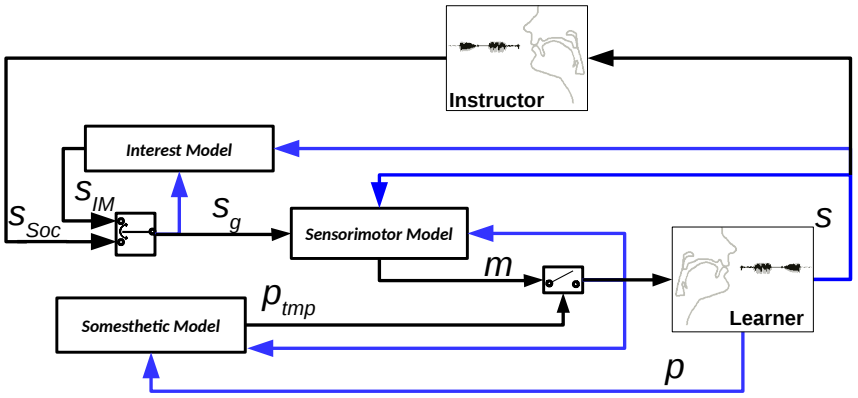

FIGURE 5.1: Diagram of the socially reinforced sensorimotor exploration architecture. Black lines represent the flow of data during each vocalization. Blue lines represent signals used to update the models. Notice that the selector switch indicates that s_g could be generated either directly from the social reinforcement (prioritized) or from the intrinsic motivations mechanism. The simple switch indicates that the somesthetic model might accept or reject a proposed interesting goal.

exploration in biological agents. The architecture is based on goal-directed motor babbling (Rolf et al., 2010). Thus, sensory goals are actively chosen according to either, a model of interest during exploration, or according to the instructor responsiveness through the course of autonomous exploration.

The extended version of the architecture represented in Figure 4.2 to include the *reformulation/imitation* mechanism is shown in Figure 5.1. The notorious difference is the insertion of a social instructor and the relation of the productions of this instructor with the mechanism to choose sensory goals. Interaction occurs when the learner production is 'enough' similar to one relevant to communication. In that case, the instructor perceives this similitude and reformulates with the relevant sensory unit. When the learner perceives an utterance by the instructor, immediately it attempts to imitate that utterance. This reformulation mechanism is similar to the one used in the *Elija* model (Howard and Messum, 2011), which was motivated by the episodes of vocal imitation observed in mother-child interaction.

In the architecture shown in Figure 5.1, the learner starts without any experience producing intended goals nor own constraints knowledge. First, the models are initialized in a first stage, either randomly or using any other criteria. Once they are initialized, the intrinsically motivated exploration begins. Then, looking at Figure 5.1, the interest model in the diagram

can propose new goals that are likely to foster the progress of the competence function, which measures the ability of the agent to reach sensory goals.

Let us suppose that the interest model proposes a goal and the instructor is not currently interacting with the learner. In this case, the selector switches to the signal provided by the interest model. Then, the proposed goal is received by the sensorimotor model; it infers which is the motor action that would produce the self-proposed sensory goal according to the current agent's knowledge. Finally, the motor action is received by the somesthetic model which predicts the nociceptive outcome of executing the proposed motor action according to the current knowledge of the agent.

If the nociceptive prediction indicates that the nociceptive signal will be triggered when executing the proposed motor action, in other words, if the learner suspects that executing the motor action could produce a 'painful' experience, then the simple switch is open, and the learner does not execute the motor command. Thence, the interest model proposes a new goal and the prediction process is repeated. On the other hand, if the nociceptive prediction suggests that there is not a risk when executing the motor action, then the simple switch is closed, and the learner executes the motor action.

When the motor action is executed, sensory outcomes are produced. The salient sensory outcomes are observed by the instructor and the learner itself, whereas the nociceptive outcome is an internal sense of the learner. At this point, the generated data is used to train the models.

Simultaneously, the instructor perceives the salient sensory production of the learner and compares it to the set of sensory units relevant to communication. The instructor selects the more similar unit and, if the Euclidean distance between the sensory production of the learner and the sensory unit relevant to communication is lower than a predefined threshold, then the instructor produces the relevant unit as a reformulation of the learner's original production. At that point, the double switch selects as a sensory goal the signal perceived by the learner from the instructor's production. The nociceptive prediction mechanism is then activated as explained before. If the somesthetic model determines that it is possible to imitate the instructor reformulation without risk of reaching an undesired motor configuration, then the imitation action is executed, finishing the imitation episode. Otherwise, if a 'painful' configuration may occur when imitating the instructor's production according to

the nociceptive prediction, then the interest model starts proposing intrinsically motivated goals again to continue with the exploration. It is important to notice that every time the instructor produces a reformulation, it decreases the social threshold for that sensory unit. In this way, the *reformulation/imitation* episodes will occur if the agent is showing progress toward mastering relevant social units. At this point, the generated data is used to train the models, and the intrinsically motivated exploration process starts again.

In the next section, we introduce the proposed approach to deal with constraints during intrinsically motivated and socially reinforced sensorimotor exploration as an algorithmic architecture. The new details to implement such an architecture, and the main differences with respect the architecture studied in Chapter 4, will also be introduced.

5.4 Architecture Implementation

The cognitive architecture described in the previous section, shown in Figure 5.1, is algorithmically formulated in this section. In the previous chapter, the elements needed to build the experimental setup were explained in detail. In this section, a quick review of the already introduced elements is performed. The new element, the instructor, which depends on the embodied system that is being studied, is described in detail first for the parabolically shaped region, and later for the ear-vocal track system, The explanation would allow any interested reader to reproduce the experiments.

In the following, the elements shown in Figure 5.1 are enlisted and briefly described as they are detailed in Section 4.4.

Embodiment. It represents the physical link between the learner and its inner and outer environment. Three elements are necessary to build an embodiment in the architecture considered. First, a motor system that allows the learner to modify its environment. Secondly, a sensory system that allows the learner to perceive the effects of its action over the outer environment. Finally, a sensory system that allows the learner to perceive 'pain' through a nociceptive signal.

Sensorimotor Model. It is a an Incremental Gaussian Mixture Model (iGMM) as introduced in Chapter 3. For this model, a map f is assumed to exist such that $\mathbf{s} = f(\mathbf{m})$.

The agent can observe $\mathbf{s} + \sigma_{\mathbf{s}}$ for any executed action $\mathbf{m}$. Thus, it is possible to find a GMM, $M_{\mathcal{SM}}$, representing the extended space $SM = \mathcal{S} \times \mathcal{M}$. This model allows us to compute the probability distribution $P(\mathbf{m}|\mathbf{s})$ applying Gaussian Mixture Regression (GMR), and later it is possible to determine which motor command $\mathbf{m}$ is the most likely command to produce a desired sensory goal $\mathbf{s}_g$, thus solving the inverse regression problem $\mathbf{m} = f^{-1}(\mathbf{s}_g)$.

Somesthetic Model For the somesthetic model, we consider the same m-dimensional motor command space $\mathcal{M}$, with $\mathbf{m} \in \mathcal{M}$, and a new binary 'pain' output space $\mathcal{P} = \{0, 1\}$, with $p \in \mathcal{P}$. If the somesthetic system detects that a harmful body configuration has been reached, then a nociceptive signal is triggered and perceived by the agent as pain, then $p = 1$, otherwise $p = 0$. A map g is assumed to exist such that $p = g(\mathbf{m})$ and the agent can observe p for each vocal experiment. Thus, it is possible to find a model $M_{\mathcal{SS}}$, with $\mathcal{X} = \mathcal{M}$ and $\mathcal{Y} = \mathcal{P}$, that allows the prediction $p = g^{-1}(\mathbf{m})$ to determine when a motor command $\mathbf{m}$ is likely to lead to a 'painful' configuration.

Interest Model for Auditory Goals The interest model for sensory goals is an element that endows the learner with the ability to select goals that maximize the expected competence progress in order to improve the quality of its sensorimotor model, resulting in a better control over it. Through this work we use the competence measure used in Moulin-Frier et al. (2013), an later adopted in Acevedo-Valle et al. (2015, 2018), written as

$$c = e^{-|\mathbf{s}_g - \mathbf{s}|}, \tag{5.1}$$

where $\mathbf{s}_g$ is the sensory goal and $\mathbf{s}$ is the actual auditory production after executing a motor command $\mathbf{m} \sim f^{-1}(\mathbf{s}_g)$. In this work, we consider the interest models provided in the `explauto` toolbox.

Instructor The instructor and the learner are endowed with the same embodiment; thus in this work, we do not consider the correspondence problem. A similar embodiment guarantees that the learner can reproduce any production that is made by the instructor. In consequence, the instructor has a sensory system capable of perceiving the effects of the learner's actions. The instructor is assumed to be an expert in producing sensory units relevant to communication purposes that are grouped into the set of sensory units $\mathbf{S}$.

Algorithm 7 Sensorimotor exploration with goal babbling, motor constraint awareness and social reinforcement.

Set $\{n_e, randomseed\}$

1: Initialize $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$
2: Initialize M_{IM} and $i \leftarrow 1$
3: **while** $i \leq n_e$ **do**
4: $\quad p_{tmp} \leftarrow 1$
5: $\quad$ **while** p_{tmp} **do**
6: $\quad\quad \mathbf{s}_{g,i} \leftarrow sample\,(M_{IM})$
7: $\quad\quad \mathbf{m}_i \leftarrow M_{SM}\,(\mathbf{s}_{g,i})$
8: $\quad\quad p_{tmp} \leftarrow M_{\mathcal{SS}}\,(\mathbf{m}_i)$
9: $\quad \mathbf{s}_i \leftarrow f\,(\mathbf{m}_i) + \sigma$ and $p_i \leftarrow g\,(\mathbf{m}_i)$
10: $\quad c_i \leftarrow (1 - p_i * \gamma) e^{-|\mathbf{s}_{g,i} - \mathbf{s}_i|}$
11: $\quad i \leftarrow i + 1$
12: $\quad train\_models()$
13: $\quad \mathbf{s}_{g,i} \leftarrow interaction(\mathbf{s}_i)$
14: $\quad$ **if** $\mathbf{s}_g \neq null$ **then**
15: $\quad\quad \mathbf{m}_i \leftarrow M_{\mathcal{SM}}\,(\mathbf{s}_{g,i})$
16: $\quad\quad p_{tmp} \leftarrow M_{\mathcal{SS}}\,(\mathbf{m}_i)$
17: $\quad\quad$ **if** $!p_{tmp}$ **then**
18: $\quad\quad\quad \mathbf{s}_i \leftarrow f\,(\mathbf{m}_i) + \sigma$ and $p_i \leftarrow g\,(\mathbf{m}_i)$
19: $\quad\quad\quad c_i \leftarrow (1 - p_i * \gamma) e^{-|\mathbf{s}_{g,i} - \mathbf{s}_i|}$
20: $\quad\quad\quad i \leftarrow i + 1$
21: $\quad\quad\quad train\_models()$

function $interaction(s)$

Define $\mathbf{S}$, $\mathbf{th_S}$, α_{th}

1: $dist = |\mathbf{s} - \mathbf{s}_s|$ **for** $\mathbf{s}_s \in \mathbf{S}$
2: **if** $min(dist) < \mathbf{th_S}[argmin(dist)]$ **then**
3: $\quad \mathbf{th_S}[argmin(dist)] * = \alpha_{th}$
4: $\quad$ **return** $\mathbf{S}[argmin(dist)]|$
5: **else return** $null$

5.4.1 Algorithm for Socially Reinforced Sensorimotor Exploration

Algorithm 7 corresponds to the cognitive architecture in Figure 5.1. The first part of the pseudo-code corresponds to the learner. Besides, the *interaction* function represents the behavior of the instructor, which can produce sensory units from a set of sensory units $\mathbf{S}$ relevant to communication purposes. The algorithm for socially reinforced exploration with goal babbling and motor constraint awareness starts with the learner having no sensorimotor control experience.

First, models $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$ are initialized in line 1 using arbitrary motor commands with small values around the neutral motor system position. In line 2, model M_{IM} is initialized using the sensory results obtained in the first line as a sensory goal.

Then, in line 6 of Algorithm 7 the agent selects a sensory goal $\mathbf{s}_{g,i}$ for the next sensorimotor experiment according to the interest model M_{IM}. With $\mathbf{s}_{g,i}$, in line 7, the sensorimotor model is used to obtain the motor command $\mathbf{m}_i$ that according to the current knowledge of the agent would produce $\mathbf{s}_i = \mathbf{s}_{g,i}$.

Given the motor command $\mathbf{m}_i$, $\mathcal{M}_{\mathcal{SS}}$ predicts the nociceptive or 'pain' prediction p_{tmp} for $g(\mathbf{m}_i)$ in line 8. That prediction indicates if the selected motor command is likely to trigger the nociceptive signal, thus causing the agent a 'painful' experience. If the pain prediction indicates that the signal p will be triggered when executing $\mathbf{m}_i$, then the agent rejects the goal, the simple switch in Figure 5.1 is open; therefore the motor command is not executed. Afterward, the interest model proposes a new goal and the prediction process is repeated until a 'safe' goal is obtained. On the other hand, if the 'pain' prediction suggests that there is no risk when executing the motor action $\mathbf{m}_i$, then the simple switch in Figure 5.1 is closed, and the agent accepts and executes $\mathbf{m}_i$.

Next, the motor command $\mathbf{m}_i$ is executed by the motor system. Afterwards, the agent observes $\mathbf{s}_i$ and p_i in line 9. In line 10, the learner evaluates the competence value c_i, which receives a penalization according to the parameter $\gamma = 0.7$ if the agent perceives 'pain' ($p_i = 1$). Finally, in line 12, the training function for models is called, and each model ($M_{\mathcal{SM}}$, $M_{\mathcal{SS}}$ and M_{IM}) is updated.

Different from Algorithm 6, in Algorithm 7, the instructor also observes the salient outcomes $\mathbf{s}_i$ produced by the learner, as indicated in line 13. As the instructor perceives $\mathbf{s}_i$, the function *interaction* is called. In function *interaction*, $\mathbf{th}_{\mathbf{S}}$ is a vector representing the social threshold for each of the sensory units in $\mathbf{S}$, whereas $\alpha_{th} \in [0, 1]$ is a scaling factor that multiplies the corresponding social threshold of a unit when it is selected within an interaction episode.

Within the *interaction* function, first $\mathbf{s}_i$ is compared to the set of sensory units relevant to communication $\mathbf{S}$ in line 1. Then, the instructor selects $\mathbf{S}[argmin(dist)] \in \mathbf{S}$, the more similar sensory unit. If the Euclidean distance between $\mathbf{S}[argmin(dist)]$ and s is lower than the corresponding threshold $\mathbf{th}_{\mathbf{S}}[argmin(dist)]$, then the instructor produces $\mathbf{s}_{IM} = \mathbf{S}[argmin(dist)]$ as a reformulation of $\mathbf{s}_i$ toward the learner, otherwise the *interaction* function returns a *null* answer.

If the *interaction* function returns a null value, then the learner continues the autonomous sensorimotor exploration in line 4. Else, *interaction* returns a sensory goal, the learner

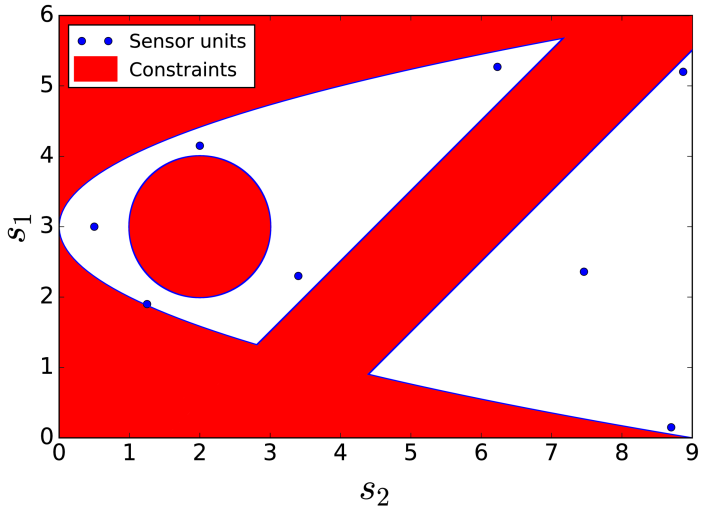

FIGURE 5.2: Parabolic shaped constrained region including sensory units relevant to communication.

perceives $\mathbf{s}_{IM}$ and chooses this sensory value as its next sensory goal $\mathbf{s}_{g,i}$ as indicated in line 13 of the general algorithm. Then, the sensorimotor model $M_{\mathcal{SM}}$ receives the sensory goal and predicts the motor command $\mathbf{m}_i$ that likely will produce the desired sensory output. Once the $\mathbf{m}_i$ is computed, the somesthetic model predicts if the execution of this motor command is likely to produce an undesired motor configuration. If the nociceptive prediction p_{tmp} from line 16 is close to 1, then the agent does not attempt to imitate the reformulation produced by the instructor. However, if p_{tmp} suggests that there is not a risk in executing m_i, then the learner executes the motor command attempting to imitate the instructor reformulation in line 18, at this point the learner perceives $\mathbf{s}_i$ and p_i, the competence is computed in line 19, finishing the *reformulation/imitation* episode. Finally, in line 21, the training function for models is called again, and the learner continues the autonomous sensorimotor exploration in line 4.

In the following, specific details of each embodiment considered for experimentation, especially those related with the instructor, are detailed.

5.4.2 Parabolic Shaped Region System Embodiment

The parabolic shaped region system, shown in Figure 5.2, was introduced in Section 3.7 and endowed with a nociceptive mechanism in Section 4.4.2. Regarding the blue marks in Figure 5.2, they represent sensory units laying intentionally close to the system constraints.

TABLE 5.1:
Sensory units relevant to communication in **S** for the parabolic shaped region system.

	m_1	m_2	s_1	s_2
1	1.9	4.1180	1.9	1.25
2	4.15	4.4142	4.15	2.
3	2.3	4.8439	2.3	3.4
4	5.27	5.4960	5.27	6.23
5	0.15	5.9496	0.15	8.7
6	2.36	5.7313	2.36	7.46
7	5.2	5.9783	5.2	8.87
8	3.0	3.7071	3.0	0.50

TABLE 5.2: Formant frequencies of German vowels (Hz).

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

An instructor able to produce those sensory units is assumed and units are assumed to be relevant to communication. The numerical values for those constraints are shown in Table 5.1.

5.4.3 Ear and Vocal Tract Embodiment

The ear-vocal tract system was introduced in Section 4.4.3. For the social interaction mechanism, in the case of prelinguistic development, this work considers an instructor with and identical embodiment as the explained in Section 4.4.3. The instructor is capable of producing coarticulated vocalizations concatenating vowels similar to German vowels. The German vowels used as a reference were taken from Birkholz (2013) and are shown in Table 5.2.

In order to synthesize these vowels with the *Maedas*' synthesizer, a constrained non-linear optimization problem was formulated for each one of them and solved using the `fmincon` function available in Matlab®. The optimization problem was solved using the static vocal tract `diva_synth` from Guenther Lab. Thus, the optimization problem for each vowel can

be written as

$$\begin{aligned} \underset{\mathbf{m}}{\text{minimize}} \quad & e_s(\mathbf{m}) = |\mathbf{s} - \mathbf{s}_v| \\ \text{subject to} \quad & \mathbf{s} = vt_s(\mathbf{m}) \\ & -3 \leq m_i \leq 3,\ i = 1, \ldots, 10, \\ & m_i = 1,\ i = 11, 12, 13, \\ & \min a_f > 0.01. \end{aligned} \tag{5.2}$$

In the optimization problem represented in Eq.(5.2), $vt_s(\mathbf{m})$ represents the first two formant frequencies produced by the Maedas' synthesizer when executing motor command $\mathbf{m}$, vector $\mathbf{s}_v$ is composed of the two formant frequencies of the vowel that is being synthesized, m_i is the i-th element of the motor command vector $\mathbf{m}$, and finally $min(a_f)$ is the minimum values of the cross section of the area function of the vocal tract, it is chosen a minimum value to this minimum, because it was observed that when $min(a_f) \sim 0^+$, even when the formant frequencies are consistent, the synthesized speech signal is not congruent. At the same time, to guarantee phonation, the motor commands related to phonation $m_{11,12,13}$ are set to 1.

As the optimization problem includes many local minima, it was run several times using different initial seeds for each vowel, and the best solution was taken. In Table 5.3, the final results regarding formant frequencies are shown. Figure 5.3 shows graphically the synthesized vowels. As it can be corroborated, these sensory units are the same units considered for evaluation in the previous chapter (see Figure 4.7).

Once the motor commands are obtained through the optimization process, the seventeen vowels are recombined to generate coarticulated movements using the embodiment explained in Section 4.4.3 and showed in Figure 4.3. Considering the case of unarticulated gestures, when the motor action of one of the time windows indicated in Figure 4.3 is set to zero, the number of sensory units relevant to communication is 323.

5.5 Sensorimotor Exploration Results

In the previous chapter, results showed the advantages of considering constraints when using intrinsically motivated sensorimotor exploration architectures. The architecture from

TABLE 5.3: Formant frequencies of German vowels. F_1 scaled to 500 Hz. F_2 scaled to 1500 Hz.

	F_1	F_2	$\tilde{F}_1$	$\tilde{F}_2$		F_1	F_2	$\tilde{F}_1$	$\tilde{F}_2$
/a:/	1.432	0.7893	1.3215	0.7958	/a/	1.388	0.8627	1.2910	0.862
/e:/	0.692	1.4813	0.5674	1.4658	/E:/	1.052	1.2787	1.0520	1.2787
/i:/	0.53	1.4527	0.5300	1.4527	/y:/	0.548	1.136	0.6206	1.0972
/o:/	0.674	0.403	0.4968	0.4701	/O/	1.068	0.6193	1.1117	0.6238
/u:/	0.576	0.4187	0.6820	0.4493	/U/	0.81	0.634	0.8100	0.6340
/2:/	0.632	0.874	0.6691	0.8740	/@/	0.87	1.076	0.8700	1.076
/I/	0.812	1.236	0.812	1.2427	/E/	1.064	1.2393	1.0640	1.2393
/Y/	0.792	0.868	0.792	0.8680	/9/	1.002	0.8893	1.002	0.8893
/6/	1.278	0.9253	1.2316	0.9175					

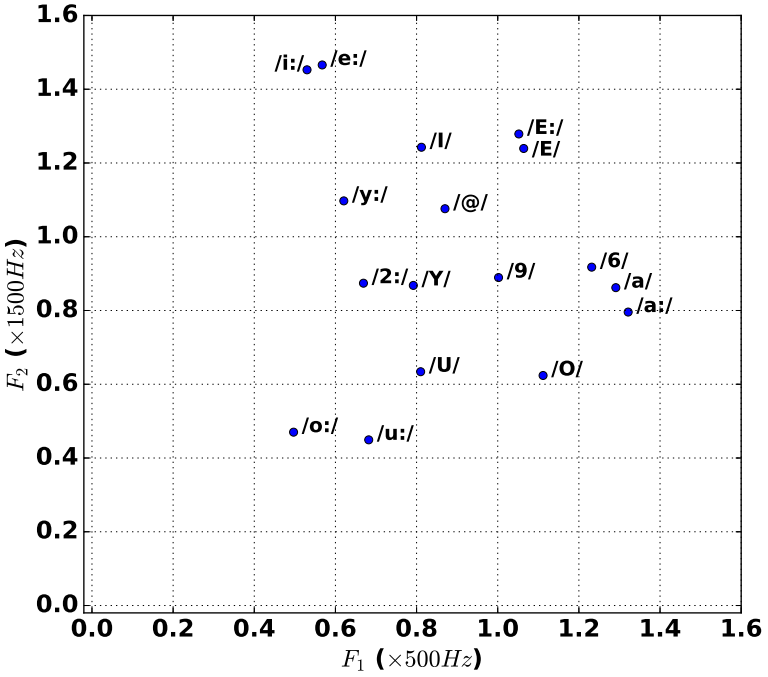

FIGURE 5.3: Fixed vocalizations obtained with `divapy`. Single auditory results for vocalizations with no 'painful' configurations used to generate 323 co-articulated gestures.

Moulin-Frier et al. (2013) was modified and served to test the hypothesis that using somesthesis to acquire motor constraint awareness would foster sensorimotor control development. In this chapter, we compare Algorithm 6 against the new proposed Algorithm 7 in order to study the advantages reported in Acevedo-Valle et al. (2017a) about including a simple *reformulation/imitation* social mechanism into a constraints aware sensorimotor exploration architecture. In the following, first, we present experimental results obtained when applying the exploration architectures to our toy example, the constrained parabolic shaped region, described in Section 4.4.2. Later, we present the experimental results obtained with the exploration architectures for the ear-vocal tract system described in Section 4.4.3.

TABLE 5.4: Simulation parameters for Algorithm 7. Parabolic Shaped Region (PSR). Ear-Vocal tract (E-VT).

Parameter	Name	PSR	E-VT
n_e	number of experiments	$10K$	$100K$
randomseed	random seed		
M_{SM}	sensorimotor model	iGMM	iGMM
K_{min}	minimum number of Gaussian components	3	3
ΔK_{max}	maximum increment of Gaussian components	5	10
K_{max}	maximum number of Gaussian components	20	30
α_{SM}	forgetting rate	0.2 to 0.05	0.2
$train_{SM}$	training step	100	400
M_{SS}	somesthetic model	wNN ($k = 3$)	wNN ($k = 3$)
M_{IM}	interest model	*discretized progress*	*tree*
$\mathbf{th_S}$	Similarity threshold for each sensory unit	all 0.3	all 0.5
α_{th}	Threshold scaling factor after social episode	$1, 0.99$	$1, 0.96$

5.5.1 Simulation Parameters

In general, the simulation parameters used for the experiments in this section are the same ones that were employed in the previous chapter. The parameter values are the result of the tuning process made through Acevedo-Valle et al. (2015), Acevedo-Valle et al. (2017a), Acevedo-Valle et al. (2018). Table 5.4 summarizes all the parameters that must be defined in order to run Algorithm 7, different from the parameters in Table 4.1. This time some parameters should be defined for the instructor agent.

Parabolic Shaped Region System

Regarding the chosen parameters for simulations with the parabolically shaped region system, they were chosen according to the discussion in Section 4.5.1. Handcrafting mechanisms were used in order to choose the sensory unit threshold $\mathbf{th_S}$ and the scaling factor α_{th} that will scale the sensory unit threshold for a unit after this unit is used in a *reformulation/imitation* episode. Several simulations were run using different values for these parameters as reported in Acevedo-Valle et al. (2017a). In the case of $\mathbf{th_S}$, it was observed that large values caused an important increment on the ratio of social experiments against intrinsically motivated experiments. When $\mathbf{th_S}$ was chosen to be small, the ratio of imitation episodes decreased considerably, and the behavior of the developmental trajectory was somewhat similar to the non-socially reinforced (autonomous) architecture.

Regarding α_{th}, its value was also chosen after running several simulations, differently from $\mathbf{th_S}$, we could not find a direct relationship between the developmental trajectory and α_{th}.

Ear-Vocal Tract System

Similar to the chosen simulation parameters for the parabolically shaped region system, a description of the parameters chosen for the simulations with the ear-vocal tract system can be found in Section 4.5.1. To tune the parameters for the instructor, in this problem a similar procedure to the toy example was performed. Different values for $\mathbf{th_S}$ and α_{th} were tested. Several simulations showed that a good value for the thresholds in $\mathbf{th_S}$ was 0.5, in order to achieve a good trade-off between the ratio of interactions and an improvement in the evaluation error against the dataset containing the sensory units relevant to communication. Regarding α_{th}, similar to the case of the toy example, we could not establish a direct relationship between the exploration performance and α_{th}. However, after running some preliminary tests, 0.96 was chosen as value to observe the behavior of the exploration when this parameter is varied.

5.5.2 Parabolic Shaped Region System

The procedure to minimize randomness in order to obtain significant results is the same as in the previous chapter. A large number of simulations were run considering fifty random seeds. Thus, for each set of chosen parameters, fifty explorations changing the random seed were run. Five scenarios are considered to experiment. First, the constraints aware architecture without social mechanism corresponding to the results in the previous chapter to use it as a baseline because its performance was the best in Chapter 4. Later, we consider the social mechanism in four different scenarios: on the one hand, when somesthesis is considered for $\alpha_{th} = 1$ and $\alpha_{th} = 0.99$, and on the other hand when somesthesis is not considered for $\alpha_{th} = 1$ and $\alpha_{th} = 0.99$. Each simulation consists of 100 experiments to initialize $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$, 100 experiments to initialize M_{IM} and 10K exploratory experiments, including both the intrinsically motivated experiments and the social experiments that may emerge during the exploration. The sensorimotor model in each simulation is evaluated against two data sets during the exploration. First, after the initialization of the models, then after every 500 samples during exploration, and finally, at the end of each simulation. The first dataset is

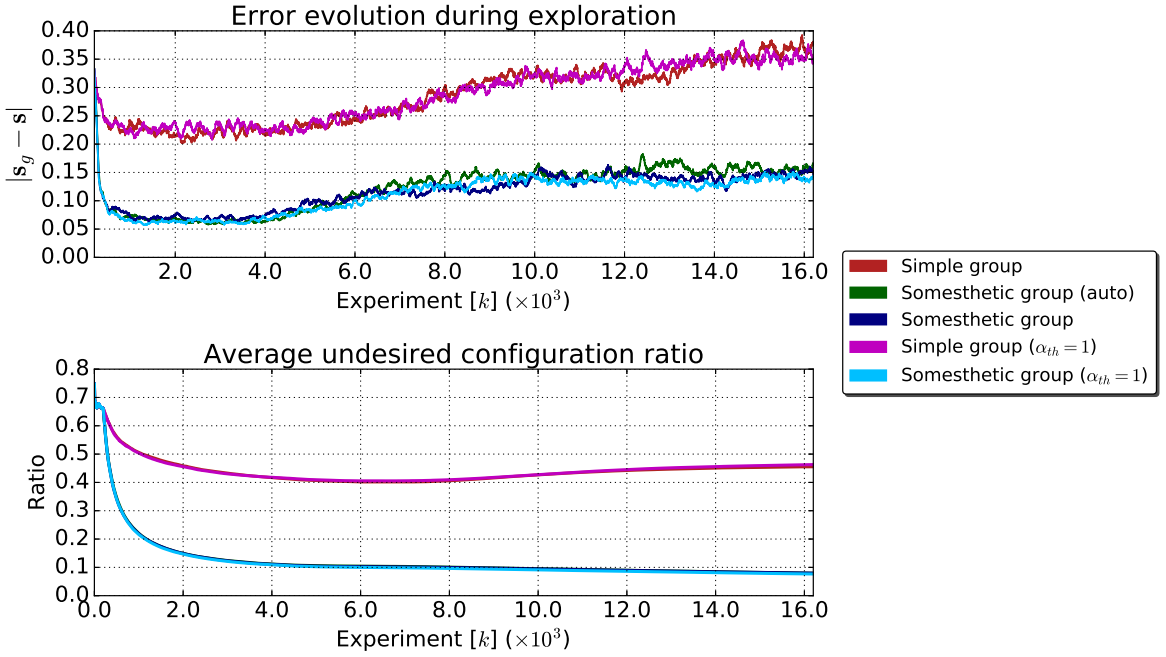

FIGURE 5.4: Results along the exploration running Algorithm 7 using the parabolic shaped region system. If not indicated $\alpha_{th} = 0.99$, and (auto) indicates that it was an autonomous exploration with Algorithm 6.

composed of 441 points evenly distributed along the allowed region of the parabolic shaped region which is shown in Figure 3.7. The second dataset contains the socially relevant sensory units shown in Figure 5.2 and Table 5.1.

Figure 5.4 shows the average results during the exploration for the five groups of simulations that were run. The plots were obtained using the average of the 50 simulations considered for each group of simulations. On the one hand, Figure 5.4 (upper) shows the average norm of the error $|\mathbf{s}_g - \mathbf{s}|$ during exploration considering a moving average window of 100 samples, defined in Eq. (4.3). On the other hand, Figure 5.4 (lower) shows the average undesired motor configuration ratio along the exploration $ucr_{av,expl}$ defined in Eq. 4.4.

Figure 5.5 shows the average ratio of *reformulation/imitation* episodes along the exploration eri_{av} defined as:

$$eri_{av}(k) = \frac{1}{n_{rs}} \sum_{i=0}^{n_{rs}} \left[\frac{1}{k} \sum_{j=0}^{k} [interaction(\mathbf{s}_{i,j}) \neq null] \right], \tag{5.3}$$

for the k-th experiment, where n_{rs} is the number of random seeds and $interaction(\mathbf{s}_{i,j})$ is the reinforcement provided by the instructor in the j-th experiment when simulating with

the i-th random seed.

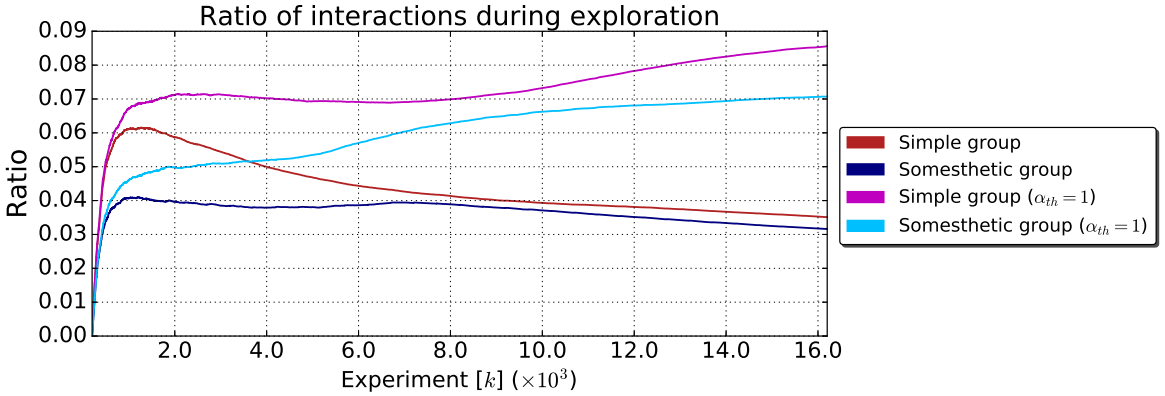

FIGURE 5.5: Interactions along the exploration running Algorithm 7 using the parabolic shaped region system. If not indicated $\alpha_{th} = 0.99$.

Results in Figures 5.4-5.5 suggest that the best performance is obtained when the somesthetic mechanism and the imitation mechanisms are considered together. In general terms, the ratio of undesired motor configurations during exploration and evaluation are lower, and the exploratory and evaluation errors, for both evaluation datasets, are noticeably smaller.

During exploration, looking at Figure 5.4, results obtained in Chapter 4 are corroborated. In all the considered simulation scenarios, it is observed a significant decrement of 'painful' configurations when somesthesis is considered. The rate of decrement is considerably larger for the group endowed with the somesthetic mechanism, regardless of whether they were endowed with the social mechanisms or not. Regarding the behavior of the exploratory error, as observed in Chapter 4, somewhere between the samples 4K and 6K, it tends to increase until reaching a sort of steady value. However, for the agents without somesthesis, it continues to increase and also coincides with a slight increment of the ratio of undesired contacts. We argue that after the learners take advantage of regions where learning may be easy, they are pushed to explore regions where the error increases to keep the progress in competence. As it is observed in Figure 5.4, for learners endowed simultaneously with the social and somesthetic modalities, the values of the exploratory error are predominately smaller than for the autonomous learners.

In Chapter 4, we argued that the error increased because the agents have already explored the regions that permit a high rate of progress in competence values and start to exploit regions where the progress is harder due to their closeness to constrained regions. Thus, there is a slight increase of undesired configurations in the simple agents, whereas the agents endowed

with the somesthetic mechanism continue decreasing the rate of 'painful' configurations but at a considerably smaller rate. On the other hand, the groups of agents endowed with the social mechanism showed better performance regarding the evolution error when the error starts increasing. The improvement in performance may be attributed to the location of the sensory units relevant to communication. Those units are close to constrained regions; thus the interaction with the instructor encourages the agent to explore hard regions of the sensory space earlier in its life achieving a more ordered incremental learning. To the observer, it might appear that all the agents, depending if they were endowed with the somesthetic mechanism or not, threw the same results regarding the evolution of the ratio of undesired motor configurations, however numerically there is a slight difference reported in Table 5.5.

Regarding the ratio of *reformulation/imitation* episodes, in Figure 5.5, we observe that agents endowed with somesthesis, in general, interact less with the instructor than the agents without somesthesis. A smaller number of interactions is due to the fact that those agents without the nociceptive feedback attempt to imitate any production produced by the instructor, regardless whether the imitation act could produce a 'painful' outcome. On the other hand, agents endowed with somesthesis recall information from the sensorimotor and somesthetic models. Thus, before imitating the instructor's production, the learner predicts if the imitation act is likely to produce a 'painful' situation, and if that is the case, then the learner prefers not to imitate the reformulation provided by the instructor. On the other hand, instructors considering a scaling factor for the social threshold $\alpha_{th} < 1$ produced significantly smaller ratios of social episodes compared to those instructors with $\alpha_{th} = 1$. However, the decrease of interactions did not considerably affect the evolution of the exploratory error nor the ratio of undesired motor configurations.

As socially relevant units are close to constraints, increments in imitations strengthen our hypothesis that the agent is exploring close to constrained regions as shown in Figure 5.5, at least in the case for $\alpha_{th} = 1$. However, it is also important to observe that when $\alpha_{th} = 0.99$, as the ratio of interactions decreases, then the expected results for the current experimental setup is that learners interacting with instructors considering $\alpha_{th} = 0.99$ will have a similar behavior to the autonomous agents in the long term as observed in Figure 5.4. In that figure, it is observed how, after 10K experiments, the exploratory error $|\mathbf{s_g} - \mathbf{s}|$ for the autonomous simulations and the simulations with $\alpha_{th} = 0.99$ keeps higher than for the simulations with $\alpha_{th} = 1$ when comparing social groups.

In the following, we study the results considering the evaluations against the two considered evaluation datasets. Figure 5.6 was obtained from the data generated during the evaluations performed over the sensorimotor model through the simulations using the dataset containing data distributed along the whole reachable space of the shaped parabolic region (Whole dataset). The figure was obtained averaging the results of the 50 simulations run for each group. First, Figure 5.6 (upper) shows the average mean evaluation error e_{av} defined in Eq. 3.12. Secondly, Figure 5.6 (lower) shows the average ratio of undesired motor configurations ucr_{av} defined in Eq. 4.6. In Figures 5.6-Figures 5.7, the size of the round markers in the plot is proportional to the standard deviation between simulations.

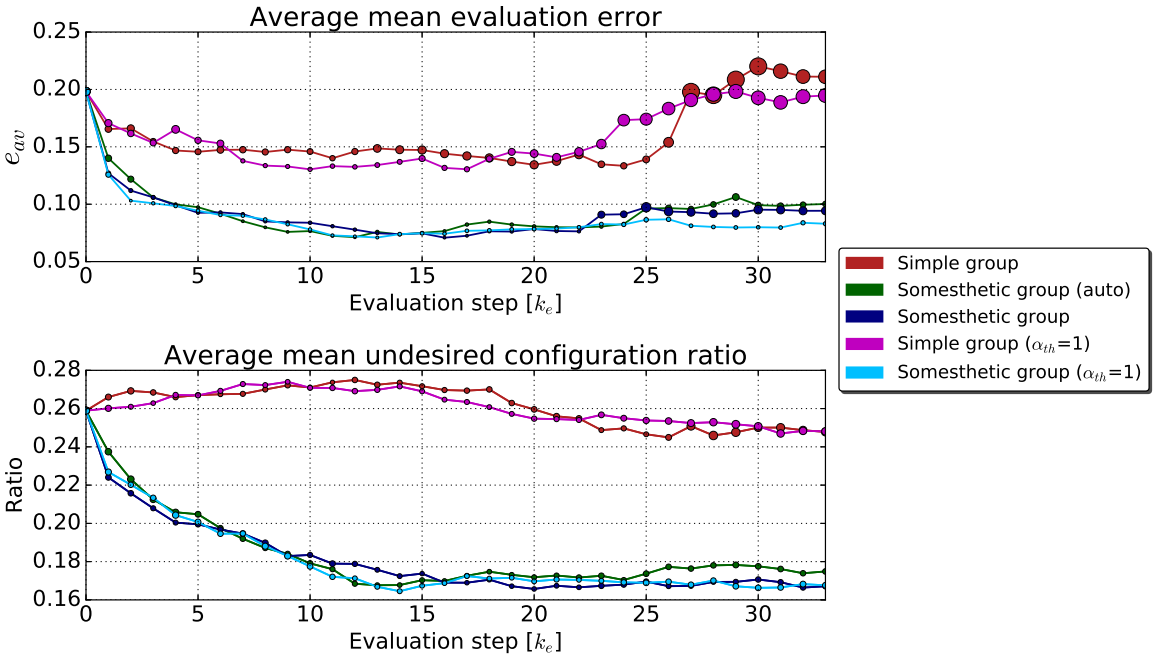

FIGURE 5.6: Evaluation evolution against the dataset evenly distributed along the reachable space of the parabolic shaped region system, running Algorithm 7. If not indicated $\alpha_{th} = 0.99$, and (auto) indicates that it was an autonomous exploration with Algorithm 6.

Similarly to Figure 5.6, Figure 5.7 was obtained from the evaluations performed over the sensorimotor model through the simulations using the data set containing sensory units relevant for socialization purposes (Social dataset). First, Figure 5.7 (upper) shows the average mean evaluation error e_{av}. Secondly, Figure 5.7 (lower) shows the average ratio of undesired motor configurations for the evaluation dataset ucr_{av}.

Finally, Table 5.5 shows some values of interest to analyze the results. The first row displays the values for the undesired motor configuration ratio through the simulation. Afterward, relevant values for the evaluation against the Whole and the Social datasets are shown: the

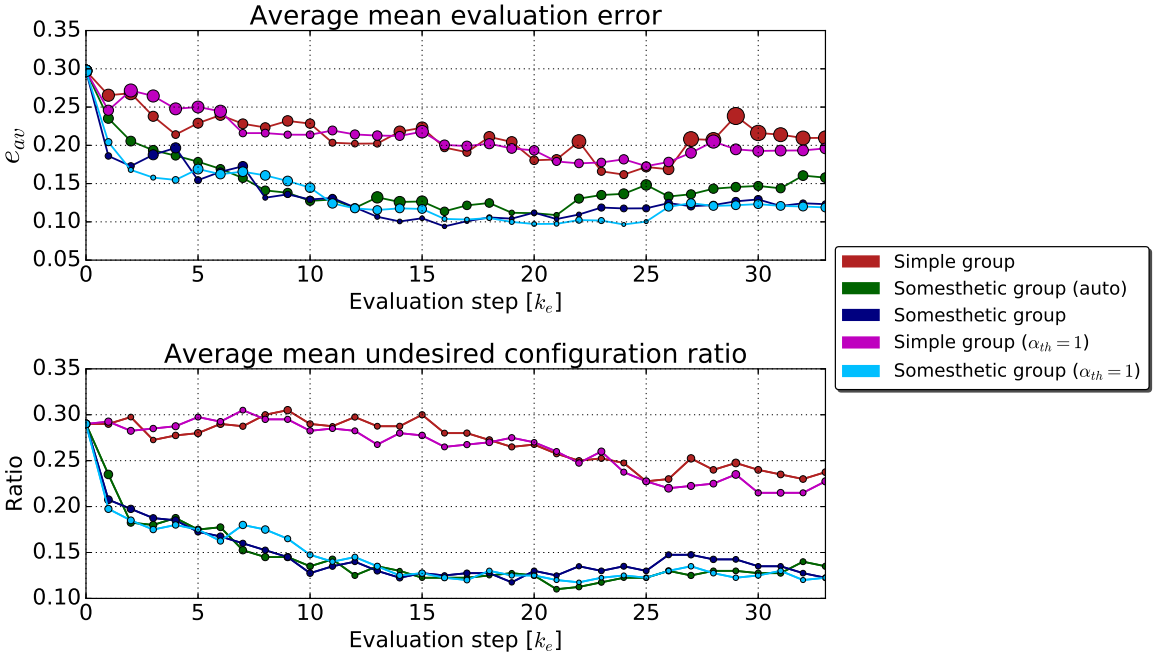

FIGURE 5.7: Evaluation evolution against the sensory units relevant for socialization pusposes of the parabolic shaped region system, running Algorithm 7. If not indicated $\alpha_{th} = 0.99$, and (auto) indicates that it was an autonomous exploration with Algorithm 6.

minimum average mean evaluation error achieved per each group of simulations $\min e_{av}$, the average ratio of undesired motor configurations when achieving $\min e_{av}$, and the evaluation step in which $\min e_{av}$ was achieved.

TABLE 5.5:
Simulation results for the constrained parabolic shaped region system using Algorithm 7.

	Algorithm 6		$\alpha_{th} = 1$				$\alpha_{th} = 0.99$			
	Somesthetic group		Simple group		Somesthetic group		Simple group		Somesthetic group	
	value	std	value	std	value	std	value	std	value	std
$ucr_{av,expl}$	0.0792	–	0.4629	–	**0.0766**	–	0.4553	–	0.0796	–
Whole evaluation dataset										
$\min e_{av}$	0.0713	0.0321	0.1304	0.0566	0.0710	0.0317	0.1334	0.0953	**0.0708**	0.0274
ucr_{av} for $\min e_{av}$	0.1684	0.0289	0.2708	0.0276	**0.1668**	0.0274	0.2497	0.0293	0.1690	0.0245
Evaluation Step	12		**10**		13		24		16	
Social evaluation dataset										
$\min e_{av}$	0.1088	0.0677	0.1728	0.1236	0.0966	0.0437	0.1618	0.1109	**0.0941**	0.0523
ucr_{av} for $\min e_{av}$	**0.11**	0.0776	0.2275	0.0990	0.125	0.0661	0.2475	0.0810	0.125	0.075
Evaluation Step	21		25		24		24		**16**	

NOTE: The table shows, in order of appearance, the average of undesired motor configuration ratio during the exploration, and considering both evaluation datasets, the minimum value of the average mean evaluation error $\min e_{av}$, the average ratio of undesired motor configurations obtained for $\min e_{av}$, and finally, the evaluation step in which this minimum was achieved.

Figures 5.6-5.7 and Table 5.5 corroborates what Figures 5.4-5.5 suggested. The best results

are obtained when the somesthetic mechanism is used together with the *reformulation/imitation* mechanism. Regarding the evolution of the average mean evaluation error e_{av} along the exploration against both datasets, in Figures 5.6-5.7, we observe that both evaluations evolve similarly, and they are similar to the one observed in Chapter 4, at the beginning the evaluation errors are large for all the scenarios. However, the standard deviation of the learners interacting with instructors and endowed with somesthesis are smaller than the others. As the agent explores its sensorimotor system, the evaluation error and standard deviation decrease with different rates for all the scenarios. When the somesthetic mechanism is considered, we observe that the improvement is steadier and faster than without the mechanism. For the toy example, results suggest that the capacity of the learners to acquire knowledge along all the regions of the reachable space depends more on the somesthetic mechanism than on the social reinforcement mechanism.

There is an important emerging question: *What happens during the evaluations after the performance reaches a minimum (considering the evaluation error)?* We observe that the evaluation errors tend to increase after a minimum is reached. This increment of the error is more notorious in agents without somesthesis, and less notorious for the socially reinforced agents with $\alpha_{th} = 1$. As it was mentioned before, it was expected to observe similar behaviors in the long term for simulations with $\alpha_{th} < 1$ and simulations without the social mechanism, as we observe in the evaluation behaviors again. The increment observed in the error is attributed to the ratio of undesired motor configurations during exploration indicated in Figure 5.4 which causes the overall knowledge that the learner has of its sensorimotor system to degrade. As the agent explores constrained regions, the redundancy of the sensorimotor knowledge increases, which makes it more difficult for the sensorimotor model to represent all the knowledge and retrieve accurate motor command inference over the whole sensorimotor space.

It should be determined whether the scaling factor α_t was relevant for the final results or not. Even though the results for all the scenarios were very similar, the large number of run simulations guarantees a certain degree of conservativeness. In fact, at this point is not easy to provide an absolute conclusion regarding α_t. First, looking at Figure 5.4 does not give any clear clue. In Figure 5.5, it is observed that the instructor must work less to achieve similar results through the development of the agent. However, considering that the autonomous learner also achieves a similar result without social feedback, this is not an

argument sufficiently strong to make a conclusion. Besides, looking at Table 5.5, it is observed that in the evaluation scenario with the Whole dataset, the minimum error is achieved by the agents with $\alpha_{th} = 1$ only after 13 evaluation steps (6K), whereas for the evaluation against the socially relevant units, the minimum is reached in 16 evaluations steps (7.5K samples) by the learners interacting with instructors considering $\alpha_{th} = 0.99$. Thus, these learners achieve to master the socially relevant units considerably earlier than the other groups of agents. Looking at Figure 5.7, it is observed that for the agents with $\alpha_{th} = 0.99$, the evaluation error and its standard deviation are decreasing until they reach their best behaviour after 16 evaluation steps, from that point the group of agents starts a sort of transient behavior that degrades their performance, as all the groups of agent do after hitting their best behavior. The relevant conclusion at this point is that $\alpha_{th} = 0.99$ allowed faster and better performance for a social development scenario with the toy example.

5.5.3 Ear-Vocal Tract System

In this chapter, experimentation for the ear-vocal tract system is divided into three groups. First, agents endowed with the somesthetic mechanism proposed in the previous chapter and simulated according to Algorithm 6, as their advantage over the agents without this mechanism was established. The second and third groups correspond to agents endowed with the social mechanism for imitation. They are simulated according to Algorithm 7, one group considers $\alpha_{th} = 1$ and the other $\alpha_{th} = 0.96$. On the other hand, different from the previous chapter, the initialization criteria for the interest model M_{IM} considers all the sensory results obtained when initializing the somesthetic model $M_{\mathcal{SS}}$ and the sensorimotor model $M_{\mathcal{SM}}$ as sensory goals. Based on the results obtained in the previous chapter, we concluded that initializing M_{IM} only with goals that did not trigger the nociceptive signal p during the initialization of $M_{\mathcal{SM}}$ performed worse when evaluating against $\mathbf{S}_{eval}$.

For each scenario of simulation, eighteen different random seeds were considered to generate random initialization sets of motor commands from uniform distributions. In total, fifty-four independent simulations were run, thirty-six using Algorithm 7 and eighteen for the group of simulations with Algorithm 6. All simulations consisted of 100K vocalizing experiments plus 2K initialization vocalizing experiments. The limits for the values of initializing motor commands related to the vocal tract articulators were $[-1, 1]$, whereas for motor commands related to the phonation parameters were $[0, 0.7]$.

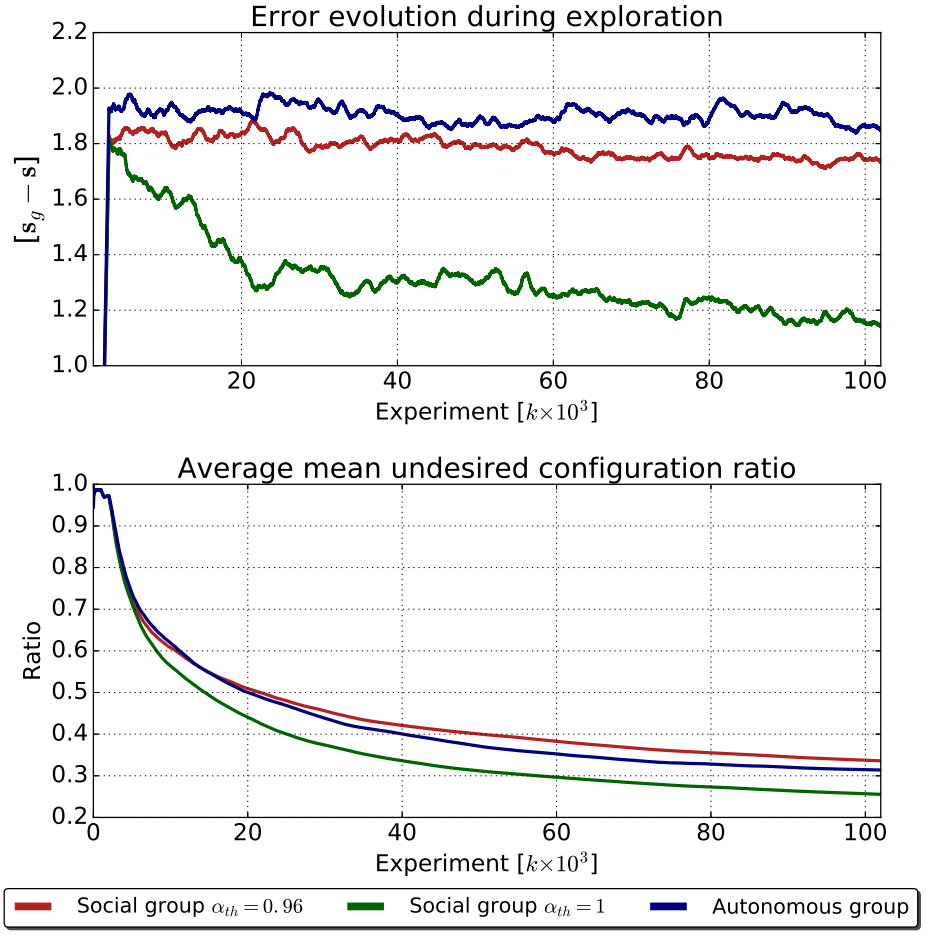

FIGURE 5.8: Results for simulation with the ear-vocal tract system using Algorithm 7. (Upper) Average sensory error during exploration using a moving average window of 1000 samples. (Lower) Average undesired motor configuration ratio evolution along the exploration.

Summarizing, for each simulation $M_{\mathcal{SM}}$ and $M_{\mathcal{SS}}$ are initialized together as indicated in line 1 of Algorithm 7 with the different initial motor command sets. Later, using the auditory results of the first stage, the interest model M_{IM} is initialized as indicated in line 2 of Algorithm 7. During the initialization of M_{IM}, $M_{\mathcal{SM}}$ is used to infer the motor actions that will likely produce the initial auditory goals. These commands are executed without considering the nociceptive prediction p_{tmp}. Afterward, the socially reinforced intrinsically motivated sensorimotor exploration is run for 100K experiments.

Finally, the sensorimotor model generated during the exploration is evaluated with respect to the 323 coarticulated gestures, which were generated from the recombination of the German vowels dataset described in Figure 5.3 and Table 5.3. This set of samples will be called evaluation dataset $\mathbf{S}_{eval}$ in the following. Evaluation against $\mathbf{S}_{eval}$ is performed every 2.5K

samples during each simulation.

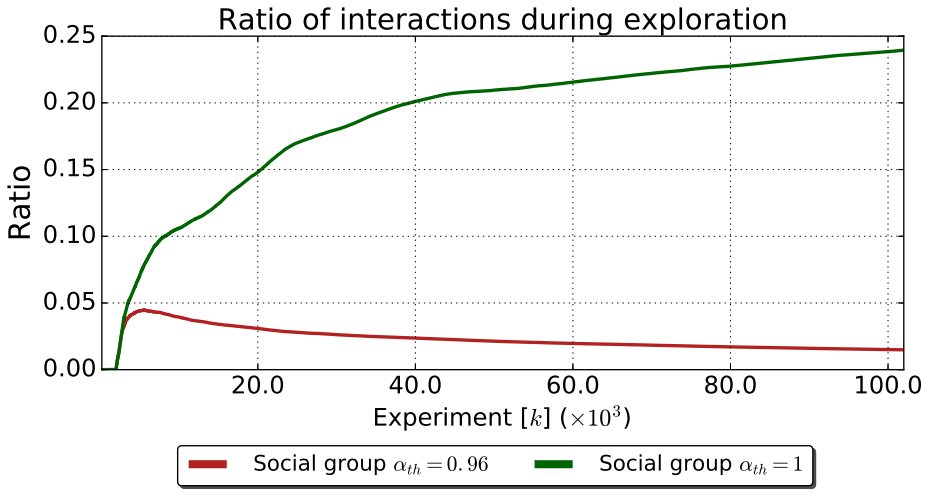

FIGURE 5.9: Results for simulation with the ear-vocal tract system using Algorithm 6. (Upper) Average mean evaluation error performed every 2.5K samples during exploration against $\mathbf{S}_{eval}$. (Center) Average sensory error during exploration using a moving average window of 1000 samples. (Lower) Average undesired motor configuration ratio evolution along the exploration.

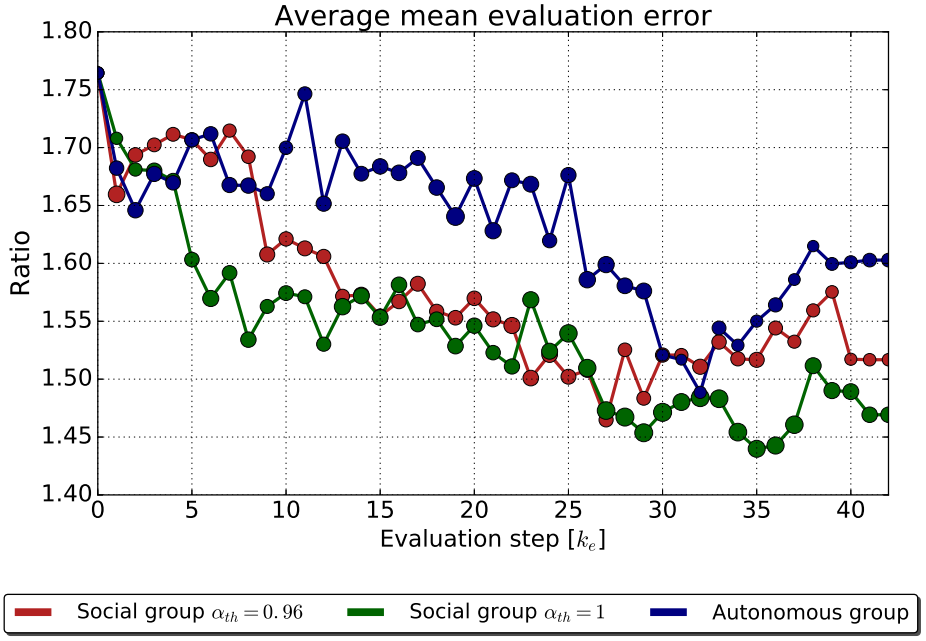

FIGURE 5.10: Results for simulation with the ear-vocal tract system using Algorithm 6. (Upper) Average mean evaluation error performed every 2.5K samples during exploration against $\mathbf{S}_{eval}$. (Center) Average sensory error during exploration using a moving average window of 1000 samples. (Lower) Average undesired motor configuration ratio evolution along the exploration.

In Figure 5.8 (upper), it is shown the moving average of the mean exploratory error along the different groups of simulations considering a window size ws of 1000 samples and defined in

Eq. (4.3). In Figure 5.8 (lower), it is shown the undesired motor configuration ratio, defined in Eq. (4.4), for each group of simulations along the exploration. Finally, Figure 5.9 shows the average ratio of *reformulation/imitation* episodes defined in Eq. (5.3).

On Competence and Social Reinforcement Impact

Results suggest that those agents that are endowed with the social mechanism perform better than the autonomous agents. From Figure 5.8 (upper), it is observed that social reinforcement produces significant differences between simulation groups regarding sensory error during explorations. In general, the results are consistent with Acevedo-Valle et al. (2017a) regarding the advantages of social reinforcement. Moreover, it is observed that the social group with $\alpha_{th} = 1$ starts reducing its exploratory error considerably faster than the autonomous group and the other social group. The tendencies regarding the average ratio of undesired configurations are similar in all the considered cases, but the social groups achieve lower amounts of undesired configurations, being $\alpha_{th} = 1$ the best scenario. The differences between simulation groups may be attributed to the results presented in Figure 5.9. Therein, as expected, it is observed that the agents with $\alpha_{th} = 1$ produced more *reformulation/imitation* episodes than the other agents. In fact, the group of social agents with $\alpha_{th} = 0.96$ shows the same tendency as the other social group at the beginning of the simulation, but as long as the value of α_{th} causes the social thresholds in the vector $\mathbf{th_S}$ to decrease every time an interaction occurs (indicated in line 3 of function *interaction* in Algorithm 7). Then, the interactions between instructor and learner occur with less frequency, causing the exploration to be more similar to the exploration made by autonomous learners.

In Acevedo-Valle et al. (2017a), we argued that the changes in the tendency of interactions from the perspective of the instructor could be an indicator of developmental changes. Therein, regarding the ratio of interactions, it was observed that at the beginning of the simulations without considering the somesthetic modality, learners start to imitate any feedback received from the instructors regardless of the possibility to reach undesired motor configurations. On the other hand, at the beginning learners endowed with the somesthetic mechanism does not imitate with the same frequency. The unwillingness to imitate is attributed to the somesthetic mechanism. If the nociceptive prediction in line 17 of Algorithm 7 indicates that an undesired configuration is likely to occur, then imitation does not occur. However, as learners continue exploring and discover regions where attempting imitation is not likely to

produce undesired motor configurations, then the amount of interactions increases dramatically. For instructors, who are unaware of learners' internal cognitive processes, the rise of the number of interactions might be seen just as a spontaneous 'desire' of learners for social interaction, in other words, instructors may interpret this change as the onset of a socially guided developmental stage.

Regarding the evaluation against the coarticulated gestures in $\mathbf{S}_{eval}$, Figure 5.10 shows the average mean evaluation error e_{ev} evolution for each simulation group. It is important to remember that the size of the markers in the plot is proportional to the standard deviation between simulations in the same group normalized according to the maximum standard deviation obtained for all the groups. In terms of progress on the ability of the agents to reproduce the evaluation dataset $\mathbf{S}_{eval}$, it is observed that tutored learners achieve better results, even though the evaluation does not display a smooth evolution. Looking at Figure 5.10, except during some periods, the autonomous agents are an upper bound and the social group with $\alpha_{th} = 1$ are the lower bound of the average mean error. The minimum value for e_{ev} is achieved by the social group with $\alpha_{th} = 1$ as reported also in Table 5.6.

Similar to Figure 4.9, Figure 5.11 was obtained using the average proportion of each type of vocalizations for each simulation group. Vocalizations are classified in three types: (a) Silent, if no phonation occurs in any of the two perception windows; (b) Unarticulated, if phonation occurs in one of the perception windows; and (c) Coarticulated, if phonation occurs in both perception windows. Plots in Figure 5.11 show the evolution of the proportional contribution of each vocalization type to the total of vocalizations through the exploration for each simulation group. The improvement in the number of phonatory vocalizations and especially the amount of coarticulated gestures is notorious when the social contingency is considered. In the last chapter, we mentioned that agents belonging to the somesthetic group avoid exploitation of silent regions. Many of these regions are likely produced by undesired motor configurations, which are avoided due to the ability to predict the nociceptive result of a given motor command. In this chapter, we have observed that the social mechanism has a dominant influence when it comes to the production of more complex vocalizations. In general, we have obtained a more clear picture of the developmental transition from silent vocalizations to unarticulated and later to coarticulated gestures.

We emphasize that unaware of the learners' internal cognitive processes, changes in the learners' behavior regarding the rate of change for the different vocalization type proportions may

be interpreted by the instructors as developmental milestones. If we observe the simulation group with $\alpha_{th} = 0.96$ in Figure 5.11, it is observed that the rate of change in the proportions is related to the ratio of interactions showed in Figure 5.9, when a socially guided developmental stage vanishes and the agent starts an autonomous exploration, then the rate of change of vocalization type proportion decreases. On the other hand, when considering the social group with $\alpha_{th} = 1$, as the social guidance does not vanish, thence the ratio of interactions keeps growing, first with a significant rate which later decreases, therefore it is expected that the proportion of unarticulated and coarticulated vocalizations remains growing as corroborated in Figure 5.11. Furthermore, the rate of growth of those vocalization types seems to be related to the rate of growth of the *reformulation/imitation* episodes ratio.

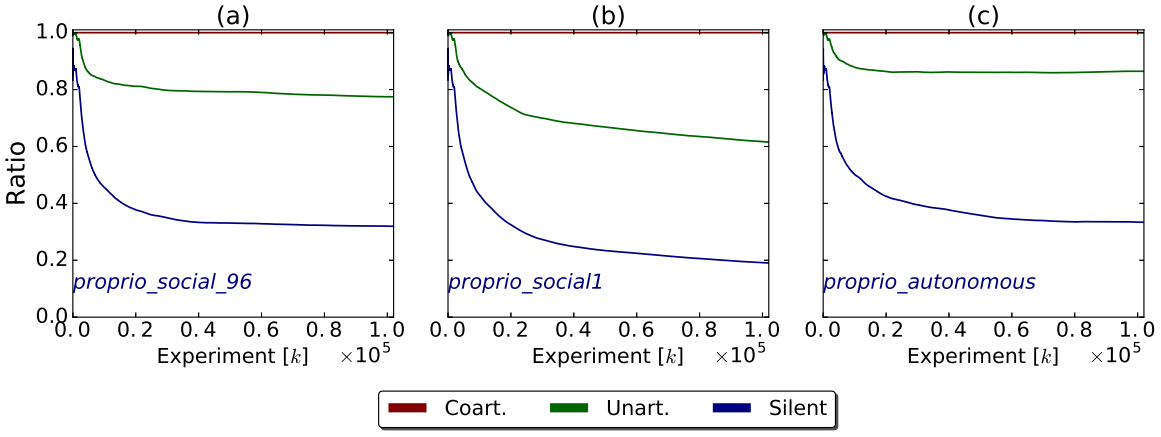

FIGURE 5.11: Proportions of vocalization classes. (a) Social group ($\alpha_{th} = 0.96$). (b) Social group ($\alpha_{th} = 1$). (c) Autonomous group.

Finally, we argue that social groups, especially that with $\alpha_{th} = 1$, achieve better performance regarding the exploratory error with respect to the autonomous group due to a significant shrank of silent and unarticulated vocalizations that lead to higher errors as discussed in Chapter 4.

On Explored Regions

Table 5.6 reinforces the arguments provided based on Figures 5.8-5.11 regarding the performance of the exploration. A series of numerical descriptors are displayed for each simulation group to draw more conclusions from the experimental results. First, the minimum value achieved for the average mean evaluation error e_{av} is displayed, visible as well in Figure 5.10. Then, the average mean exploratory error for each group of simulations is also provided.

TABLE 5.6: Exploration results for the era-vocal tract system using Algorithm 7.

Group	autonomous	$\alpha_{th} = 0.96$	$\alpha_{th} = 1$
$\min(e_{av})$	1.4885	1.4648	1.4398
Average $\lvert\mathbf{s}_g - \mathbf{s}\rvert$ exploration	1.8741	1.7578	1.2949
$ucr_{av,expl}$	0.3139	0.3359	0.2554
Unart. vocalization ratio	0.5305	0.4549	0.4249
Coart. vocalization ratio	0.1359	0.2256	0.3847
Convex hull volume	1.0430	1.0074	0.9992

Note: The table shows, in order of appearance, the minimum average mean evaluation error, the average mean exploratory error, the average 'painful' articulation ratio during the exploration, the average ratio of unarticulated and coarticulated vocalizations along the simulations, and the volume of the convex-hull encapsulating the frequency components of the explored auditory data.

Moreover, it is also shown in Table 5.6 the average 'painful' vocalization ratio produced during exploration for each group of simulations. Afterward, it is shown the average proportion of unarticulated and coarticulated vocalizations achieved by each group of simulations. In general, supporting the conclusions borrowed from the previous figures, the numerical descriptors also suggest that the results obtained with the learners endowed with the social modality perform better than the agents endowed only with the somesthetic modality.

Finally, Table 5.6 contains the volume of the convex hulls described by the explored data for each simulation group. As in Chapter 4, those volumes are obtained using the sensory data generated during each simulation and considering only the formant frequencies (F_{11}, F_{21}, F_{12}, and F_{22}) of the auditory result. Python's `scipy` library is used to compute the convex hull that encapsulates the data for each simulation. Then, the volume of the convex hull for the simulations within each simulation group is averaged to obtain the descriptor displayed in the table.

Regarding the trade-off between exploration and exploitation, experiments in this chapter show that learners without social reinforcement perform slightly better than the agents with the social mechanism regarding the size of the convex hull described by the explored sensory data. This result might be attributed directly to the fact that the imitation episodes attract the exploration to certain regions of the sensory space, whereas the autonomous learners remain free to spend more experiments in different regions of the sensory space. If imitation episodes lead to more exploitation experiments, then it is not surprising that the group of simulations with $\alpha_{th} = 1$ is the one producing the smallest convex hull, as it is the

one producing more imitation episodes. Even though there is a shrinkage of the convex hull volumes, the ratio of (1.0074/1.0430) for the group of simulations with $\alpha_{th} = 0.96$ with respect to the autonomous group represents a shrinkage of 3.4%, whereas the ratio for the group with $\alpha_{th} = 1$ is (0.9992/1.043) representing a shrinkage of 4.2%. Therefore, the shrinkage of explored regions over the sensory space is not significant compared to the improvement observed in the performance to produce sensory goals and to the increment of more complex phonatory productions.

Similar to the analysis provided in Chapter 4, Principal Component Analysis (PCA) is performed to analyze the results obtained running Algorithm 7 with the ear-vocal tract system. The PCA is performed over the formant frequencies dimensions of the sensory space ($F_{1,1}$, $F_{2,1}$, $F_{1,2}$, and $F_{2,2}$). As in the previous chapter, first, we concatenate all the data obtained with the three groups of simulations to guarantee that the results are comparable. Then, PCA is performed over the concatenated data. Looking at the PCA results, it is observed that the data can be represented keeping the two first principal components, those components keep a total of 97.27% (50.36% + 46.91%) of the data variance with respect to the original data. The first and second principal components are:

$$\begin{aligned} \vec{pca}_1 &= \begin{bmatrix} 0.4196 & 0.4844 & 0.4890 & 0.5917 \end{bmatrix}, \\ \vec{pca}_2 &= \begin{bmatrix} 0.4936 & 0.5878 & -0.3991 & -0.5015 \end{bmatrix}, \text{ and} \\ \mu &= \begin{bmatrix} 0.4619 & 0.5420 & 0.5251 & 0.6461 \end{bmatrix}, \end{aligned}$$

where μ is the estimated mean for all the sensory data used to perform the PCA procedure. Once PCA was performed, the sensory data for each group was transformed from 4-D data into 2-D data, thence Kernel-Distribution Estimation (KDE) was performed using Gaussian-Kernels in order to compare the explored regions obtained for each simulation group, and Figure 5.12 was obtained from this KDE.

In general, the results in Figure 5.12 indicate that the agents explored similar regions, but with different intensity. In both principal components, there are three main zones of exploration and the distribution of the data seems to be kind of uniform between the regions. As in the previous chapter, two high peaks appear in the regions where the silent vocalizations are located. Considering the PCA obtained in this chapter, for a null phonatory input the

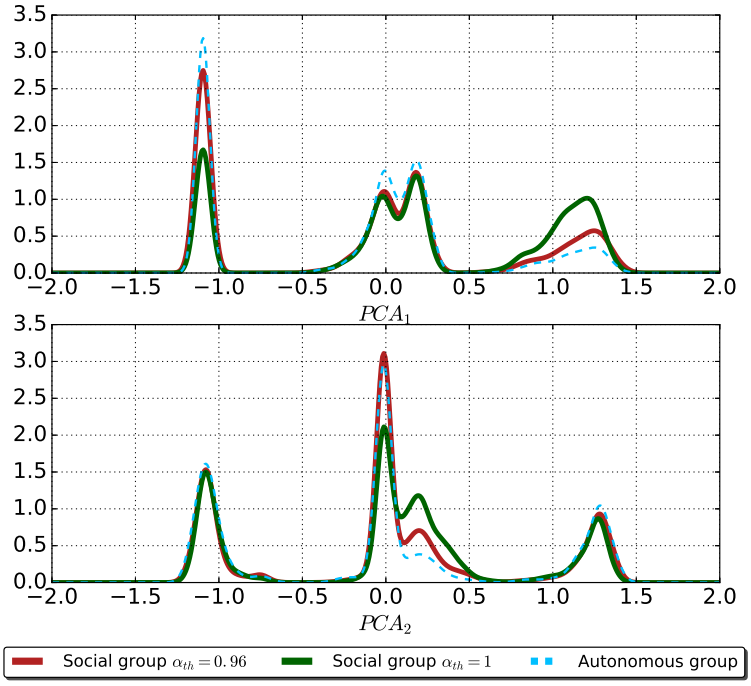

FIGURE 5.12: Estimated data density distributions using Gaussian kernels for all the data obtained for each group of simulations considering two PCA components.

PCA transformation produces

$$\mathbf{s}[0, 1, 2, 4] = [0,\ 0,\ 0,\ 0] \xrightarrow{\text{PCA transform}} [-1.0955,\ -0.0130].$$

As expected, due to the proportion of each type of vocalizations for each group (see Figure 5.11), the autonomous learners and the social learners with $\alpha_{th} = 0.96$ explored more intensively around the silent regions. On the other hand, the exploration performed by the social learners with $\alpha_{th} = 1$ look very balanced along the three regions where, in general, exploration occurs. The achieved uniformity by this social group is a good explanation to the significant increment of coarticulated vocalizations.

In Figure 5.13, data distributions along the principal components are estimated only considering the 'unpainful' vocalizations along the explorations for each simulation group. To make the results comparable, as in Chapter 2, data distributions are scaled proportionally to the number of 'unpainful' vocalization per group divided by the number of 'unpainful' vocalizations for the group that produced the most vocalizations of that kind (social group with $\alpha_{th} = 1$ as observed in Table 5.6). In Figure 5.8, it is also observed that the final ratio of undesired motor configurations is similar for the three groups of simulations. Thus,

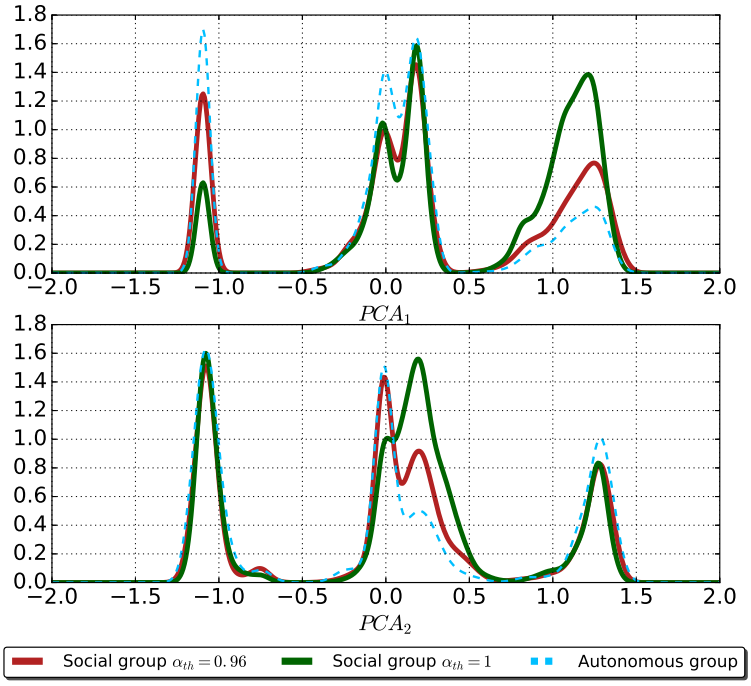

FIGURE 5.13: Estimated data density distributions using Gaussian kernels for the 'unpainful' data obtained for each group of simulations considering two PCA components.

it is obvious that even though the distributions of 'unpainful' vocalization are not equally distributed along the principal component projections for all the simulation groups, at least they should have similar volumes, which is corroborated in Figure 5.13. The differences in the shape of the distributions are still around the silent regions of both principal components, where the autonomous learners and the social group with $\alpha_{th} = 0.96$ explore with more intensity, whereas the social group with $\alpha_{th} = 1$ explores with more intensity around $PCA_1 = 1.2$ and $PCA_2 = 0.2$. This result might be related to high ratio of interactions produced by this group during the sensorimotor exploration, to corroborate this hypothesis Figures 5.14-5.16 were generated.

In Figure 5.14, data distributions along the principal components are shown only considering the phonatory vocalizations along the explorations for each simulation group. The estimated distributions are scaled proportionally to the simulation group that produced the most phonatory results (social group with $\alpha_{th} = 1$ as observed in Table 5.6). Given the proportions obtained for each vocalization type with each group of simulations according to Figure 5.11 and Table 5.6, it is not surprising to observe what the results for the distribution of phonatory exploration data indicate. In general, the social group with $\alpha_{th} = 1$ explored more intensively the phonatory regions, followed by the other social group. There is a slight

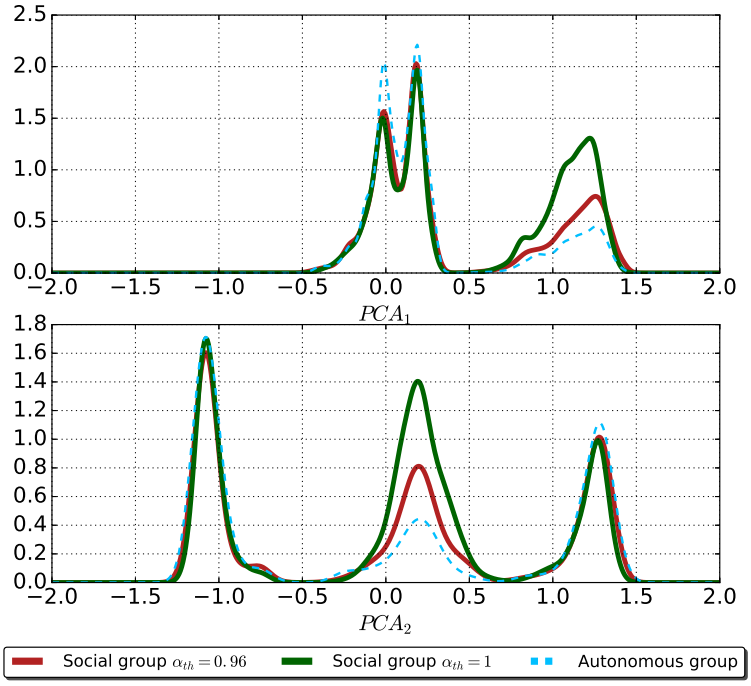

FIGURE 5.14: Estimated data density distributions using Gaussian kernels for the phonatory data obtained for each group of simulations considering two PCA components.

difference in the proportion of unarticulated gestures. As observed in Table 5.6, the social group with $\alpha_{th} = 0.96$ obtains the highest proportion of this vocalization type, 45.49% over the 42.49% obtained by the social group with $\alpha_{th} = 1$. In Figure 5.14, this difference is attributed to small sections of the principal component projections where the line describing the distribution of the explored data distribution for the social group with $\alpha_{th} = 0.96$ is higher than the other social group. For example, in the case of PCA_1 around 0 and 1.5, and for PCA_2 around -0.7 and 1.5.

Similar to Figure 4.14, Figure 5.15 was generated performing Gaussian-KDE to the dataset $\mathbf{S}_{eval}$ containing coarticulated gestures considering German vowels represented in Table 5.3. Once the distributions in Figure 5.15 were estimated, they were used to filter the distributions in Figure 5.14 in order to obtain Figure 5.16. The latter figure establishes a relation between the explored phonatory regions over the sensory space with the regions relevant for socialization. After filtering the estimated distributions, it is observed that the social groups of simulations have a clear dominance over the autonomous agents when it comes to explored regions relevant to social purposes. Which is not surprising due to the social reinforcement; logically, the group of simulations with $\alpha_{th} = 1$ explored with more intensity these relevant regions as it produces considerably more *reformulation/imitation* episodes.

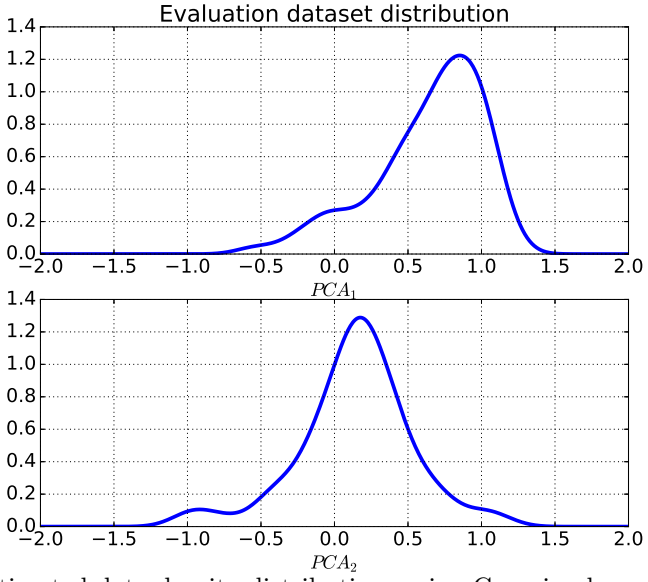

FIGURE 5.15: Estimated data density distributions using Gaussian kernels for the German vowels dataset considering two PCA components.

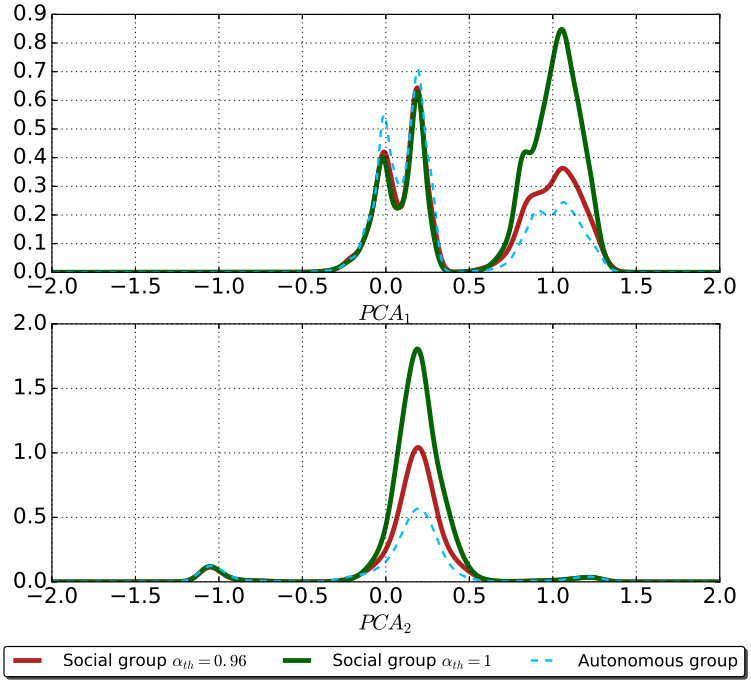

FIGURE 5.16: Estimated data density distributions using Gaussian kernels for the phonatory data obtained for each group of simulations considering two PCA components and filtered with the evaluation dataset's distribution.

5.6 Discussion

In this chapter, we have extended the results obtained with the architecture presented in Chapter 4. That architecture is an intrinsically motivated exploration architecture used

to imitate the sensorimotor exploratory behaviors in humans considering motor constraints. The novelty in this chapter is related to the integration of the concept of social reinforcement into the architecture. This reinforcement is observed naturally occurring in interactions between infants and caregivers and has an important impact on early social prelinguistic development according to developmental psychology studies mentioned as well. In general, we propose the use of *reformulation/imitation* episodes based on evidences provided in Gros-Louis et al. (2016) about mothers responding as if children were approximating a word may support language development (*imitation/expansion* responses).

Most of the conclusions obtained in the previous chapter and Acevedo-Valle et al. (2017a) were corroborated. From the experimental results, it is concluded that the socially reinforced architecture has evident advantages over autonomous architecture when looking at the evolution of exploratory and evaluation errors. When using the architecture with the ear-vocal tract system considering an instructor expert on German vowels, experimental results suggest that social reinforcement is crucial to the emergence of more complex vocalizations (coarticulated gestures) as emphasized in Gros-Louis et al. (2014). The novel architecture is compared with the best results obtained in Chapter 4, where somesthetic senses and intrinsic motivation roles were studied.

As mentioned in Calinon et al. (2007), in social artificial learning approaches, a set of generic questions have been formulated. For instance, a social learning approach faces these questions: *What to imitate?*, *How to imitate?*, *When to imitate*, and *Whom to imitate?*. Demiris and Meltzoff (2008), indirectly, formulates the same questions when analyzing the different approaches in developmental robotics studies of imitation mechanisms. In this work, we established that imitation occurs in a bidirectional manner, from the perspective of the instructor, it imitates any utterance by the learner that is similar to a socially relevant utterance, reformulating the original learner's utterance as that relevant utterance. On the other hand, from the perspective of the learner, it will establish as a sensory goal any feedback received from the instructor. However, imitation will occur only under the condition that the learner has some knowledge on how to imitate without the risk of reaching a 'painful' motor state. The question *Whom to imitate?* is not considered in this work.

As in Chapter 4, in the case of the simple toy example it was corroborated that for all the considered dimensions of evaluation, agents endowed with the somesthetic mechanism perform better. However, the comparison between the autonomous somesthetic groups and

the two somesthetic groups considering the social mechanism is not straightforward as they do not produce radical differences in their behavior. However, looking at the nuances of the results, it is possible to observe that the social groups of simulations perform slightly better as suggested by the minimum achieve evaluation errors reported in Table 5.5. Moreover, if the best performance regarding the social evaluation is achieved sooner when $\alpha_{th} < 1$ it may be interpreted as an opportunity for development. If the instructor detects that the agent has successfully developed in producing those sensory units, then it may be the door to attempt more complex interactions.

Regarding the ear-vocal tract system, comparing with the results in Chapter 4, it was observed that six simulations per scenario were not enough to establish reliable differences between simulation scenarios. Therefore, for each of the considered scenarios, we ran eighteen simulations. Through this work, it has been insisted about the high dimensionality and redundancy of vocal spaces, and the fact that many vocalizations lead to non-phonatory vocalizations, especially considering our `divapy` implementation of an ear-vocal tract system. With the modifications to the intrinsically motivated sensorimotor architecture proposed in Chapter 4, where motor constraint awareness is integrated using a somesthetic sense, a large amount of exploitation over regions of vocalizations producing 'silent' sensory results was avoided. Thus, the learners can spend more energy exploring in regions were complex vocalizations can be learned; thus increasing the proportion of unarticulated and coarticulated gestures compared to autonomous agents. In this chapter, through the introduction of a social modality working through *reformulaton/imitation* episodes, it is achieved a radical improvement regarding the proportion of complex vocalizations produced. In fact, the group of simulations considering $\alpha_{th} = 1$ achieved 20.67% more phonatory vocalizations, and 180% more coarticulated vocalizations considering absolute numbers. In other words, for each coarticulated vocalization produced by the autonomous learners, the social learners with $\alpha_{th} = 1$ produced almost three coarticulated vocalizations, which is a big improvement. Therefore, the social contingency enables artificial agents to spend more energy exploring and exploiting complex phonatory regions. In this sense, apart from being shaped by constraint awareness, we showed how sensorimotor exploration could be driven to sensory regions that generate more complex behaviors by interactions with an expert instructor, where imitation interactions occur when an instructor reformulates learner's spontaneous sensory experiments similar to sensory units relevant to communication.

Regarding the emergence of vocal developmental stages, in Chapter 4 we observed a clear developmental trajectory in the somesthetic groups as shown by Figure 4.9. The results considering the somesthetic modality were considerably better than agents under the architecture from Moulin-Frier et al. (2013). In this chapter, Figure 5.11 shows a beautiful picture of developmental transition with a significant improvement compared to the autonomous architectures. It is observed how the social modality is crucial for the onset of stages with complex vocalizations. Improvement is notorious even though we consider a simple imitation scenario, where the instructor has a rather simple behavior compared to the complex interactions that are observed between caregivers and infants as described in Gros-Louis et al. (2014).

We consider that an important fact to be remarked is that, besides exploring regions relevant to socialization, the social learners do not show any relevant handicap when compared with the autonomous agents, just a small shrinkage in the size of the convex hull described by the explored regions over the sensory space. On the advance toward the study of vocal development using artificial agents, apart from the contribution of the somesthetic mechanism to learn vocal spaces in interesting and less redundant regions, we also show the effects that social interactions have over the uniformity of explored sensory regions. In this sense, exploration focuses on the exploitation of relevant regions to communication as infants do from the onset of Canonical Babbling as indicated by Kuhl (2004).

As hypothesized at the end of the Chapter 4, in sensorimotor 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. In this chapter, experimental results have established the importance of studying mechanisms of social development in parallel to vocal development and other sensorimotor developmental processes. We conclude that the study of artificial vocal development should evolve in two directions, a more realistic speech architecture, and more complex social scenarios. Our experimental setup is coherent with respect to evidence suggesting that newborns can imitate static gestures, in fact, infants seem to be able to identify the means of achieving the end-state when they see the end-state (Demiris and Meltzoff, 2008). The experimental results have demonstrated some advantages of our architecture over similar ones; however, there is still a prioritized requirement to consider unstructured vocalizations and more complex social scenarios attempting to cover

other categories of maternal response and infants' vocalization directionality as defined in Gros-Louis et al. (2014).

Chapter 6

Conclusions and Future Work

"Poole and Bowman could talk to Hal as if he were a human being and he would reply in the perfect idiomatic English he had learned during the fleeting weeks of his electronic childhood."

— Arthur C. Clarke, *2001. A Space Odyssey.*

This research has presented a study with experimental results of the role of constraints and social interaction during exploratory sensorimotor behaviors; special emphasis was placed on prelinguistic vocal development. Based on available methodologies of developmental robotics, especially those of Moulin-Frier and Oudeyer (2013b), Oudeyer et al. (2007) and Howard and Messum (2007), and in coherence with a large body of developmental psychology studies, we have provided a study to understand possible mechanisms underlying developmental progression. We have shown that intrinsically motivated sensorimotor exploration can be enriched with more perceptual modalities, as somesthesis and social imitation mechanisms, to obtain a picture of developmental change that allows the emergence of complex behaviors. In the case of vocal prelinguistic development, we observed a nice picture of vocal gestures complexity evolution, which is necessary for the emergence of spoken language in children.

6.1 General Conclusions

The first objective of this work was to collect and understand the series of studies that led to the findings in Moulin-Frier et al. (2013), which was performed in Chapter 2. We

went through a series of studies regarding embodiment, intrinsic motivations, sensorimotor exploration, but also studies dealing with sensorimotor behaviors and development of the speech production and perception systems in children. Furthermore, we briefly studied some theories that attempt to formalize the perceptuo-motor link in speech. Regarding the perceptuo-motor link, we argue that one of the contributions of this work, and those in early vocal development studies, is the fact that if we can have models that are candidates to represent the perceptuo-motor human speech system, they can be used to improve dialogue-based human-machine interaction systems.

Our second and third objectives were to reproduce and expand the results from Moulin-Frier et al. (2013). This part of the work can be divided into different stages. In a first stage we used the experimental setup information available in Moulin-Frier and Oudeyer (2013b) and Moulin-Frier et al. (2013) to code an implementation in Matlab® of the architecture where the sensorimotor models were Gaussian Mixture Models (GMMs) trained with the generative method from Calinon and Billard (2007), as explained in Chapter 3. The result was a working intrinsically motivated sensorimotor exploration architecture. On the generated implementation we integrated our first contribution. Based on studies from developmental psychology and neurophysiology, we hypothesized that constraints could play an important role in the course of sensorimotor control learning. Consistent with the embodiment paradigm, we generated active learning mechanisms to endow artificial agents with constraint awareness along the motor inference processing path. Two articles were published with the results obtained at this stage of the project; these are Acevedo-Valle et al. (2015, 2018).

In the second part of our research, pursuing our second and third objectives, we migrated our experimental platform to Python. During this stage, we developed a new incremental learning platform for GMMs and Gaussian Mixture Regression (GMR) based on geometric properties of Gaussian distributions. We presented our approach for GMMs in Chapter 3. We showed there are some cases in which the generative method from Calinon and Billard (2007) works better and others where our growing GMMs method works better, that depends on the learning application. In this stage we also implemented a version in Python of the Maeda's synthesizer (`divapy`), a vocal-tract initially implemented in Matlab® by the Guenther Lab and widely used in the literature as a simulator of the ear-vocal tract system. After migrating to Python, we also were able to take some advantages of the `explauto` library, which is

a specialized tool for autonomous sensorimotor exploration problems (Moulin-Frier et al., 2014).

The last part of our research pursuing our second and third objectives is contained in Chapter 4. Therein, we obtained results using the current version of the proposed architecture. Regarding constraint awareness, our results strongly suggested that they should be taken into account during sensorimotor learning as suggested by Corbetta et al. (2014). Motor constraint awareness improves performance in both respects, the size of the explored sensory regions and the performance to reach self-generated goals. Considering constraints also produced better results when evaluating against some selected datasets. Constraint awareness was achieved by providing a somesthetic mechanism. Artificial agents were endowed with the capacity to generate a somesthetic model and use it to predict the somesthetic consequences of motor actions. If an agent knew that an action execution could likely lead to an undesired state and produce the perception of 'pain', then it could avoid the execution of such action.

From the new results introduced in Chapter 4 and those from Acevedo-Valle et al. (2015, 2018), several improvements to the state of the art studies can be pointed out. First, for the simple sensorimotor system introduced in Chapter 3, we observed that endowing agents with motor constraints improved the results in all the considered dimensions to evaluate performance. Secondly, for the ear-vocal tract system, we argued that sensorimotor redundancy was reduced when somesthesis was included. Thus, the number of exploratory experiments decreased, and the number of exploitation experiments increased over 'unpainful' phonatory vocalization regions. In consequence, sensorimotor models achieved better fitting to the regions of interest where motor constraints are met, and phonation occurs.

Our fourth and last specific objective was to study the role of social interactions in early vocal development and integrate a social modality to our sensorimotor exploration architecture. As in the case of language, speech cannot be learned in isolation. We hypothesized that social mechanisms of interaction such as imitation are essential for infants, and also should be for robots, during the period when they engage in sensorimotor exploratory behaviors. We assumed that instructors, experts in sensorimotor behaviors that are relevant to socialization, could interact with developmental agents to help them in the task of mastering socially relevant gestures. We focus on the imitation mechanism through *reformulation/imitation* episodes arguing that they allow infants and robots to acquire knowledge from other humans

or robots incrementally. The first stage of this part of the project was published in Acevedo-Valle et al. (2017a) and the second stage is contained in Chapter 5 and our submitted work Acevedo-Valle et al. (2018). Therein, we first introduced a socially reinforced intrinsically motivated exploration architecture with constraint awareness, where learners use an imitation mechanism as human and birds do during early development to foster development.

We integrate the social modality into our architecture based mainly on the results by Gros-Louis et al. (2016), where the authors remarked the relevance of *imitation/expansion* episodes between children and mothers as one of the most important social interactions shaping development. We also got inspiration from Howard and Messum (2011), Miura et al. (2012) and other works where the social modality is considered to study vocal development. However, we considered that learners start developing the social interactions can occur since the beginning of the life agent if the conditions required for the interaction occurs, which means that the somesthetic and sensorimotor models will be shaped by the social modality. In other words, we did not consider hard coding to trigger cognitive, motor, perceptual, nor social capabilities, neither the onset of developmental stages by hard coding. In contrast, we considered that the cognitive capabilities of the artificial agent develop in parallel and it is the deeply intertwined developmental process which produces notorious developmental changes. In fact, our results suggest that it may be the reaching of a developmental milestone in one modality that produces abrupt developmental changes in the other. For instance, there are some regions of the sensory space that are not explored by a learner, until its somesthetic model indicates that the motor regions that are likely to be related to that sensory regions are not likely to produce undesired 'painful' states.

In this work, we considered that imitation occurs in a bidirectional manner: from the perspective of the instructor, it imitates any utterance by the learner that is similar to a socially relevant utterance, reformulating the original learner's utterance as that relevant one. On the other hand, from the perspective of the learner, it will establish as a sensory goal any feedback received from the instructor. However, imitation will occur only under the condition that the learner has some knowledge on how to imitate without the risk of reaching a 'painful' motor state. From the results in Acevedo-Valle et al. (2017a, 2018) and Chapter 5, it can be concluded that considering *reformulation/imitation* episodes, the social reinforcement has an important role in the course of intrinsically motivated learning.

Considering socially reinforced explorations in the case of the simple toy example, the comparison against the autonomous exploration was not straightforward as the performance did not drastically change as it did just by considering the somesthetic mechanism. However, the nuances of the results, plus the number of simulations executed to obtain representative results, suggested that the social groups of simulations perform slightly better. Regarding the ear-vocal tract system considering an instructor expert on German vowels, experimental results suggested that social reinforcement is crucial to the emergence of coarticulated gestures as emphasized in Gros-Louis et al. (2014). First, the motor constraint awareness explained in Chapter 4 had already reduced the amount of exploitation over regions of vocalizations producing 'silent' sensory results. However, through the introduction of the *reformulaton/imitation* episodes, it was achieved a significant increment of the 'phonatory' vocalization proportion. Besides exploring regions relevant to socialization, and produce significantly more 'phonatory' vocalizations, the social learners did not show any relevant handicap when compared with the autonomous agents, just a small shrinkage in the size of the explored regions over the sensory space.

There were two more objectives for this thesis. First, it was to generate sensorimotor exploration algorithms in such a way they could be easily applied to any sensorimotor system and not only to the ear-vocal tract system. We did so, and moreover, we presented a simple toy example to illustrate interested researchers about the kind of problems in which our architectures can be used. Regarding the last objective, it was to generate open source codes that can be used by any interested researcher willing to work in similar applications. As a part of this project we presented the work Acevedo-Valle et al. (2017b) and generated two open source Python tools[1,2].

6.2 Final Conclusion

On the advance toward the study of vocal development using artificial agents, we have contributed investigating the role of somesthetic and social mechanisms during intrinsically motivated sensorimotor exploration. Motor constraint awareness, integrated through somesthesis, assists artificial learners to explore vocal spaces in interesting and less redundant

[1] https://github.com/yumilceh/igmm
[2] https://github.com/yumilceh/divapy

regions. Moreover, we have also shown the effects that social interactions have over the uniformity of explored sensory regions; thus exploration focuses on the exploitation of relevant regions to communication as infants do from the onset of canonical babbling as indicated by Kuhl (2004).

The positive impact of motor constraint awareness on the considered problem is undeniable. The impact caused by the social evaluation is open to more interpretations. In this sense, we argued that one will choose an architecture and parameter values according to the problem at hand. We consider a scenario where a social similarity threshold is considered to trigger the *reformulaton/imitation* episodes. That threshold is scaled by a factor lower than one every time an imitation episode occurs. This mechanism is similar to the 'scaffolding' process observed in caregivers/children interactions, where the support provided by the caregiver to the child is gradually reduced as the child makes progress towards succeeding in the task being learned.

To the values considered for the scaling factors, we found that, especially in the toy example, considering a right value could foster the progress to master the socially relevant sensory units. In other words, one could achieve faster progress toward lower errors which may be interpreted as an opportunity for development. If the instructor detects that the agent has successfully learned to produce those sensory units, then it may be the door to attempt more complex interactions. However, regarding the overall performance, when the scaling factor was not considered (or was equal to one), we obtained the best results regarding vocal developmental stage transitions.

Developmental stage transitions were the main object of study in Moulin-Frier et al. (2013). Therein, the authors found that intrinsic motivations could be a good candidate to explain the developmental transitions observed in infants' vocalizations: in a first stage, 'silent' vocalizations are dominant, then gradually unarticulated and coarticulated vocalizations emerge, at the beginning being the earlier kind of phonatory vocalizations dominant, but gradually this dominance is taken by coarticulated gestures, as observed in infants (Kuhl, 2004). In Chapter 4, comparing agents simulated with the architecture from Moulin-Frier and Oudeyer (2013b), we observed a more evident developmental transition when somesthesis is considered. In Chapter 5, we obtain a final picture of developmental stage transitions of this work. Therein we showed a picture of developmental transition with a significant improvement compared to the autonomous architectures considered in this work. In that

chapter, it is observed that the social modality is a crucial issue for the onset of stages with complex vocalizations. The amount of coarticulated vocalizations produced by socially reinforced learners is considerably higher than in other cases. This result was partially attributed to the number of interactions generated, especially when the scaling factor was not considered. Improvement is notorious even though we consider a simple imitation scenario, where the instructor has a rather simple behavior compared to the complex interactions that are observed between caregivers and infants, as described in Gros-Louis et al. (2014).

Summarizing, somesthetic and social contingencies enable artificial agents to spend more energy exploring and exploiting complex phonatory regions during sensorimotor exploration. In this sense, we showed how sensorimotor exploration, thus sensorimotor knowledge, is shaped by constraint awareness. Moreover, exploration is also driven by interactions with expert instructors toward sensory regions that generate more complex behaviors that later are used for complex communication. In general, we have observed that besides intrinsic motivations, it is important to emphasize the relevance of other mechanisms, somesthetic perception and social episodes among them. As the literature coming from neuroscience and developmental psychology suggests, developmental robotics studies should aim at a holistic approach where we study mind and body together; thus we also deal with perceptual and motor capabilities together. We argue that the success of building a robot that develops as a human does and to borrow insights regarding the developmental process in infants will rely upon the capacity of roboticists to generate systems in which different perception, motor, cognitive and social contingencies develop together.

We argue that one of our more important contributions, compared to other similar architectures, is that we did not prespecify the onset of different developmental stages or socially guided stages. Instead, we build a system in which all the modalities develop together. This requirement is emphasized by works like Howard and Messum (2011) and Kröger et al. (2009), where it is mentioned that likely self-learning and social learning occur in parallel, even though in practice those works considered that both stages occur in series. Moreover, we proposed a possible interconnection that may exist between somesthetic, auditory, and intrinsic motivations that may explain different processes occurring within an infant's mind during a simplified scenario similar to Canonical Babbling.

In building artificial developmental systems, as we have done through this work, many questions arise regarding nature and nurture. Demiris and Meltzoff (2008) asked *Where do we*

start? What bootstrapping mechanisms should we strive to implement in robotic systems and how flexible are such systems as a result? In general, we have studied speech emergence in a simplified scenario according to behavioral and physiological evidence using a developmental approach. As suggested by Asada (2016), we focused on the role of embodiment and social interactions in the course of development, where elementary cognitive skills are built. Our scenario considers an agent that through interactions with other peer and subjected to an imitative reflex incrementally build models motivated by an intrinsic drive to improve performance, allowing to empower its mental and behavioral structure. However, the agent is endowed with the capacity of choosing when to imitate according to its current sensorimotor knowledge, as mentioned by Oudeyer et al. (2007).

6.3 Future Work

Developmental robotics is a relatively young field of research, and the way to go in order to achieve its main goal is still long. We are at the beginning of a quest aimed at building robots with the capacity of developing in an autonomous open-ended manner as humans do. Through this work, we have contributed with a small piece of a puzzle that must be contextualized to the current state of the art in the quest of combining natural cognitive studies, especially those related to ontogenetic and epigenetics, with artificial cognition technologies. The ways in which this research may continue are diverse and discussed below.

We conclude that the study of artificial vocal development should evolve in two directions, a more realistic speech architecture, and more complex social scenarios. Our experimental setup is coherent with respect to evidence suggesting that newborns can imitate static gestures. In fact, infants seem to be able to identify the means of achieving the end-state when they see the end-state (Demiris and Meltzoff, 2008). The experimental results have demonstrated some advantages of our architecture over similar ones. However, there is still a prioritized requirement to consider unstructured vocalizations and more complex social scenarios attempting to cover other categories of maternal response and infants' vocalization directionality as defined in Gros-Louis et al. (2014). In the following, we discuss a series of elements that we consider of interest for further studies.

First of all, we would like to mention two interesting projects that took as a starting point the results in Moulin-Frier et al. (2013) and were carried on in parallel to our project, those

works are Forestier and Oudeyer (2017) and Najnin and Banerjee (2017). In Forestier and Oudeyer (2017), the authors focused on the role that situated vocal development may have in the emergence of vocal gestures to label objects which were part of an interaction scenario where the caregiver was considered with a similar embodiment to that of the learner. The learner could interact with objects in three different manners, reaching the object with its arm, or use its arm to reach a tool an use it to reach a farther object or ask the caregiver to hand over the object. The labeling learning mechanism could be considered similar to the one of Howard and Messum (2011), but it is implemented similarly to Moulin-Frier et al. (2013) and up to a certain point to our approach. The main difference with respect to our work is that Forestier and Oudeyer (2017) considers that the learner also chooses over which sensorimotor model explore during intrinsically motivated learning.

In the sense of exploring over different sensorimotor models, we defend that it is one improvement we could do to our architecture but with other purposes. Taking advantage of the complete somesthetic information we can borrow from the vocal tract (tactile and proprioceptive) we could build two extra sensorimotor models and then use ideas like those from Forestier and Oudeyer (2017) and Navarro-Guerrero et al. (2017b) to refine sensorimotor knowledge to achieve better performance when reaching sensory goals.

Najnin and Banerjee (2017) emphasized the study of developmental transitions between types of vocalizations as Moulin-Frier et al. (2013). The authors also made a contribution that should be explored in the course of our experiment as discussed in our works Acevedo-Valle et al. (2017a), Acevedo-Valle et al. (2018), which is a more realistic modeling framework for vocalizations and speech. In the case of Najnin and Banerjee (2017), they proposed two interesting elements. First, the use of Mel-Frequency Cepstral Coefficients that are widely used along the scientific literature and practical applications as descriptors for speech signals. Second, Najnin and Banerjee (2017) proposed that the agent learns the timing of commands execution. We defend that further studies must consider a more realistic speech modeling in which we can introduce the concept of consonants in a more realistic way than the one assumed by Moulin-Frier et al. (2013) and Najnin and Banerjee (2017). Moreover, we claim that motor timing must be taken into account in further studies as it is something children must also learn during early development. Learning motor timing could be constrained by fatigue, breathing, and cognitive processes.

Another critical element that further studies in early vocal development must take into account is to advance toward achieving more realistic scenarios as observed in humans interactions, in which adults talk to infants with adult language, that is, words or sentences (Yoshikawa et al., 2003), a line of research that has been already partially explored by Howard and Messum (2014), Kröger and Cao (2015). In achieving more complex interactions and more complex knowledge, we should advance toward understanding how the information captured in the sensorimotor models through this work may affect perception. If the learned structures become more advanced than the current ones, then roboticists might build models that could contribute to answering how the motor system may be recruited for perceiving speech as hypothesized by Galantucci et al. (2006) and Schwartz et al. (2012). In finding suitable candidates for this recruiting process we could contribute to the improvement of Automatic Speech Recognition systems, Natural Language Processing systems, and Speech Synthesizers.

Through our investigation, we have also identified other topics that are of interest in the quest of studying prelinguistic vocal development. For example, a deeper analysis of the learning processes underlying the non-auditory development related to mastication, deglutition and crying from the cognitive and developmental perspectives should be completed. This knowledge could contribute to generating more complex somesthetic architectures. Regarding maternal responsiveness, seven categories of maternal verbal response are distinguished in Gros-Louis et al. (2014): acknowledgments, attributions, directives, naming, play vocalizations, questions and *imitation/expansions*. Integrating different works, e.g., as this work with Forestier and Oudeyer (2017), we could achieve a platform in which all the kind of responses and its role in early sensorimotor exploration can be studied.

We have argued that the success on the quest of building a complex social robot with human-like cognitive skills should be approached in an interdisciplinary way. However, as we consider that it would not be enough, then we also encourage roboticists to work on the integration of motor, perceptual, cognitive and social systems in a developing agent where all the systems develop in a parallel as they do in our children.

Appendix A

A Maeda's Vocal Tract Pythonic implementation: divapy

Guenther Lab provided an implementation of the Maeda's synthesizer (Maeda, 1982, 1989) in Matlab®. The experiments in Acevedo-Valle et al. (2015, 2018) considered that implementation for experimentation purposes. On the other hand, Moulin-Frier et al. (2014) bridged the implementation in Matlab® to Python by means of the library *pymatlab*[1] with the disadvantage of being slow as the synthesizer was running on a parallel session in Matlab®. As a part of the project, our objective was to use and generate open-source software tools; moreover, tools like *explauto* are only available in Python. Therefore, we migrated our implementation from Matlab® to Python. In order to solve this bottleneck due to the execution of the synthesizer in Matlab®, we translated the code provided by Guenther Lab into a functional Python package called *divapy*[2]. The key to achieving a fast implementation in Python was to consider the library *numpy-groupies*[3], which is *"a library of optimized tools for doing things that can roughly be considered 'group-indexing operations'"*.

The vocal tract consists of ten articulators and three phonation parameters. The Python package *divapy* can provide different outputs once a sequence of articulatory positions and phonation values is established:

- Outline of the face and vocal tract.

[1] https://pypi.org/project/pymatlab/
[2] https://github.com/yumilceh/divapy
[3] https://github.com/ml31415/numpy-groupies

- Area function describing the shape of the vocal tract.
- Sequence of somatosensory signals based on tactile and proprioceptive information.
- Sequence of formant frequencies.
- Speech sound signal.

The on-line repository contains examples regarding the usage of the most important functions of the class `Diva` contained in the package. Once an object of the class has been created the methods defined for that object are described below[4].

get_audsom Returns the sequence of formant frequencies, somatosensory signals, outlines, and area function of a given sequences of articulatory positions and phonation values.

Input:

art `([ndarray] n_samples x 13)` Sequence of articulatory positions and phonation values.

scale (optional) `(bool)` Indicates if the auditory output is scaled to the values proposed by Guenther Lab.

Output:

aud `([ndarray] n_samples x 4)` Sequence of formant frequencies (F_0, F_1, F_2, and F_3).

som `([ndarray] n_samples x 8)` Sequence of somatosensory signals.

outline `([ndarray] n_samples x max_outline_shape_dim)` Sequence of the vocal-tract's outline.

af `([ndarray] n_samples x max_af_shape_dim)]` Sequence of the vocal-tract's shape descriptor (area function).

get_sound Returns the sound wave produced by a given sequence of formant frequencies.

Input:

art `([ndarray] n_samples x 13)` Sequence of articulatory positions and phonation values.

[4] `array` and `ndarray` are defined within the `numpy` package: http://www.numpy.org/

Output:

s (`[array]`) Sound signal.

get_sample Given an articulatory configuration and phonation values, this function returns the formant frequencies, somatosensory signals, outlines, and area function.

Input:

art (`[array] 1 x 13`) Articulatory positions and phonation values.

Output:

aud (`[ndarray] 1 x 4`) Formant frequencies (F_0, F_1, F_2, and F_3).

som (`[ndarray] 1 x 8`) Somatosensory signals.

outline (`[ndarray] 1 x outline_shape_dim`) Vocal-tract's outline.

af (`[ndarray] 1 x af_shape_dim)]`) Vocal-tract's shape descriptor (area function).

d Value used internally by the class.

plot_outline Given an articulatory configuration and phonation values, this function plots the outline of the vocal tract.

Input:

art (`[array] 1 x 13`) Articulatory positions and phonation values.

axes Axes in which the outline should be plotted.

Output:

axes Updated axes.

play_sound Plays the soundwave obtained with `get_sound` if any speakers are available.

Input:

s (`[array]`) Sound signal.

Note: `art[:10]` are the articulatory positions and `art[10:]` are the phonation values.

Finally, to assess the improvement of performance in comparison with the original implementation with *pymatlab* some tests were run. In the first test, we ran 1000 experiments

similar to the ones during the sensorimotor exploration simulations, described in Figure 4.3. The execution time for 1000 experiments with *divapy* was 64.34 seconds, whereas with *pymatlab* was 345.59 seconds. Finally, we obtained the sound waves for 1000 vocalizations, the execution time with *divapy* was 266.05 seconds, whereas with *pymatlab* was 242.66 seconds.

For the first run-time experiment, which is the one required for the experimentation in this thesis, we run the function **get_audsom**, and we made some efforts to optimize it, that is the reason we achieve an impressive improvement. On the other hand, for the function **get_sound** we did not make any effort to optimize the function as for now we only use the auditory output. However, if we want to work with the raw speech signal, it would be necessary to invest some time to improve the implementation of this function and obtain better performance.

Bibliography

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

J. M. Acevedo-Valle, V. V. Hafner, and C. Angulo. Social reinforcement in intrinsically motivated sensorimotor exploration for embodied agents with constraints. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, 2017a.

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

J. M. Acevedo-Valle, C. Angulo, and C. Moulin-Frier. Autonomous discovery of motor constraints in an intrinsically motivated vocal learner. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):314–325, June 2018. ISSN 2379-8920. doi: 10.1109/TCDS.2017.2699578.

J. M. Acevedo-Valle, V. V. Hafner, and C. Angulo. Social reinforcement in artificial prelinguistic development: A study using intrinsically motivated exploration architectures. *IEEE Transactions on Cognitive and Developmental Systems*, 2018. [Submitted].

C. Angulo, R. A. Téllez, and D. E. Pardo. Internal representation of the environment in cognitive robotics. *International Journal of Robotics & Automation*, 24(3):214, 2009.

M. Anusuya and S. Katti. Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, 6(3):181–205, 2009. ISSN 00189219. doi: 10.1109/PROC.1976.10158.

M. Asada. Modeling Early Vocal Development Through Infant–Caregiver Interaction: A Review. *IEEE Transactions on Cognitive and Developmental Systems*, 8(2):128–138, June 2016. ISSN 2379-8920. doi: 10.1109/TCDS.2016.2552493.

M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37(2):185–193, 2001.

M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: a survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34, 2009.

J. Baraglia, J. L. Copete, Y. Nagai, and M. Asada. Motor experience alters action perception through predictive learning of sensorimotor information. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 63–69, 2015.

A. Baranes and P.-Y. Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013. ISSN 0921-8890.

R. M. Beckstead. *A Survey of Medical Neuroscience*. Springer New York, New York, NY, 1996. ISBN 978-0-387-94488-3. doi: 10.1007/978-1-4419-8570-5.

M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, et al. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007.

P. Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603, 2013.

G. M. Bodner. Constructivism: A theory of knowledge. *Journal of chemical education*, 63 (10):873, 1986.

A. Bouchachia and C. Vanaret. Incremental learning based on growing gaussian mixture models. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 47–52. IEEE, 2011.

C. Breazeal and B. Scassellati. Robots that imitate humans. *Trends in cognitive sciences*, 6 (11):481–487, 2002.

S. Calinon. *Robot programming by demonstration.* EPFL Press, 2009.

S. Calinon and A. Billard. Incremental learning of gestures by imitation in a humanoid robot. In *Proceeding of the ACM/IEEE international conference on Human-robot interaction - HRI '07*, page 255, New York, New York, USA, 2007. ACM Press. ISBN 9781595936172. doi: 10.1145/1228716.1228751.

S. Calinon, F. Guenter, and A. Billard. On Learning, Representing, and Generalizing a Task in a Humanoid Robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 37(2):286–298, April 2007. ISSN 1083-4419. doi: 10.1109/TSMCB.2006.886952.

A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195, September 2010. ISSN 1943-0604. doi: 10.1109/TAMD.2010.2053034.

C. Chen, N. Zhang, S. Shi, and D. Mu. An efficient method for incremental learning of gmm using cuda. In *2012 International Conference on Computer Science and Service System*, pages 2141–2144, August 2012. doi: 10.1109/CSSS.2012.532.

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

Y. Demiris and A. Meltzoff. The robot in the crib: a developmental analysis of imitation skills in infants and robots. *Infant and Child Development*, 17(1):43–53, January 2008. ISSN 15227227. doi: 10.1002/icd.543.

K. Ejiri. Relationship between rhythmic behavior and canonical babbling in infant vocal development. *Phonetica*, 55(4):226–237, 1998.

P. M. Engel and M. R. Heinen. Incremental learning of multivariate gaussian mixture models. In *Brazilian Symposium on Artificial Intelligence*, pages 82–91. Springer, 2010.

M. K. Fagan and J. M. Iverson. The Influence of Mouthing on Infant Vocalization. *Infancy*, 11(2):191–202, March 2007. ISSN 15250008. doi: 10.1111/j.1532-7078.2007.tb00222.x.

C. B. Ferrell and C. C. Kemp. An ontogenetic perspective to scaling sensorimotor intelligence. In *in 'Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium', AAAI*. Press, 1996.

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

M. Forster. Aging Japan: Robots may have role in future of elder care, 2018. URL https://www.reuters.com/article/us-japan-ageing-robots-widerimage/aging-japan-robots-may-have-role-in-future-of-elder-care-idUSKBN1H33AB [Online; visited 2018-05-13].

B. Franklin, A. S. Warlaumont, D. Messinger, E. Bene, S. Nathani Iyer, C.-C. Lee, B. Lambert, and D. K. Oller. Effects of parental interaction on infant vocalization rate, variability and vocal type. *Language Learning and Development*, 10(3):279–296, 2014.

B. Galantucci, C. A. Fowler, and M. T. Turvey. The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, 13(3):361–377, 2006.

A. Gepperth and B. Hammer. Incremental learning algorithms and applications. In *European Symposium on Artificial Neural Networks (ESANN)*, 2016.

M. H. Goldstein, A. P. King, and M. J. West. Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences*, 100(13):8030–8035, 2003.

M. H. Goldstein, J. A. Schwade, and M. H. Bornstein. The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child development*, 80(3):636–644, 2009.

M. A. Goodrich and A. C. Schultz. Human-Robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction*, 1(3):203–275, 2007. ISSN 1551-3955. doi: 10.1561/1100000005.

A. Gopnik, A. N. Meltzoff, and P. K. Kuhl. *The scientist in the crib: What early learning tells us about the mind*. Perennial New York, NY, 2001.

J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11): 585–593, 2013.

J. Gros-Louis, M. J. West, M. H. Goldstein, and A. P. King. Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, 30(6):509–516, 2006.

J. Gros-Louis, M. J. West, and A. P. King. Maternal responsiveness and the development of directed vocalizing in social interactions. *Infancy*, 19(4):385–408, 2014.

J. Gros-Louis, M. J. West, and A. P. King. The influence of interactive context on prelinguistic vocalizations and maternal responses. *Language Learning and Development*, 12(3): 280–294, 2016.

F. H. Guenther. Cortical interactions underlying the production of speech sounds. *Journal of communication disorders*, 39(5):350–365, 2006.

F. H. Guenther, S. S. Ghosh, and J. A. Tourville. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, 96(3):280–301, 2006.

Guenther Lab. DIVA Source Code. *Speech, Language & Hearing Sciences Department in Boston University's College of Health & Rehabilitation: Sargent College.* URL http://sites.bu.edu/guentherlab/software/diva-source-code/ [Online; visited 2018-02-23].

V. V. Hafner and G. Schillaci. From field of view to field of reach-could pointing emerge from the development of grasping. *Frontiers in Computational Neuroscience*, 17, 2011.

E. Hall. Language Development in Children (Course Materials). California State University, Northridge. URL http://www.csun.edu/~vcoao0el/de361/de361ov_folder/index.htm [Online; visited 2018-05-18].

J. Hill, W. R. Ford, and I. G. Farreras. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49(0):245 – 250, 2015. ISSN 0747-5632.

M. Hollins. Somesthetic senses. *Annual review of psychology*, 61:243–271, 2010.

I. S. Howard and P. Messum. A Computational Model of Infant Speech Development. *In XII International Conference "Speech and Computer"*, pages 756–765, 2007.

I. S. Howard and P. Messum. Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, 15(1):85–117, 2011.

I. S. Howard and P. Messum. Learning to Pronounce First Words in Three Languages: An Investigation of Caregiver and Infant Behavior Using a Computational Model of an Infant. *PLoS ONE*, 9(10):e110334, October 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0110334.

H.-C. Hsu and A. Fogel. Infant vocal development in a dynamic mother-infant communication system. *Infancy*, 2(1):87–109, 2001. ISSN 1532-7078. doi: 10.1207/S15327078IN0201_6.

L. M. Hulit, M. R. Howard, and K. R. Fahey. *Born to talk: An introduction to speech and language*. Merrill New York, NY, 2011. ISBN 978-0-205-62752-3.

S. Hyken. Google Introduces Lifelike AI Experience With Google Duplex, 2018.
URL https://www.forbes.com/sites/shephyken/2018/05/13/google-introduces-lifelike-ai-experience-with-google-duplex/#3520fc234dcf
[Online; visited 2018-05-13].

IEEE RAS. Cognitive Robotics - IEEE Robotics and Automation Society - IEEE Robotics and Automation Society, 2017.
URL http://www.ieee-ras.org/cognitive-robotics [Online; visited 2018-05-16].

T. Ito, M. Tiede, and D. J. Ostry. Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences*, 106(4):1245–1248, 2009.

J. M. Iverson. Developing language in a developing body: The relationship between motor development and language development. *Journal of child language*, 37(02):229–261, 2010.

S. N. Iyer and D. K. Oller. Prelinguistic vocal development in infants with typical hearing and infants with severe-to-profound hearing loss. *The Volta Review*, 108(2):115, 2008.

H. Jiang. Confidence measures for speech recognition: A survey. *Speech communication*, 45 (4):455–470, 2005.

F. Kaplan and V. V. Hafner. The challenges of joint attention. *Interaction Studies*, 7(2): 135–169, 2006.

R. D. Kent, P. R. Mitchell, and M. Sancier. Evidence and role of rhythmic organization in early vocal development in human infants. *Advances in psychology*, 81:135–149, 1991.

S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester. Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121(2):723–742, 2007.

R. L. Klatzky, S. J. Lederman, and J. M. Mankinen. Visual and haptic exploratory procedures in children's judgments about tool function. *Infant Behavior and Development*, 28(3):240–249, 2005.

B. J. Kröger and M. Cao. The emergence of phonetic-phonological features in a biologically inspired model of speech processing. *Journal of Phonetics*, 53:88–100, Novemeber 2015. ISSN 00954470. doi: 10.1016/j.wocn.2015.09.006.

B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube. Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793–809, 2009.

P. K. Kuhl. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843, 2004.

E. H. Lenneberg, N. Chomsky, and O. Marx. *Biological foundations of language*, volume 68. Wiley New York, 1967.

M. Lewis, J. Haviland-Jones, and L. Barrett. *Handbook of Emotions, Third Edition*. Guilford Publications, 2008. ISBN 9781606238035.

A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

A. M. Liberman and D. H. Whalen. On the relation of speech to language. *Trends in cognitive sciences*, 4(5):187–196, 2000.

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological review*, 74(6):431, 1967.

M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics: a survey. *Connection Science*, 15(4):151–190, December 2003. ISSN 0954-0091. doi: 10.1080/09540090310001655110.

D. Luo, F. Hu, Y. Deng, W. Liu, and X. Wu. An infant-inspired model for robot developing its reaching ability. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 310–317, September 2016. doi: 10.1109/DEVLRN.2016.7846840.

S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1 (3):199–229, 1982. ISSN 0167-6393. doi: 10.1016/0167-6393(82)90017-6.

S. Maeda. Compensatory articulation in speech: analysis of x-ray data with an articulatory model. In *Eurospeech*, pages 2441–2445, 1989.

A. N. Meltzoff, R. A. Williamson, P. J. Marshall, W. Prinz, M. Beisert, and A. Herwig. Developmental perspectives on action science: Lessons from infant imitation and cognitive neuroscience. *Action science: Foundations of an emerging discipline*, pages 281–306, 2013.

S. Millar. Network models for haptic perception. *Infant Behavior and Development*, 28(3): 250–265, 2005.

M. Mirolli and D. Parisi. Towards a vygotskyan cognitive robotics: The role of language as a cognitive tool. *New Ideas in Psychology*, 29(3):298 – 311, 2011. ISSN 0732-118X. doi: http://dx.doi.org/10.1016/j.newideapsych.2009.07.001. Special Issue: Cognitive Robotics and Reevaluation of Piaget Concept of Egocentrism.

K. Miura, Y. Yoshikawa, and M. Asada. Vowel Acquisition Based on an Auto-Mirroring Bias with a Less Imitative Caregiver. *Advanced Robotics*, 26(1-2):23–44, January 2012. ISSN 0169-1864. doi: 10.1163/016918611X607347.

R. K. Moore. Twenty things we still don't know about speech. *Proc.CRIM/FORWISS Worshop on 'Progress and Prospects of Speech Research and Technology*, 1994.

A. F. Morse and A. Cangelosi. Why are there developmental stages in language learning? a developmental robotics model of language development. *Cognitive Science*, 41:32–51, 2017. ISSN 1551-6709.

C. Moulin-Frier and P.-Y. Oudeyer. Exploration strategies in developmental robotics: A unified probabilistic framework. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 1–6, 2013a. doi: 10.1109/DevLrn.2013.6652535.

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

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

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

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

B. Mutlu, N. Roy, and S. Šabanović. Cognitive Human–Robot Interaction. In B. Siciliano and O. Khatib, editors, *Springer Handbook of Robotics*, pages 1907–1933. Springer, Cham, 2016.

S. Najnin and B. Banerjee. A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production. *Speech Communication*, 92:24–41, September 2017. ISSN 01676393. doi: 10.1016/j.specom.2017.05.002.

S. M. Nasir and D. J. Ostry. Speech motor learning in profoundly deaf adults. *Nature neuroscience*, 11(10):1217–1222, 2008.

N. Navarro-Guerrero, R. J. Lowe, and S. Wermter. The effects on adaptive behaviour of negatively valenced signals in reinforcement learning. In *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 148–155, 2017a. ISBN 978-1-5386-3715-9.

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

C. M. Norris. Exercise therapy. In *Managing Sports Injuries*, pages 84–110. Elsevier, 2011. doi: 10.1016/B978-0-7020-3473-2.00009-5.

D. K. Oller and R. E. Eilers. The role of audition in infant babbling. *Child development*, pages 441–449, 1988.

D. K. Oller, E. H. Buder, H. L. Ramsdell, A. S. Warlaumont, L. Chorna, and R. Bakeman. Functional flexibility of infant vocalization and the emergence of language. *Proceedings of the National Academy of Sciences*, 110(16):6318–6323, 2013.

P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, April 2007. ISSN 1089-778X. doi: 10.1109/TEVC.2006.890271.

Oxford Dictionaries. English by Oxford Dictionaries, 2018. URL https://en.oxforddictionaries.com/definition/cognition [Online; visited 2018-05-15].

L. Pape, C. M. Oddo, M. Controzzi, C. Cipriani, A. Förster, M. C. Carrozza, and J. Schmidhuber. Learning tactile skills through curious exploration. *Frontiers in neurorobotics*, 6:6, 2012.

Patricia Kuhl. The linguistic genius of babies. TED Talk, 2010. *TED. Ideas worth spreading.* URL https://www.ted.com/talks/patricia_kuhl_the_linguistic_genius_of_babies [Online; visited 2018-05-18].

J. Perkell, F. Guenther, H. Lane, M. Matthies, Y. Payan, P. Perrier, J. Vick, R. Wilhelms-Tricarico, and M. Zandipour. The sensorimotor control of speech production. In *Proceedings of the First International Symposium on Measurement, Analysis and Modeling of Human Functions*, pages 21–23, 2001.

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

R. Pfeifer, M. Lungarella, and F. Iida. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318(5853):1088–1093, 2007.

S. Prescott and S. Ratté. Somatosensation and Pain. In *Conn's Translational Neuroscience*, pages 517–539. Elsevier, 2017. ISBN 9780128025963. doi: 10.1016/B978-0-12-802381-5.00037-3.

D. Purves, G. J. Augustine, D. Fitzpatrick, L. Katz, A.-S. Lamantia, J. O. Mcnamara, M. Williams, W. C. Hall, A.-S. Lamantia, J. O. Mcnamara, and S. M. Willians. *Neuroscience*. Sinauer Associates, 2nd edition, 2001. ISBN 0878937250. doi: 978-0878937257.

V. C. Ramenzoni and U. Liszkowski. The social reach. *Psychological Science*, 27(9):1278–1285, 2016. doi: 10.1177/0956797616659938. PMID: 27481910.

R. Rayyes, D. Kubus, C. Hartmann, and J. Steil. Learning inverse statics models efficiently. *arXiv preprint arXiv:1710.06463*, 2017.

A. Ribes, J. Cerquides, Y. Demiris, and R. L. de Mantaras. Active learning of object and body models with time constraints on a humanoid robot. *IEEE Transactions on Cognitive and Developmental Systems*, 8(1):26–41, March 2016. ISSN 2379-8920. doi: 10.1109/TAMD.2015.2441375.

G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

M. Rolf. Goal babbling with unknown ranges: A direction-sampling approach. In *Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on*, pages 1–7. IEEE, 2013.

M. Rolf, J. J. Steil, and M. Gienger. Goal babbling permits direct learning of inverse kinematics. *IEEE Transactions on Autonomous Mental Development*, 2(3):216–229, 2010.

R. Saegusa, G. Metta, G. Sandini, and S. Sakka. Active motor babbling for sensorimotor learning. In *Robotics and Biomimetics, 2008. ROBIO 2008. IEEE International Conference on*, pages 794–799. IEEE, 2009.

F. Sallustro and C. W. Atwell. Body rocking, head banging, and head rolling in normal children. *The Journal of Pediatrics*, 93(4):704 – 708, 1978. ISSN 0022-3476. doi: 10.1016/S0022-3476(78)80922-6.

G. Sandini, G. Metta, and J. Konczak. Human sensori-motor development and artificial systems. In *on Artificial Intelligence, Robotics, and Intellectual Human Activity Support for Applications*, pages 303–314, 1997.

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

M. Schmerling, G. Schillaci, and V. V. Hafner. Goal-directed learning of hand-eye coordination in a humanoid robot. In *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2015 Joint IEEE International Conference on*, pages 168–175. IEEE, 2015.

J. L. Schwartz, L. J. Boë, N. Vallée, and C. Abry. The dispersion-focalization theory of vowel systems. *Journal of phonetics*, 25(3):255–286, 1997.

J. L. Schwartz, A. Basirat, L. Ménard, and M. Sato. The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25 (5):336–354, 2012.

D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

P. Shaw, J. Law, and M. Lee. Representations of body schemas for infant robot development. In *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2015 Joint IEEE International Conference on*, pages 123–128. IEEE, 2015.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

The Economist. Does growing up poor harm brain development?, 2018. URL https://www.economist.com/news/united-states/21741586-team-scientists-undertakes-ambitious-experiment-which-could-change-thinking-about [Online; visited 2018-05-13].

J. J. Thompson, N. Sameen, M. B. Bibok, and T. P. Racine. Agnosticism gone awry: Why developmental robotics must commit to an understanding of embodiment and shared intentionality. *New Ideas in Psychology*, 31(3):184 – 193, 2013.

M. Tomasello and M. Carpenter. Shared intentionality. *Developmental Science*, 10(1):121–125, January 2007. ISSN 1363755X. doi: 10.1111/j.1467-7687.2007.00573.x.

K. M. Tooley and K. Bock. On the parity of structural persistence in language production and comprehension. *Cognition*, 132(2):101–136, 2014.

S. Tremblay, D. M. Shiller, and D. J. Ostry. Somatosensory basis of speech production. *Nature*, 423(6942):866–869, 2003.

P. Wagner, Z. Malisz, and S. Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. ISSN 0167-6393. doi: 10.1016/j.specom.2013.09.008.

A. S. Warlaumont. Salience-based reinforcement of a spiking neural network leads to increased syllable production. *2013 IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL 2013 - Electronic Conference Proceedings*, 2013. doi: 10.1109/DevLrn.2013.6652547.

A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller. Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, 38(0):64 – 75, 2013a. ISSN 0893-6080.

A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller. Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, 38(0):64–75, February 2013b. ISSN 08936080. doi: 10.1016/j.neunet.2012.11.012.

A. D. Wilson and S. Golonka. Embodied Cognition is Not What you Think it is. *Frontiers in Psychology*, 4:1–13, February 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00058.

Y. Yoshikawa, M. Asada, K. Hosoda, and J. Koga. A constructivist approach to infants' vowel acquisition through mother–infant interaction. *Connection Science*, 15(4):245–258, December 2003. ISSN 0954-0091. doi: 10.1080/09540090310001655075.