Self-Organization in the Evolution of Speech

Second edition[1]

Pierre-Yves Oudeyer

September 16, 2019

[1] This document is a draft of a new edition of this book, to be published in 2020 by Oxford University Press.

To Adèle, Théophile, Arthur and Cécile

Contents

Preface
New Languages for Explaining Life

As the twentieth century was just beginning, only a few years after the discoveries of Darwin, another scientific giant revolutionized the life sciences. D'Arcy Thompson, a Scottish biologist, mathematician and classics scholars, proposed the use of a new language for explaining the origins of life forms, and that language was mathematics. Mathematics had always been used for measuring, counting and describing organisms and their structures. But causal explanations, including the theory of natural selection at the time, were formulated verbally, using natural language. Aware that the great variety of mechanisms underlying life forms were often complex biophysical processes that could not be adequately described using the existing terminology of natural selection alone, Thompson showed how it was also essential to use mathematics to *express* and *explain* processes of morphogenesis. Any explanation of why shellfish, jellyfish, the wings of dragonflies, radiolaria, rams' horns, or the skeletons of dinosaurs are the way they are had also to be couched in equations expressing processes of growth that are constrained by the laws of physics. It was not enough simply to point to reproductive advantages when seeking to explain how these forms came about.

Later, around the 1950s, a second language was added to the natural scientist's explanatory toolbox, namely the language of algorithms. Some mathematical models of the formation of structures, both in physics and biology, were too complex for their behaviours to be predicted by mathematical reasoning alone. The appearance of computers, along with theoretical advances in the algorithms that enabled computers to function, opened up a new possibility: describing processes of morphogenesis as algorithmic processes of interaction between the constituent parts of a system and simulating them digitally. This approach proved to be particularly useful in the study of morphogenetic phenomena in systems comprising large numbers of components interacting in a nonlinear fashion. In such systems, order will often appear spontaneously. Macro-structures emerge out of micro-structures that contain no macro-level plan: this is self-organization. Alongside the physicists Lorenz and Fermi who used the computer to model climatic phenomena and interactions between magnetized particles, two of the great architects of computing followed in Thompson's footsteps. The first of these was Alan Turing, who used algorithms to study how chemical reactions can spontaneously create patterns such as spots, stripes or spirals on the skins of leopards and fish, and to model the way in which the spherical symmetry of an egg cell can be broken during gastrulation, thus allowing the embryo progressively to take shape. Turing's work, like Thompson's, came to have a major influence on the development of embryology and of biology (even where this influence was not direct). As for the second influential figure, John Von Neumann, he first showed how algorithmic structures could reproduce themselves, and in so doing gave rise to profound new questions about the nature of life.

The brain, the cognitive representations and the behaviours that it generates are probably the most complex structures known to us. It is therefore unsurprising that today large communities of researchers use mathematics, algorithms and computers to model them. One of the most remarkable human faculties is language. Where does it come from? How was it created? How do new languages come about? A number of scientists, including Michael Studdert-Kennedy, Bjorn Lindblöm, James Hurford and Luc Steels, have put forward the idea that new languages can organize themselves in the course of local interactions between individuals, in the same sort of way that patterns on the skins of leopards are formed out of molecular interactions.

Luc Steels, in particular, as a pioneer and visionary, has shown how algorithmic models, together with their robotic extensions that take full account of the role of the body, constitute an essential scientific language in the understanding of the origins of natural language. It is from him that I learned this fascinating new perspective that forms the basis for the work presented in this book. I had the privilege of being part of his team at the Sony Computer Science Laboratory in Paris, where a large part of the scientific foundations of this work were laid down. To him I owe a huge debt of thanks.

This book focuses on some specific questions concerning the pillar of language that is speech: how did the structures of speech, vowels and the combinatorial re-use of phonemes in syllables come about? How can a community of individuals "invent" a new repertoire of vocalizations? What part could self-organization have played in the morphogenesis and evolution of speech? How does speech develops in infants, and how does the infant discovers that speech sounds can be used to influence others, i.e. for language? How does speech development relates to the evolution of speech? Using the methods that I have just outlined, the book examines these questions through the prism of embodied computational models. Readers hoping to find definitive answers will nevertheless be disappointed, for these are big questions, and much work remains to be done. They will, however, see how this new language of algorithmic and robotic models makes it possible to formulate questions in a new way, to look at them in a new light and to begin to perceive the contours of new hypotheses about the origins of speech and language.

Some of the work described here was begun in 1999 at the Sony Computer Science Laboratory and published as the first edition of this book in 2006[1]. The models described in the 2006 edition were principally those that I describe in chapters 6, 7 and 8 of this new edition. They are concerned with the mechanisms that enable a population of individuals to form and share culturally novel combinatorial speech sound systems. They also address the question of how one can explain both statistical regularities and diversity over the world's inventories of speech systems. Individuals in these populations are equipped with neural network models of multimodal perception, decision and unsupervised learning, as well as models of the vocal tract and the auditory system. They interact through hearing each other's random but systematic vocal babbling sounds, progressively attuning their initially unorganized vocalizations towards organized systems that share many fundamental properties with the speech sounds of today's world languages.

This work was subsequently extended substantially in several directions, which dictated the rewriting of several parts and the addition of much new material in the current edition. On one hand, discussions of how these models enable us to study old and new evolutionary hypotheses about the origins of speech has been enriched across the book, especially in

[1] Oudeyer, P-Y. (2006) Self-Organization in the Evolution of Speech, Oxford University Press

chapters 9, 10 and 11. On the other hand, some of the insights resulting from the models presented in chapters 6, 7 and 8 gave rise to a whole new series of research projects that I have been pursuing with several colleagues, and which are now reported in chapter 10 and 11. As these models showed the key importance of spontaneous vocal babbling in the self-organization of shared systems of speech sounds, this led us to study the evolutionary and developmental origins of spontaneous exploration. Here, vocal learning is conceptualized within the more general epigenetic framework of the sensorimotor and social development of the child. One key question was to understand better whether systematic vocal exploration, as displayed by most children, was the result of a language-specific motivational mechanism, or whether it could be the outcome of a more general motivational mechanism for exploration.

In this perspective, Frédéric Kaplan and I designed and studied algorithmic models of intrinsically motivated learning, a form of autonomous learning that that leads organisms to explore their own body and its relations with its environment "out of curiosity". This enabled us to demonstrate, as described in Chapter 10, how babbling and vocal imitation could be a collateral effect of more general developmental mechanisms, and opened up some novel hypotheses about the evolution of speech and language. With Clément Moulin-Frier and Mai Nguyen, I examined more closely how certain structures of vocal development observed in infants might emerge from these mechanisms of curiosity-driven exploration through a process of self-organization. Then, with Sébastien Forestier, we showed that children's discovery of language as a tool to influence the behaviour of others could self-organize automatically out of intrinsically motivated goal exploration. In intrinsically motivated goal exploration, organisms explore actively by generating and sampling their own goals with the intrinsic objective of discovering causal structures of their environment. These models, who were initially developped in this context of modeling child and language development, later on developped and provided foundations for new approaches to autonomous artificial intelligence and machine learning (Baranes and Oudeyer, 2013; Oudeyer, 2018; Laversanne-Finot et al., 2018), as well as new new theories to understanding curiosity in neuroscience (Kaplan and Oudeyer, 2007; Gottlieb et al., 2013; Gottlieb and Oudeyer, 2018).

These computational models of the role of self-organization and spontaneous exploration in the evolution of speech can be seen as complementary to the work of a number of other research teams who have followed, or are following, similar paths, and who have contributed to creating the vision that this book sets out to present. I am thinking in particular of Ahmed-Reda Berrah, Pierre Bessière, Louis-Jean Boë, Julien Diard, Hervé Glotin, Rafael Laboissière, Clément Moulin-Frier and Jean-Luc Schwartz in Grenoble; Chistophe Coupé, Jean-Marie Hombert, Egidio Marsico and François Pellegrino in Lyon; Bart de Boer in Brussels; Catherine Browman, Louis Goldstein and Michael Studdert-Kennery at Yale; James Hurford in Edinburgh.

The preparation of this book and the recent work that I present owe much to my colleagues at Inria. In particular, the Flowers team around me has provided invaluable support, of both a scientific and a practical nature, and for this I thank them all.

Publishing a science book addressing questions that are fundamental and at the frontiers of science, for a readership from a variety of backgrounds and with a variety of interests, can be a challenging task. I had the good luck to encounter people who, each in their own way, made it possible. They all took the time to immerse themselves in the project and their enthusiasm helped me greatly. James Hurford and Michael Studdert-Kennedy provided to me the encouragements and oppportunity to write and publish this book at Oxford Universty Press for the first edition. James Hurford also worked on the english translation of the first edition, while David Lewis worked on the translation of the second edition. The new edition has also benefitted from the support of Mikhail Gromov and Michel Cassé, whose work bringing the mathematical sciences closer to biological and human sciences echoes the universalism of D'Arcy Thompson.

Finally I want to thank my wife Cécile and my children Adèle, Théophile and Arthur, for the energy and the light they give me each day and which have been my guide throughout.

List of Figures

Chapter 1

The Self-Organization Revolution in Science

1.1 Self-organization: a new light on nature

Nature, especially inorganic nature, is full of fascinatingly organized forms and patterns. Mountains have the same silhouette when viewed at the scale of a rock, a summit, or a whole mountain range. Sand dunes often arrange themselves in long parallel stripes. Water crystallizes into symmetrical serrated flakes when the temperature is right. And when water flows in rivers and hurtles over cataracts, trumpet-shaped vortices appear and the bubbles collect together in structures which are sometimes polyhedral. Lightning flashes draw plant-like branches in the sky. Alternating freezing and thawing of the rocky ground of the tundra leaves polygonal impressions in the earth. The list of these forms rivals many human artefacts in complexity, as can be seen in Figure 1.1. And yet they are not designed or conceived by anyone or anything, not even natural selection, Dawkins' (Dawkins, 1982) 'blind watchmaker'. What, then, are the mysterious factors that explain their existence?

All these organized structures have something in common: they are the macroscopic outcomes of local interactions between the many components of the system from which they emerge. Their global organizational properties are not, however, to be found at the local level. Indeed the properties of the shape of a water molecule, as well as of its individual physico-chemical properties, are qualitatively different from the properties of ice crystals (see Figure 1.2), whirlpools, or polyhedral bubbles. The polygonal impressions in the tundra do not correspond with the shape of the stones composing them, and have a spatial organization quite different from the temporal organization of freezing and thawing. This is the hallmark of a phenomenon that is essential to the origins of forms in nature – self-organization.

This fundamental concept is the touchstone of the paradigm shift undergone by the sciences of complexity in the 20th century. Following the pioneering and visionary work of D'Arcy Thompson (Thompon, 1917), whose work in the early years of the 20th century centered on self-organized phenomena but did not then qualify them as such, the concept was recognized and gained momentum in the 1950s, driven by such figures as Alan Turing, William Ross Ashby, Heinz von Foerster, Ilya Prigogine, Francesco Varela et René Thom (Turing, 1952; Ashby, 1952; Nicolis et Prigogine, 1977; Kauffman, 1996; Ball, 2001). Ever since Newton, good science was supposed to be reductionist, and consisted in decomposing

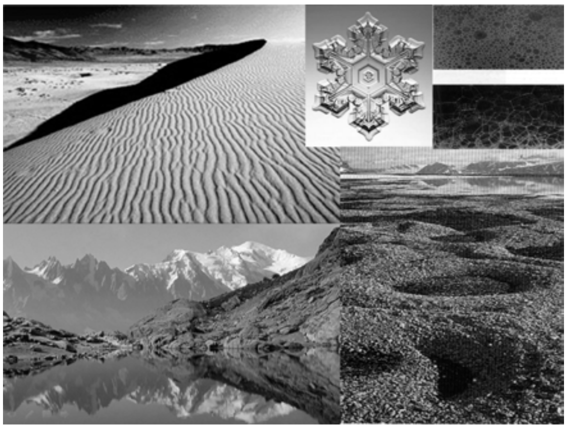


Figure 1.1: Nature is full of organized forms and patterns without there being anywhere any plans which might have served to build them; they are said to be self-organized. Here, parallel stripes running through sand dunes, water bubbles on the surface of liquid which has been stirred up and the polyhedral structures which are left when they dry out, an ice crystal, mountains whose shapes are the same whether one views them on the scale of a rock or a whole peak (Photos: Nick Lancaster, Desert Research Institute, Nevada; Burkhard Prause, University of Notre Dame, Indiana; Bill Krantz, University of Colorado).

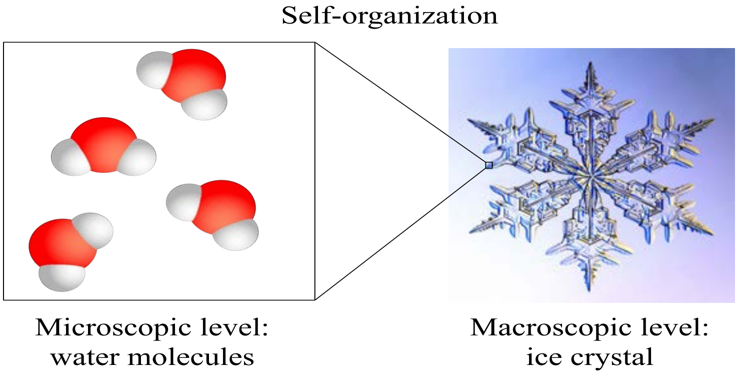


Figure 1.2: Self-organization of ice crystals. The macroscopic structure of the crystal self-organizes out of interactions between water molecules, whose microscopic structure is qualitatively different.

natural systems into simpler subsystems. For example, to understand the functioning of the human body, it was appropriate to study the respective parts, such as the heart, the nervous system, or the limbic system. Moreover, it did not stop there, and study of the nervous system, for example, was subdivided into study of the cortex, of the thalamus, or of the peripheral motor innervations, and each of these sub-parts was studied by hyper-specialists in separated dedicated university department. This method has obviously enabled us to accumulate an impressive bank of knowledge. But the prophets of complexity have broken up this paradigm. Their credo is “the sum of the parts is greater than the parts taken independently”. This is because nature is composed of complex systems with many interacting subsystems, and complex systems have a very strong tendency to self-organize. This includes even systems in biology, in which the ascendancy of natural selection is not total, but must work alongside self-organization (Kauffman, 1996; Ball, 2001).

This is why it now seems that many natural systems cannot be simply explained by a reductionist study of their parts. One of the most emblematic examples is that of the collectively built artefacts of insect societies (Camazine et al., 2003). Termites make immense nests, rising several metres above the ground, and with an architecture which is reminiscent of human structures, as Figure 1.3 shows. To try to explain how these structures are built, the study of individual termites, for example the precise study of all their neural wiring, is absolutely not sufficient. One could know everything about the anatomy and the brain of a termite, without ever understanding how their nests are built; because indeed no termite has the equivalent of a plan, even partial, of the superstructure. The know-how that they possess is infinitely more basic; it is of the type “if I come across a lump of earth, pick it up and place it where the pheromone signal is strongest”. The superstructure is rather the result of the dynamic interactions in the environment of thousands of termites, in the same way as the symmetrical structure of ice crystals is the result of the interactions of water molecules under certain conditions of pressure and temperature, and not a projection to the macroscopic level of structures already present at the microscopic level.

Concrete examples like this, of the use of the concept of self-organization, and of explanations of natural forms in terms of systemic properties, are now abundant, and are at the heart of the most advanced research of more and more physicists and biologists. Further examples include hunting or foraging patterns of bees and ants, the dynamic shapes of shoals of fish or flocks of birds, symmetrical patterns on butterfly wings, the regular spots on a leopard’s skin, the stripes of fish and shellfish, the magnetization of magnets, the formation of whirlpools in rivers, the birth of galaxies, demographic oscillations in predator-prey ecology, the formation of patterns in bacterial cultures and chemical reaction-diffusion systems, crystallization, lasers, superconductivity, the distribution of the sizes of avalanches, autocatalytic chemical systems, the formation of lipid membranes or the dynamics of traffic jams on freeways.

The sciences of complexity have thus demonstrated the fundamental usefulness of the concept of self-organization for the explanation of natural phenomena involving physical structures and certain biological structures characterizing the morphology or behaviour of simple animals like insects. We are now at the dawn of a new decisive phase in this scientific revolution: researchers in complexity theory are beginning to tackle the understanding of mankind itself, using these new tools. The understanding of the vital functions of the human body was the first to be transformed by the emerging wave of what is called “integrative” or “systemic” biology (Chauvet 1995; Kitano, 2002; Noble, 2006; Wilkinson, 2011; Klipp et al., 2016). Rather than concentrating on each organ in isolation, there is now an attempt to understand their complex interactions in an organism considered as a whole, in which each element is integrated with the others, at different scales in time and space. This has opened up new theoretical vistas, which began with Alan Turing’s mathematical models of morphogenesis (Turing, 1952), and more recently for the understanding of cancers (Kitano, 2004) and the functioning of the heart, as demonstrated by Denis Noble, one of the pioneers of the biology of systems (Noble, 2006), as well as other organs (Werner et al., 2017).

The advocates of self-organization do not stop there: the human brain, and thus the phenomena of sensation and thought, are also under the strong influence of features of

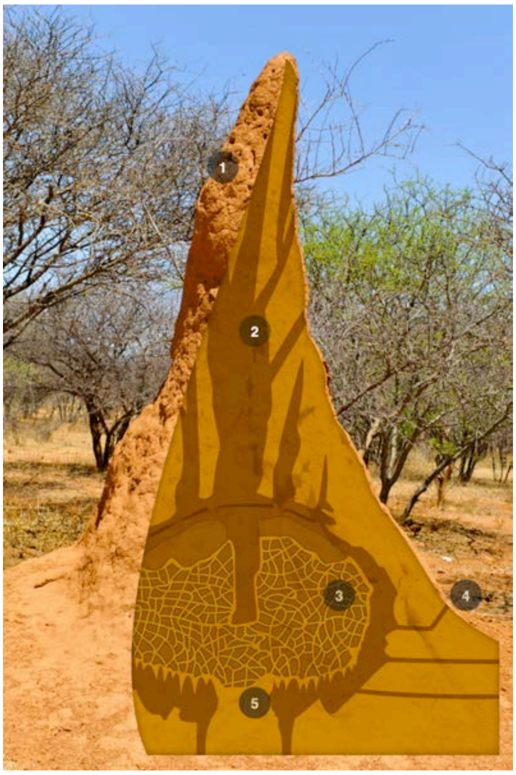


Figure 1.3: The architecture of termite nests is the self-organized results of the local interactions among thousands of individuals, none having any notion of an overall plan.

spontaneous organization in their structure. Indeed, the brain, composed of billions of neurons dynamically interacting among themselves and with the outside world, is the prototype of a complex system. For example, as we will show in this book, self-organization could be at the heart of the capacity of our brains to categorize the perceived world, that is, to organize the continuous flux of perceptions into atomic psychological objects.

But the main subject of this book goes beyond speculation about the brain as a self-organized system: we are today on the brink of a major advance in science, that of a naturalized understanding of what makes humans unique: their culture and their language. While culture and language have been the subjects of investigation by social sciences for centuries, our understanding of them has so far never been anchored in their material, biological substrate, i.e. in the set of all human brains in ongoing complex dynamic interactions in the physical world. The tools of complexity are now starting to make that possible. This book provides an illustration through one specific example: the origin and shaping of speech, one of the pillars of human language. Speech is the outward form and vehicle of language, consisting of combinatorial systems of sounds shared within language communities. In order to better understand the revolution that this question is now undergoing, we will first take a broad look at its historical development.

1.2 Language origins: a flourishing research field

It is an obvious fact, but nonetheless mysterious for that: humans speak. It is their main activity, an activity that sets them apart from the rest of the animal kingdom. Human language is a communication medium of unequalled complexity. It is a conventionalized code which lets individuals share their ideas and emotions with others, lets them talk not only about the colours in the sky above their heads, but also of distant landscapes, of past events, of how they imagine the future, of mathematical theorems, of invisible properties of matter, and of language itself. Besides that, each language defines a system which is peculiar to its speakers, an original way of organizing sounds, syllables, words, and sentences, and of spelling out the relationship between these sentences and the concepts which they convey. Around six thousand languages are spoken in the world today, and there is great diversity among them (Hagège, 2006; Hombert, 2009; Fitch, 2011). These languages are in a permanent state of change (Labov, 1994; Hombert, 2005), with some dying and others being born. The number of languages which have existed is estimated at over half a million. It is hard to imagine humanity without language. And yet, a long time in the past, humans did not speak.

This raises one of the most difficult questions in science: how did humans come to talk? A further question follows naturally: how do languages evolve?

These two questions, of the origin of the language faculty and of the evolution of languages, have been focussed on by many thinkers in centuries gone by, especially in the nineteenth century. They are prominent in Darwin's speculations (Darwin, 1859). Many such theories were developed without benefit of any empirical or experimental constraint. They were equally devoid of reasoned arguments and scientific method, to the point where the Linguistic Society of Paris ruled in 1866 that such questions should no longer be raised in the context of scientific discussion. This ruling initiated a century of almost total lack of progress in research in this domain.

Advances in neuroscience, cognitive science and genetics towards the end of the twentieth century have put these questions back into the centre of the scientific arena. On the one hand, modern neuroscience, thanks especially to new techniques of cerebral imaging, and cognitive science have made enormous progress in understanding the general functioning of the brain, and especially in the way in which language is acquired and processed in the brain. These developments have started to bring the study of language into the arena of the natural sciences, grounding the abstract systems which linguists describe in the biological and physical matter of which humans and their environment are composed. In short, natural sciences have taken over questions which were previously in the domain of the human and social sciences. This new light on the workings of language and the brain have provided researchers with the constraints whose absence undermined the speculations about the origins of language of the nineteenth century.

Progress in genetics has moreover turned the spotlight on neo-Darwinian theories of evolution, both confirming some of its foundations (with the discovery of genes, for example, along with some of their mechanisms of variation), and allowing its predictions to be tested, often successfully, thanks to the sequencing of the genomes of different species of animals so that we can reconstruct their phylogenetic trees and trace their evolutionary history. In particular, the sequencing of the human genome, along with that of other animals such as chimpanzees and monkeys makes it possible to specify the relationships between humans and their ancestors. Thus, driven by vigorous evolutionary biology, which simultaneously provides an impressive body of observations and a solid explanatory framework, the question of human origins has become a central theme in science. And quite naturally, the origins of language, this being one of the distinctive features of modern humans, has become, as in the nineteenth century, a beacon-like subject of research.

1.2.1 Interdisciplinarity

There is an emerging consensus among researchers who are today getting down to questions of the origin of the human language faculty and the evolution of languages – this research must be interdisciplinary. It in fact poses a puzzle with immense ramifications which go beyond the competence of each individual discipline on its own. Firstly, it is because out of the two big questions regarding the origin and the evolution of language there arise a number of subsidiary questions that are highly complex in themselves: What, in fact, is the language faculty? What is a language? How are sounds, words, sentences and representations of meaning related to each other? How does the brain represent and learn these sounds and sentences and the concepts which they convey? What are the respective roles of nature and nurture? What is language for? How does a language form and then change in the course of successive generations of speakers? What do we know of the history of each particular language? Why are the language faculty and languages the way they are? Why do we see universal tendencies and at the same time great diversity in languages? How does language influence the way we perceive and understand the world? Is language, like the appearance of eyes, the result of a phylogenetic evolution, or is it rather, like writing, a cultural invention? Is language an adaptation to a changing environment? An internal change improving an individual's chances of reproduction? Is it an exaptation, a side effect of changes which were not at first tied to communicative behaviour? What are the evolutionary prerequisites which paved the way for the capacity of speech?

Ranged against the diversity of these questions is an even greater diversity of research disciplines and methods. Linguists, even though they may continue to provide crucial data on the history of languages, are no longer the main actors. Developmental and cognitive psychology and neuropsychology carry out behavioural studies of language acquisition and language pathology, and these often reveal cognitive mechanisms involved in language processing. Neuroscience, especially with equipment for brain imaging allowing us to see which brain regions are active for given tasks, attempts to find neural correlates of verbal behaviour, to discover its organization in the brain. Some researchers also study the physiology of the vocal tract, to try to understand how we produce speech sounds. The physiology of the ear, the essential receptor in the speech-decoding chain (or vision, in the case of signed languages), is also a focus of research. Archeologists examine fossils and artefacts left by the first hominids, and try on the one hand to deduce our anatomical evolution (especially of the larynx) and on the other hand to get an idea of what activities they were engaged in (what tools did they make? and how did they use them? and what can these tools tell us about the degree of cognitive development?). Anthropologists do fieldwork on isolated peoples, and report on cultural differences, especially those related to languages and the meanings they convey. Primatologists try to evaluate the communicative capacities of other primates and to compare them with our own. Geneticists on the one hand sequence the human genome and that of potential ancestral species when it is possible to specify their phylogenetic relatedness, and on the other hand use genetic information from different people across the planet to help in reconstructing the history of languages, which is often correlated with the genetic history of their speakers.

Language thus involves a multitude of components interacting in complex ways in parallel on several scales of time and space. There is the ontogenetic timescale of the growth of an individual person, the glossogenetic or cultural timescale of the evolution of cultures, and the phylogenetic timescale of the evolution of species. Now, as we saw in the first part of this chapter, even though each of these components still needs to be studied independently in order to reduce the complexity of the problem, it is also fundamental to study their interactions. Many of the properties of language are probably not directly encoded by any of the components involved, but are the self-organized outcomes of the interactions of the components. These self-organizational phenomena are often complicated to understand or to foresee intuitively, and complicated to formulate in words.

1.2.2 Computer, mathematical and robotic modelling

This is why research into the origins of language today brings together computer scientists, mathematicians and specialists in robotics who build operational models of these interactions between the components involved in language, conceptualized as a complex system (Steels, 1997; Coupé et al., 2010; Christiansen and Chater, 2016). An operational model is one which defines the set of its assumptions explicitly and then computes or simulates the consequences of these assumptions, proving that a particular set of conclusions is entailed. There are two main types of operational model, that differ in the formal language used and in the uses to which they are put.

The first type involves abstracting a certain number of variables from the phenomenon of natural language formation and expressing how these variables are related and how they change over time in the form of mathematical equations. Usually this comprises systems of coupled differential equations and makes use of the framework of dynamic systems theory, and sometimes statistical physics, as for example in the work of Nowak et al. and Loreto et al. who studied the mechanisms by which lexical conventions are formed (Nowak et al., 2002; Loreto et al., 2012). When these models are sufficiently simple, their behaviour can be predicted analytically and formally proved from their mathematical structure alone. However, the abstractions and assumptions on which such proofs are based are sometimes far removed from physical, cognitive and social realities. Moreover, the formal language of mathematical relations is not always very suitable for explaining processes at work in nature (or in culture).

A second type of model has thus come into existence, where processes of morphogenesis are formulated in terms of algorithms. These algorithms are themselves expressed, in practice, using computer programming languages. Using this kind of formal language to describe a natural or a cultural process has two solid advantages. First, the great expressivity of these languages lets them formulate highly complex processes concisely (Dowek, 2011). Secondly, in the case of phenomena whose behaviour is very difficult to predict analytically from a series of equations, it can be possible to calculate this behaviour automatically through simulation. The programs can be run on computers in what is a simulation of a morphogenetic process, and researches can observe how the simulated system behaves with different parameters.

In research into the origin and evolution of language, this approach often involves building artificial systems in which individuals (their bodies, brains and behaviour), their interactions and their environment are modelled by programs. It is an approach that we will look at in detail in the following chapters, and which has already yielded decisive results that have opened the way to solving some fundamental unanswered questions. The pioneering work of Steels using this kind of algorithmic model (Steels, 1997; Steels, 2003; Steels, 2012; Steels,

2016) showed how linguistic conventions linking words to meanings can form spontaneously through pairwise interactions between individuals. Together with a number of colleagues he then extended these models so as to allow the formation of systems of shared semantic categories, that is to say methods by which individuals within the same community come to classify sense data in the same way (Steels, 2003; Kaplan, 2001). This work sheds light on the similarities and dissimilarities to be found among the world's languages in the way colours, for example, are named and categorized (Steels and Belpaeme, 2005). More recently these models have been applied to the formation of grammatical structures (Spranger et Steels, 2012; Steels, 2016), complementing other pioneering work from Hurford, Kirby, Smith and their colleagues on the evolution of syntax (Kirby and Hurford, 2002; Kirby et al, 2014), but also that of Coupé and Hombert (Coupé and Hombert, 2005). At the same time, other models have been developed for investigating the formation of speech structures, including those of Berrah et al. (Berrah et al., 1996; Berrah and Laboissière, 1999), Browman and Goldstein (Browman and Goldstein, 2000), de Boer (de Boer, 2001), Oudeyer (Oudeyer, 2005a,c) and Moulin-Frier et al. (Moulin-Frier et al., 2008).

Some of these algorithmic models have in recent years been hybridized for implementations that combine the use of computers and robotics. These models, instead of representing the brain-body-environment system purely as computer algorithms, make use of programs running on computers that are embedded in physical robots. Here, only the "brain" is formulated algorithmically, while the body is modelled using mechatronic elements, and the environment approximates to humans' real environment. This approach is currently at the heart of developmental robotics, where there is intense activity around the modelling of how sensorimotor, cognitive and social skills develop (Oudeyer, 2010), particularly regarding the acquisition and evolution of language (Steels, 2003; Cangelosi et al., 2010; Cangelosi and Schlesinger, 2015; Oudeyer et al., 2018). Examples are discussed in Chapter 10, describing experimental work using robots endowed with a system of intrinsic motivations that spontaneously explores their own body out of a form of artificial curiosity, and progressively learn to move about, manipulate objects, produce vocalizations and predict their effect on other individuals.

Building this kind of robotic model is interesting from several points of view. First, regarding language, it addresses the *symbol grounding problem*, in other words the fundamental problem of how symbols are grounded in the physical world. The problem here, so eloquently formulated by Steven Harnad (Harnard, 1990) is understanding how the symbols commonly used for describing and modelling languages (such as words and grammatical rules) can become meaningful in the physical and social reality of a real organism. In particular, it involves a capacity to link the abstract world of symbols to the concrete world of numerical and chemical quantities that are perceived and manipulated by the brain and the body in context. It is hard to see how models that are purely algorithmic, implemented entirely on computers that *in fine* represent the world in a symbolic, discrete way, might usefully inform any questions about symbol grounding. Hybrid algorithmic-robotic models, on the other hand, that by definition are at the cutting edge between the symbolic and the physical worlds, are extraordinary tools for studying this problem (Steels et Kaplan, 2001; Steels, 2012). Here algorithmic models of language are confronted with physical and social reality, and whether or not there is an effective grounding of symbols – a strong constraint on the plausibility of these models – can be tested empirically. Chapter 4 describes the Talking Heads experiment, performed by Steels and his colleagues, in which a group of robots invent and progressively negotiate a new language that enables them to communicate about objects in their environment.

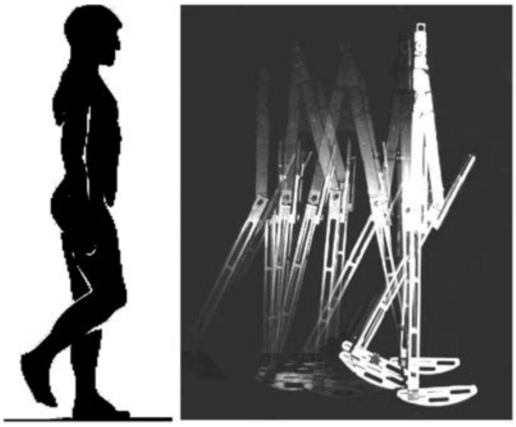


Figure 1.4: Tad McGeer's passive robot Walker (on the right). A mechanical structure with neither computer nor internal power supply spontaneously exhibits a natural, balanced walking movement when placed on a slope, showing that the interaction between the morphology of the body and the force of gravity can be enough to generate structured movement. This has helped us understand the extent to which the body itself could have engendered walking in humands. In the same way, it is important to take account of the properties of the body when modelling the formation of systems of language, and of speech in particular. Here, robotic models are an invaluable aid. (photo: Tad McGeer)

There is another very good reason for using robots. As Esther Thelen and Linda Smith, for example, have argued in their theory of development (Smith and Thelen, 2013), the formation of behavioural and cognitive structures results from dynamic interaction between the brain, the body and the environment. The body and the environment, whose physical substrate gives rise to particular properties of structure generation, have a crucial role. Embodiment, that is to say the material composition and the geometry of a body and its sensorimotor system, can dramatically simplify the acquisition of certain behaviours. Concerning bipedalism, for example, Tad McGeer showed that a purely mechanical leg structure (see Figure 1.4), with no computer and not even any power supply, could with the right geometry start to move forward in steps that have many properties in common with human walking (McGeer, 1993). Concerning the acquisition of language, Chen Yu and Linda Smith showed how geometrical hand-eye relations and the physical manipulation of objects could create favourable situations for learning the meanings of first words (Yu and Smith, 2012; Smith et al., 2018). Thus, the body carries out physically a type of information processing, sometimes referred to as "morphological computation" (Pfeifer et al., 2007). In this context, robots make it possible to model mechatronically – in a straightforward, realistic way – interactions between the brain, the body and the environment that would be far too complex or even impossible to model algorithmically. In the models of morphogenesis of speech that will be described in chapters 6, 7 and 8, we will see how simulating the physical properties of the vocal and auditive systems can provide a means of investigating the specific role of the body in the formation of systems of vocalizations among a population of individuals.

In other spheres, many other examples can be found today of robotic models being used to enhance understanding of animal and human behaviour [1], concerning such varied phenomena as navigation and phototropy in insects, control of locomotion in dolphins, distinguishing between self and non-self in human infants, but also the impact of the visual system on the formation of linguistic concepts.

1.2.3 Exploring the evolution of speech *in silico*

The work to be presented in this book belongs in this methodological tradition of building and exploring artificial models *in silico*. It will concentrate on the origin of one particular aspect of language: speech sounds and the systems of vocalizations that are formed out of them. Vocalizations, as explained in detail in the next chapter, constitute a conventional code providing each language with a repertoire of forms for conveying its messages. This code, which has both an acoustic and an articulatory element, organizes sounds into categories which are special to each linguistic community, and regulates the ways in which they can be combined (these rules of sound-syntax are also cultural conventions). The system is thus discrete and combinatorial. Without such a code, which could also be implemented in a manual modality for signed languages, there could be no form, no content, and hence no linguistic communication. How could such a code have arisen? In particular, how were the first codes able to form themselves before there was any conventional linguistic communication, of which they are prerequisites? Why are the sound patterns of human languages the way they are? These are the questions we will discuss in this book. They might be seen as very ambitious, since they include many complex aspects, both individual and social, but at the same time very modest in relation to the overall programme of research into the origin of language and the evolution of languages. In fact, they only concern the origin of one prerequisite of language among many others (such as the capacity to form symbolic representations or the pragmatic capacity to infer the intentions of others by means of behavioural signals).

This book does not aim to suggest definitive answers. The origin of speech holds many mysteries that science is only just beginning to uncover. The artificial systems that I present and the *in silico* investigations that these systems allow are aimed first and foremost at developing our intuitions. Through them, complexity can be handled in a concrete fashion, doors opened that previously seemed firmly shut, new hypotheses explored and, perhaps more importantly, new questions formulated that can change our perspectives and shed fresh light on our quest for the origin of speech. As Niels Bohr famously said,"Every sentence I utter must be understood not as an affirmation but as a question.".

Chapter 2 presents an overview of speech structures as we understand them today and some of the challenges to be overcome in understanding how speech originated. Chapter 3 locates the problems of the origin of speech in the general framework of the origins of form in biology. It discusses the interrelationships linking the phenomena of self-organization and natural selection, two creators of forms in the living world. This leads to the question of how we should expect any potential explanation of the origin of living forms to be structured. Chapter 4 makes a detailed survey of the literature, reviewing answers that have already been proposed, and defends the approach that is at the heart of this book. Chapter 5 describes my methodology, that is the building of artificial systems, along with the goals and

[1] See for example my review (Oudeyer, 2010).

scientific philosophy which drive it. Chapter 6 gives a formal description of an initial artificial system, and explains its operation. We see the emergence of a discrete, combinatorial speech code shared by a population of individuals who at first only pronounce unstructured, holistic vocalizations, and do not follow any rules of coordinated interaction. We will discuss, in particular, the role which morphological constraints on the vocal and perceptual apparatus may or may not play in the formation of speech codes. Chapter 7 presents a variant of this artificial system in which, in contrast to Chapter 6, there is no assumption that individuals are capable from the outset of retrieving articulatory representations of the sounds which they hear; this capacity will be learned, thanks to a quite generic neural architecture. This system will also use a model of the human vocal tract for producing vowels, which will enable us to specify the analogy between artificial and human systems: it will be shown that the statistical regularities which characterize the vowel systems of populations of artificial individuals are very similar to those of the vowel systems of human languages. Chapter 8 presents an extension of the artificial system of Chapter 6, and shows how rules of sound syntax, termed “phonotactic rules”, can arise. Chapter 9 discusses novel evolutionary scenarios that these *in silico* experiments might suggest. In particular, we will explain how conventionalized, combinatorial speech codes could have emerged spontaneously out of two pre-linguistic mechanisms. The first of these mechanisms is vocal imitation, which has functions outside language. The second is the spontaneous exploration of the body from intrinsic motives that push the organism to discover its sensorimotor space out of pure curiosity, and in particular its vocal space through spontaneous babbling. Chapter 10 focuses on these mechanisms of curiosity-driven learning, studying how they could play a crucial role in the origins of speech and the discovery of communication in individuals. It outlines computational models of intrinsically motivated spontaneous exploration, and presents a series of experiments with embodied agents. These experiments show how a series of developmental stages of increasing complexity can self-organize, enabling individuals to make ratcheting discoveries. These discoveries include how to move their body, how to interact with objects, and how to produce speech, participate in speech interaction with others, and finally use speech as a tool to act and achieve joint tasks with others others. Together with self-organization arising from coupled neural maps and bodies, curiosity-driven learning is shown to be a key driving force for the origins of cognition. This opens new perspectives to understand natural intelligence, as well as to build flexible autonomous artificial intelligence.

Chapter 2

Human Speech Structures

Language, and more generally communication, involves the transmission of information between individuals. This needs a physical medium. The medium we currently use most is the sound of the human voice, and it is structured in a very particular way. The capacity for producing, perceiving and structuring the stream of sounds is called 'speech' [1]. Other media are manual signs (for sign languages) and writing. The sounds which humans use for speaking are organized into a structure, sometimes referred to as the speech "code", which provides a repertoire of forms that are used as a physical vehicle for information. This code is a prerequisite for communication in language. Without such a repertoire of forms, whether acoustic or gestural, there is no way of transferring information between individuals. The speech code is mainly a matter of convention, and there is an enormous diversity among languages. The code regulates the manner in which vocalizations are organized (which is discrete and combinatorial), the way in which sounds are classified, and the ways in which they can be sequenced (defining the rules of sound-syntax). We will now describe how this speech code is organized, focusing first on how sounds are produced and perceived.

2.1 The instruments of speech: the vocal tract and the ears

We are dealing with a complex musical instrument: the vocal tract. This is organized into two subsystems (see Figure 2.1): one generates a sound wave, the other shapes it. The first system is subglottal. Together, the lungs and the diaphragm cause air to pass through the trachea, which vibrates the larynx. The larynx is an assemblage of cartilages and muscles, which generates a sound wave when it vibrates. The generated sound is made up of a great number of frequencies. The second system, the supra-laryngeal, is a tube stretching from the larynx upward and forward, dividing into the nasal and oral passages, and ending at the lips and nostrils. The organs of the glottis, the velum, the tongue (tongue body and tongue tip) and the lips make it possible to modify the shape of this tube, in particular its length and volume. This change in shape results in the weakening or strengthening of certain frequencies in the sound signal. Thus the production of sound with the vocal tract resembles the production of sound with a flute: you blow at one end, and air passing

[1] Note that the term "speech" is reserved here for acoustic and articulatory events independent of meaning and is thus not a synonym for "language".

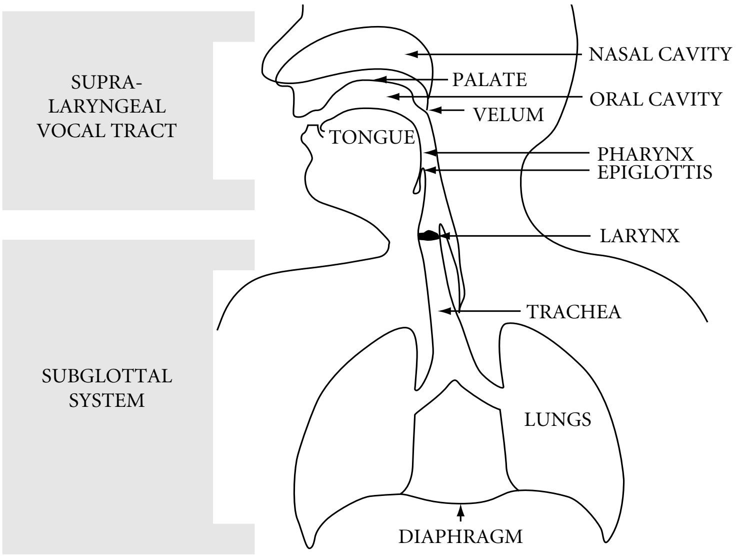


Figure 2.1: The vocal tract is organized into two subsystems: the subglottal system which provides a source of sound, and the supra-laryngeal system whose changeable shape makes it possible to shape this sound wave (adapted from Goldstein, 2003b).

through the 'whistle' makes a sound composed of many frequencies, which can be modified by blocking the holes in the body of the instrument. Thus, speech reduces to moving the various organs of the vocal tract.

The perception of sounds is carried out by the ear, in particular the cochlea (Figure 2.2). The cochlea is the device which lets us register a certain number of parameters of sound. Among these parameters is the decomposition of a sound into its component frequencies, or harmonics. In fact, every complex sound can be seen as a superposition of sine waves, each with a given frequency and amplitude. In mathematical terms, this is called a decomposition into a Fourier series. The cochlea performs an approximation to this decomposition, due to the basilar membrane. This membrane increases in thickness, and in its thinnest parts it responds better to high frequencies, while in its thicker and heavier parts it responds rather to low-frequency stimuli, corresponding better to its own inertial properties (Figure 2.3) . Nerve cells called "hair cells" are linked to this membrane to gather information about the stimulus. This information is passed by a network of fibres to the central nervous system.

2.2 How our brain plays the instruments of speech

How do the organs of the vocal tract control the flow of sound? What representations does our brain use to produce sounds? How is the connection between production and perception managed? These are the questions addressed by articulatory phonology (Browman and Goldstein, 1986), which is the approach adopted in this book.

The central concept of articulatory phonology is the *gesture*. A gesture, a unit of action, is the coordination of a certain number of organs (e.g. the tongue, the lips) to bring about a constriction in the vocal tract. A constriction obstructs the passage of the sound wave. It

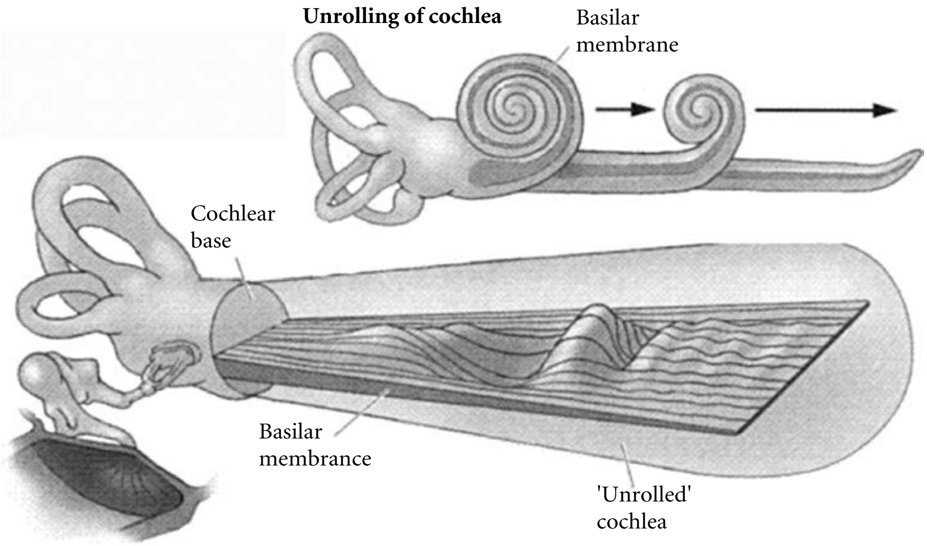


Figure 2.2: The cochlea, speech perception organ. Its basilar membrane can decompose the sound wave into a Fourier series, that is it can calculate the amplitude of its various harmonics (adapted from Sekuler and Blake, 1994).

is a narrowing of the vocal tube. For example, the words “boat”, “parameter” and “method” all begin with a closure of the lips. A gesture is specified not by the trajectory of one or more organs, but by a constriction target defined by a relationship between organs. For example, the opening of the lips is a constriction variable which can be implemented by the movement of three organs, the upper and lower lips and the jaw (each controlled by a set of muscles). An articulatory target (a constriction defined by a relationship among organs) can be realized by several different combinations of movements by the organs.

The constriction variables, or systems of constriction, used to specify gestures are the position of the larynx, the velum, the tongue-body, the tongue tip and the lips. Each of these constriction variables can be controlled by the movement of several organs and the many muscles which activate them. Figure 2.4 schematizes the constriction variables along with the organs which can implement them.

Each constricting organ can produce gestures whose constriction varies along two continuous dimensions: place and manner of articulation. Among the places of articulation, that is the places in the vocal tract where the narrowing occurs, the following can be mentioned: bilabial, dental, alveolar, palatal, velar, uvular and pharyngeal. Figure 2.5 gives further examples. Narrowing can be achieved in a variety of that include stops (e.g. [d]), fricatives (e.g. [z]) and approximants (e.g. [r]). Gestures which involve a severe narrowing or a total blockage are called consonantal, and those which involve a wider articulation are called vocalic, i.e. they are vowels.

When we speak, several gestures can be performed in parallel. Our vocalizations are thus the parallel temporal combination of several gestures. This combination can be represented by a “gestural score”, by analogy with musical scores (the gestural score shown in Figure 2.6 is for the word “pan”). Gestures in five constricting systems (velum, tongue tip, tongue-

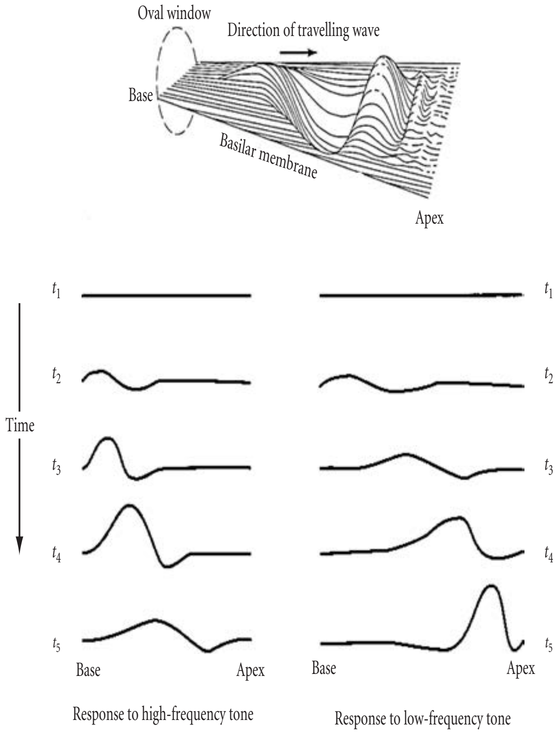

Oval window
Direction of travelling wave
Base
Basilar membrane
Apex
Time
t_1
t_2
t_3
t_4
t_5
Base
Apex
Response to high-frequency tone
Response to low-frequency tone

Figure 2.3: The basilar membrane decomposes the signal into its harmonics: it increases in thickness, and in its thinnest parts responds better to high frequencies, while in its thicker and heavier parts it responds rather to low-frequency stimuli, corresponding better to its own inertial properties (adapted from Escudier and Schwartz, 2000).

body, lips, glottis) are represented on five different lines. The shaded boxes represent the time intervals during which the gestures of each constrictor are active in the vocal tract. The labels on the boxes indicate the place and manner of constriction.

Some phonologists use the concept of a phoneme to describe the sounds found in words. This presupposes that words can be segmented into sequences of units called phonetic segments (although these do not necessarily correspond to the letters of the word when written). A unit is an element that distinguishes two words such as “bar” and “par”. It is possible to define phonemes in terms of the gestures that produce them and the way in which these gestures are organized: a phoneme can be seen as a set of gestures (often just one) which systematically recur in many words in a regular scheme of coordination.

So when we produce an utterance, it is gestures that are specified. Commands are sent to the organs so that they they can implement the corresponding constrictions. Speech production is thus organized on two levels: the level of the commands which define the gestural score, and the level of implementation. The first level is intrinsically discrete (but not necessarily discrete in the way that vocalizations as a whole are discrete, as we will see). The second level is intrinsically continuous, since it corresponds to a trajectory of the physical organs. The acoustic wave which is produced is related in a fixed but complex manner to this trajectory. In fact, the nonlinear physics of the vocal tract are

	tract variable	articulators involved
LP	lip protrusion	upper and lower lips, jaw
LA	lip aperture	upper and lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

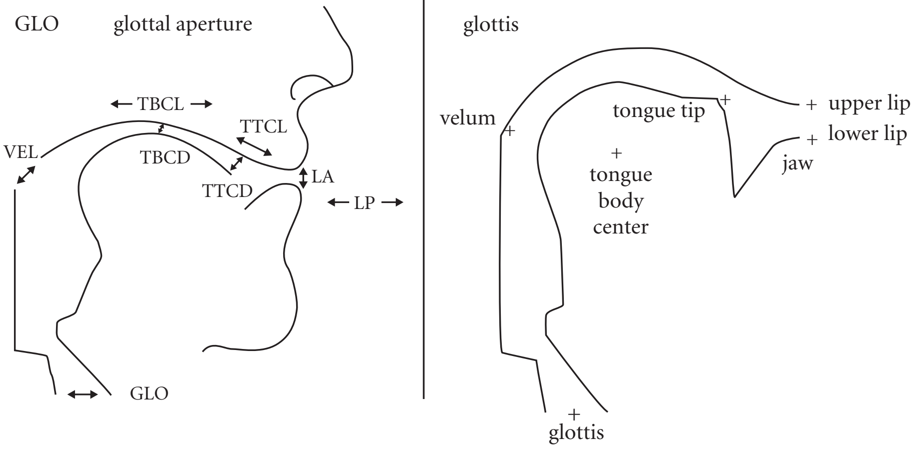


Figure 2.4: Constriction variables are used to specify gestures. Each of these variables can be controlled by the movement of several organs and many muscles which activate them (adapted from Goldstein, 2003b).

such that many different vocal configurations produce the same sound; or yet again certain configurations which are very close in articulatory terms are very different acoustically. The electro-mechanical properties of the cochlea also complicate the function which computes the correspondence between the perception of a sound and the motor program which implements it.

Articulatory phonology theory hypothesizes that gestures and their coordination are represented in the brain not only for controlling the production but also for the perception of sounds. In fact, gestures are the representations which make a connection between perception and production possible. The Motor Theory of Speech Perception (Liberman and Mattingly, 1985) takes this idea further in proposing that it is muscular representations that are used by the brain for classifying sounds – although this theory has been contested (Moulin-Frier et al., 2012). Thus the brain should be capable of transforming gestural representations into muscular representations (to control the speech organs) and vice versa. Figure 2.7 summarizes this organization.

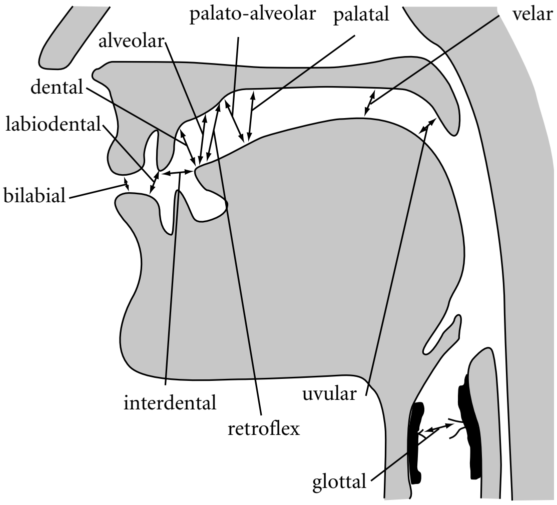


Figure 2.5: Places of constriction range between the larynx and the lips (adapted from Escudier and Schwartz, 2000).

2.3 The organization of the speech code

Comparison of the gestural scores forming words shows striking regularities both within one language and between languages.

2.3.1 Discreteness and combinatoriality

The complex vocalizations that we produce are coded phonemically. This has two implications: First, in any language, the articulatory and acoustic continuum which defines gestures is broken up into discrete units; secondly, these units are systematically re-used to construct the representations of the next higher linguistic level, such as the syllable level. This is what we mean here by combinatoriality.

It would be possible to imagine that each syllable is specified by a gestural score consisting of unique gestures or unique combinations of gestures. For comparison, there are writing systems which use a unique holistic symbol for each syllable (Nakanishi, 1998). These are syllabaries, as opposed to alphabets. In fact, in contrast with writing systems, all human languages have repertoires of gestures and combinations of gestures which are small in relation to the repertoires of syllables, and whose elements are systematically re-used to make syllables. In the languages of the $UPSID_{451}$ database (UCLA Phonological Segment Inventory Database) initially elaborated by Maddieson (Maddieson, 1984) and containing 451 languages, the average is about thirty phonological segments per language. More precisely, 22 consonants and five vowels are the most frequent counts, as figures 2.8 and 2.9 show. This small number (22) has to be contrasted with the considerable number of phonemes that we can possibly produce: for one thing, the gestures can vary the place of their constriction continuously from the larynx to the lips; they can also continuously vary the manner of articulation (that is, the shape and degree of narrowing). Thus, as phonemes are combinations of gestures, it is clear that the combinatorial possibilities are immense. Moreover, we have

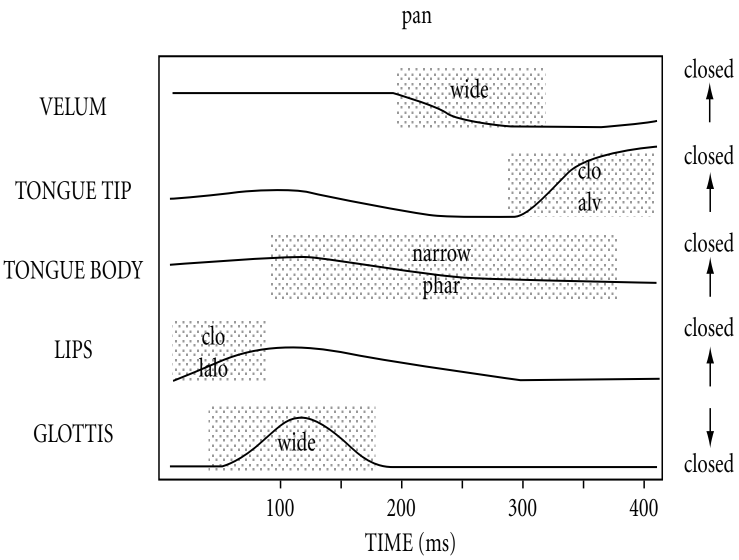


Figure 2.6: An example of a gestural score: the word "pan" (adapted from Goldstein, 2003b).

examples of languages like !Xu (a member of the Khoisan family of languages in southern Africa) that, impressively, use more than 140 phonemes (but such languages are very rare). The total number of phonemes in the UPSID database is 920. Nevertheless, despite all these possibilities, any given language will generally use only a handful of phonemes, recombining them in a systematic fashion to form a few hundred syllables that are in turn recombined to form tens of thousands of words.

This phenomenon of systematic re-use also applies to the gestures themselves, which are recursively constructed on this principle. In fact we have just explained that the place and manner of constriction which specify a gesture may vary continuously. Now in a given language, only a small number of places and manners occur and are re-used (varying the combinations, for sure) to make the gestures. Gestures could in principle have places of articulation which are peculiar to each, while still being re-used in many syllables, but this is not what happens. For example, for each separate manner of articulation, 95 percent of languages only use three places.

Thus, from the vast continuous space of possible gestures, speech carves out basic building blocks which it re-uses systematically. The phonemic and gestural continuum becomes discretized. Speech, already discrete in its manner of control involving motor commands specifying articulatory targets, is now also discrete in terms of the systems it uses (that is, the possible articulatory targets in a language are limited to small finite numbers, although physically they could be distributed across the whole articulatory continuum) and combinatorial. There are thus two remarkable points: first, the discretization of the continuous space of gestures; and second, the three levels of systematic re-use of resources:

- places and manners of articulation are re-used to make gestures
- gestures and their combinations are re-used to form syllables
- syllables are re-used to form words

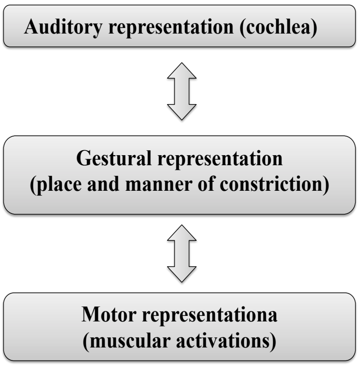

Auditory representation (cochlea)
Gestural representation
(place and manner of constriction)
Motor representationa
(muscular activations)

Figure 2.7: The brain handles three representations in the perception and production of speech: an auditory representation, a muscular representation and a gestural representation, which is the representation used to classify sounds. The brain can pass from one representation to another, according to the theory of articulatory phonology (adapted from Goldstein, 2003b).

2.3.2 A code shared culturally

All the speakers of a given language perceive and classify sounds in more or less the same way. However, speakers of different languages can perceive and classify sounds very differently. Every linguistic community has its own system for categorizing vocalizations, which makes it a form of cultural convention. We shall now examine this property of human speech in more detail.

For a start, speech perception is characterized by a psychological uniformity among the members of a linguistic community, despite the great physical variability of speech. Physically different sounds associated with different trajectories of the organs of the vocal tract can in fact correspond to the same sound, psychologically speaking, as the [d] in "idi", "ada" or "odo" for speakers of English. The sounds are different because the gestures which specify the [d] are superimposed on the gestures which specify the [i], [a] and [o]: the articulatory targets are temporarily in competition with each other, and the organs have to make a compromise to satisfy them as well as possible. The result is that the specifications of the gestures are not exactly met, but are approximated by the low-level muscular system. This is the phenomenon of co-articulation, which explains the acoustic and motor variability of sounds. Co-articulation occurs not only when the contexts of a segment change, but also when the rhythm of speaking changes. Thus each phoneme in a language can be realized in several different ways while still remaining identifiable by speakers of the language. The variants of a phoneme which speakers are able to recognize are called allophones. It is in the space of gestures that variations are the weakest: co-articulation can bring into play very variable muscular or acoustic trajectories, which nevertheless more or less preserve the articulatory targets in terms of the relationships among the organs. This is why gestural representation is central to speech. However, even if the level of motor commands is invariant, the level of their realization involves variability which is managed in precise and particular ways in each language. In a given language, all speakers decide in the same way which sounds are variants of the same phoneme and which are variants of different phonemes. This organization of the space of sounds is culturally specific to each language. For example, native speakers of Japanese identify the [r] of "read" and the [l] of

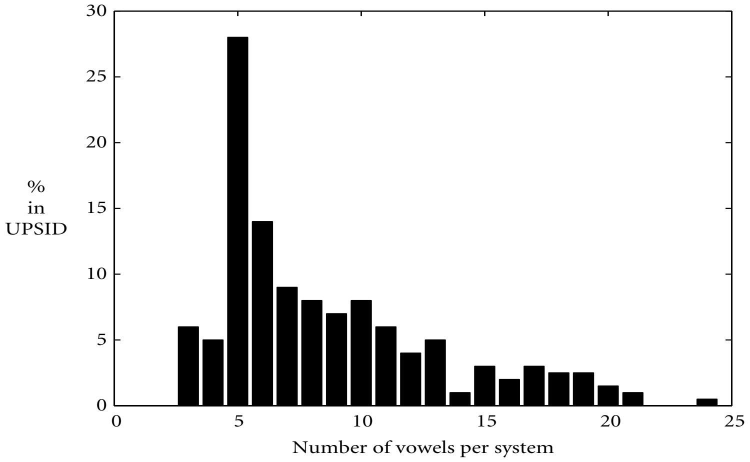


Figure 2.8: Distribution of counts of vowel inventories in the languages of the UPSID database (adapted from Escudier and Schwartz, 2000, Vallée, 1994).

“lead” as allophones, that is they classify these two sounds under the same heading, whereas for English speakers these are two distinct categories.

Not only do speakers of the same language have a shared way of classifying sounds which is specific to that language community, but they also have a shared way of perceiving sounds subjectively: they have the same “sensations” of sounds. Speakers of other languages will have different sensations. This is shown by the “perceptual magnet” effect (Kuhl et al., 1992). When subjects are asked to judge the similarity between two sounds (on a scale of 1 to 10), and where these sounds are regularly spaced at a given distance D in a physically defined space (e.g. a spectrum of amplitudes of Fourier components), it is noticed that when two sounds belong to the same phonemic category, the similarity reported by the subjects is greater than that which they report for two sounds which do not belong to the same phonemic category but which are nevertheless the same distance apart on the physical scale. To summarize, intra-categorial perceptual differences are diminished, and inter-categorial differences are augmented. It is a kind of perceptual deformation or acoustic illusion (“perceptual warping”), in which the centres of categories perceptually attract the elements of the categories like magnets. Once again, this a culturally specific phenomenon: the perceptual deformations are particular to each language. Figure 2.10 shows this effect.

2.3.3 Regularities in human languages

Statistical study of languages shows universal tendencies in the inventories of phonemes and the gestures which compose them. Certain phonemes are very frequent, while others are very rare: 87 percent of languages in UPSID have the vowels [a], [i] or [u], while only five percent have [y], [œ] or [ɯ]. More than 90 percent of languages have [t], [m] and [n] in their inventories, as Figure 2.11 shows. It is the same with gestures, and in particular the places and manners of articulation; 15 percent of languages use alveodental and bilabial places, but only 3 percent use retroflex and uvular places. Similarly, 38 percent of languages include plosives, while only 3.9 percent have trills, taps or flaps (Vallée, 1994; Schwartz et al., 1997b).

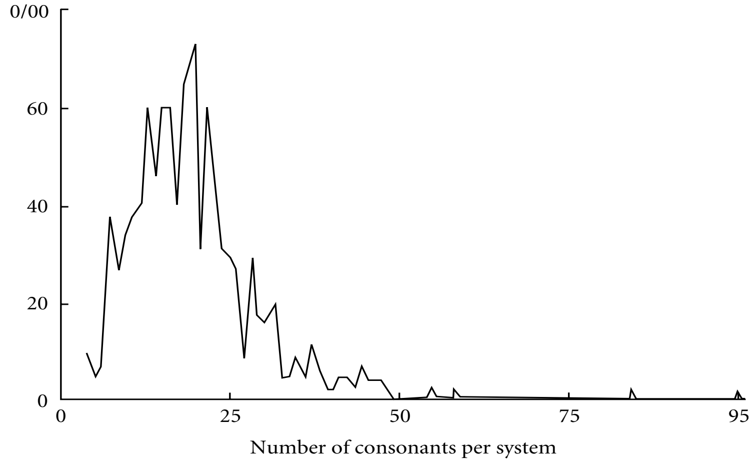


Figure 2.9: Distribution of counts of consonant inventories in the languages of the UPSID database (adapted from Escudier and Schwartz, 2000, Vallée, 1994).

The regularities do not apply only to phonemes individually, but also to the structure of phoneme inventories. This means for example that if a language has a front unrounded vowel of a certain height, like the [e] in "bet", it will also usually have a back rounded vowel of the same height, like the [aw] in "hawk". Thus the presence of certain phonemes is correlated with the presence of other phonemes. Some vowel systems are also very common, while others are rarer. The five-vowel system /[i], [e], [a], [o], [u]/ is found in 28 percent of languages, as Figure 2.12 shows (Vallée, 1994).

There are also regularities governing the ways in which phonemes combine. In a given language, not all possible phoneme sequences are allowed. Speakers know this (whether consciously or unconsciously), and if they are asked to invent a new word, it will be made up of non-arbitrary phoneme sequences (some will never be used). For example, in English "spink" is a possible word, while "npink" or "ptink" are not possible. Here again, the rules governing the possible ordering of phonemes, called phonotactics, are cultural and particular to each language. In Tashliyt Berber, "tgzmt" and "tkSmt" are allowed, but they are not allowed in English. Generally speaking, at the level of the syllable, which can be broken down into an ordered sequence of phonemes, we remark that the phonemes of a language can only occupy specific slots in this sequence. There are languages like Japanese where syllables can comprise two phonemes only; consonants are only allowed in the first slot and vowels in the second slot (these syllables are termed "CV"). Sometimes it is groups, or clusters of phonemes that can only occur in certain places in the sequence.

Moreover, there are phoneme combinations which are statistically preferred over others in the languages of the world. All languages use syllables of the CV type, while many do not allow consonant clusters at the beginnings of syllables. Statistically, languages prefer CV syllables, then CVC syllables, followed by CC, CCV, and CVVC/CCVC/CVCC. Syllables tend to begin with a phoneme with a high degree of constriction, and then to let the degree of constriction decrease until the middle of the syllable, and then to increase the level of constriction again up until the last phoneme of the syllable. This is known as the "sonority hierarchy".

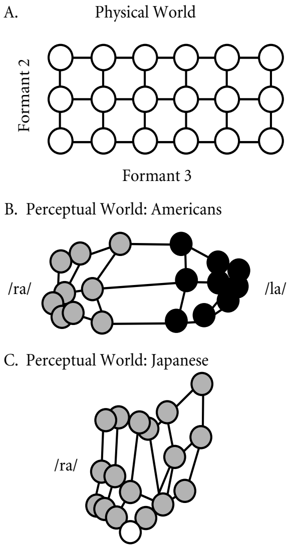


Figure 2.10: Kuhl et al. required subjects to rate pairs of consonants for similarity on a scale of 1 to 10, where these consonants varied continually between [r] and [l] (they were followed by the [a] vowel), and their values are represented by the circles in the figure above (Kuhl et al., 1992) . One group of subjects were American, the other group were Japanese. It is possible to derive from Kuhl's results a graphic representation showing the subjective way in which they perceived each consonant. This graph of their subjective perception is given in B for the Americans and in C for the Japanese. Note that the Amrericans subjectively perceive two categories of sound in this continuum, whereas the Japanese only perceive one. Moreover, within the neighbourhoods of [r] and [l] for the Americans, the sounds are subjectively more similar to each other than they are when measured objectively in physical space (adapted from Kuhl et al., 1992).

2.4 Diversity across human languages

Some of the regularities mentioned in the previous section are systematic and common to all languages. This is the case with the re-use of phonemes and gestures, as well as their discretization; it is also the case with the shared classification of sounds by speakers of each language, as well as with acoustic illusion phenomena related to speakers' knowledge of the phoneme inventories of their languages.

On the other hand, regularities in the inventories of phonemes and gestures, as well as the phonotactic preferences, are only statistical. This sheds light on a striking aspect of the speech systems of the world – their diversity. Recall that the languages of the *UPSID* database include 177 vowels and 645 consonants (and again, this classification groups together phonemes which are not exactly identical). While the average is five vowels per language, some have more than 20; while the average is 22 consonants per language, some have only six (e.g. Rotokas, a language of Papua New Guinea) or have 95 (e.g. !Xu). As we have seen above, the ways in which sounds are classified also vary greatly; Chinese, for example, uses tones, that is musical pitch, to differentiate sounds, something which the ears of English speakers have difficulty in catching.

C	**%**	**C**	**%**
t	97.5	g	56.2
m	94.4	ɳ	52.7
n	90.4	?	48.0
k	89.5	tʃ	41.8
j	84.0	ʃ	41.6
p	83.3	F	40.0
w	76.8	dz	34.9
s	73.5	ɲ	31.3
d	64.7	t^{s}	29.3
b	63.8	k^{h}	22.9
h	62.0	p^{h}	22.4
l	56.9	v^{r}	21.1

Figure 2.11: Distribution of consonants in the UPSID database (adapted from Escudier and Schwartz, 2000; Vallée, 1994).

2.5 Origins, growth and form: three fundamental questions

The previous sections have described some of the essential structures of human speech and looked at their role in human languages today. Possession of a speech code is a basic attribute of modern humans. The origin of these structures (which we might also term "traits", or "forms") is a fundamental question that is being addressed by research into the evolution of language, in the same way that the origin of morphological structures such as the hands, or traits such as bipedalism are addressed by the science of biology.

The speech code is very complex, and seeking to explain its origin implies answering several questions corresponding to different perspectives. For a start, we have seen that it is a conventional system – it is linked to norms formed by the interaction of individuals during the course of their lives. A number of works, and in particular approaches that have used algorithmic models based on the notion of language games (Steels, 1997; Kaplan, 2000 ; de Boer, 2001; Oudeyer et Kaplan, 2007; Steels, 2012; Moulin-Frier et al., 2012) have shown how a linguistic norm can come into being and evolve among a population of agents whose environment includes preexisting conventionalized and ritualized protocols of interaction that can structure language learning.

The separate question of how the very first norms were established, at a time when ritualized and structured interactions did not exist and thus no language games were possible, remains largely unexplored. And this question applies particularly to the formation of the first speech codes. We have seen how speech codes provide modern humans with a repertoire of forms for transmitting information within a framework of conventionalized communication. Now, without a speech code (or manual sign code), it is impossible to have ritualized communicative interaction, because this requires a shared repertoire of forms. So there is the question of how a speech code could arise without there already being conventionalized or ritualized interactions, that is to say without presupposing the existence of any linguistic convention shared by a whole community.

Next, we need to be able to explain why humans speech codes are the way they are.

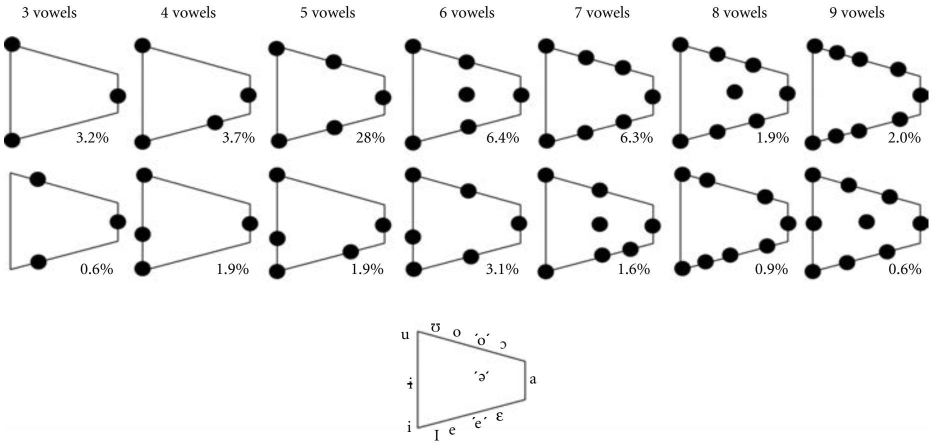


Figure 2.12: Distribution of vowel systems in the languages of the UPSID database. The triangle represents two dimensions characterizing vowels, the first and second formants (formants are the frequencies for which there is a peak in the energy spectrum of the signal, where the harmonics have the greatest amplitude). (adapted from Escudier and Schwartz, 2000; Vallée, 1994).

Looking at such systems from the outside (as a linguist does), how does it come about that speech is discrete, and carves basic building blocks out of the articulatory continuum and re-uses them systematically? How come sounds are grouped into categories? Why are there preferences for places and manners of articulation and for phonemes in the repertoires of the world's languages? Why are there syntactic rules governing the formation of syllables? Why are some rules preferred over others? Why do we have both regularity and diversity at the same time?

Looking at such systems 'from the inside', as a psychologist or neuroscientist does, parallel questions arise: How can speakers acquire a sound system as they develop? Which sensorimotor or cognitive mechanisms will be used? How is the repertoire of gestures learned? How is that that sounds are perceived subjectively, with acoustic illusions?

The question of the origin of the first conventionalized speech code, the question of the general form of this code in contemporary languages, and the question of the ontogenetic acquisition mechanisms for this code are normally dealt with by quite independent research communities, and are not taken on together. This, however, is what we will attempt in this book, in the modest framework of the computational models that will be presented. One of the things we will try to show, moreover, is that taking these problems into account simultaneously can be done in a reasoned way and can yield theories which are not too complex and illuminate them with a new and original light. It is by trying not to isolate the separate questions too much from the start, keeping in mind the systemic and complex nature of speech, that one may show that the complexity of these problems can after all be reduced [2].

The questions which arise concerning the origin and form of the speech code are analogs of the questions which arise in general for biologists on the origin of the shapes, structures and characteristics of living organisms. It is essential to consider these questions within a

[2] Keeping a certain level of complexity in problems at a global level evidently leads to reducing complexity at the local level: Chapter 5 will discuss the advantages and disadvantages of this approach.

general framework of the origin of forms in nature. This allows one to specify what kinds of answers are required and what pitfalls one should avoid. For this reason, before continuing to spell out the problems of the origin of the speech code and to develop precise theories about it, we will take a step back and reflect on the mechanisms responsible for the origin of forms in nature in general. The next chapter will therefore set out the phenomenon of self-organization, characterizing a certain number of form-creating mechanisms particularly responsible for shaping living organisms. Above all, we will attempt to present a reasonably argued relationship between the concept of self-organization and that of natural selection. This will allow us to outline the structure of the arguments necessary for explaining the origin of living forms.

Chapter 3

Self-Organization and Evolution of Life Forms

3.1 Self-organization

In nature, self-organization can be seen in the spontaneous formation of organized forms and patterns whose organizational properties are qualitatively different from those present at the local level. For example, the global symmetry of an ice crystal is not a direct transposition of the form or the properties of the interacting water molecules of which it is composed – there is a qualitative difference. Self-organization is a fundamental concept of modern science[1] that can be observed in a whole variety of natural systems, whose form-creating mechanisms are not necessarily instances of natural selection (for example, the mechanisms involved in the formation of inorganic structures). There are certainly no universal principles which enable us rigourously to unite all self-organizing systems, but nevertheless a number of components do recur with some frequency: breaking of symmetry, tension between forces, positive feedback loops, the presence of attractors in dynamic systems, a flow of energy through dissipative structures, nonlinearities and bifurcations, and noise (see Philip Ball's *The Self-Made Tapestry* (Ball, 2001) for a variety of examples). The next sections give two classic examples of systems which have this property of self-organization: Rayleigh-Bénard convection and the ferro-magnetization . Both examples are taken from the inorganic world, and are therefore illustrations of mechanisms of morphogenesis that do not involve natural selection.

3.1.1 Rayleigh-Bénard convection: Structure formation out of equilibrium

The first example is the formation of so-called Rayleigh-Bénard cells. This phenomenon arises when a thin layer of liquid is placed on the level top of a stove. If the liquid is not heated, then it is in a state of equilibrium in which none of its particles move. The properties of the system are homogeneous; the temperature is the same throughout. The system is symmetrical at the macroscopic level. Now if the stove is gently heated from below, the heat

[1] Contrary to what one sometimes reads in the literature, self-organization is not a mechanism, but a property of a growing system, in the same way than symmetry of size can be a property of a system.

will be transferred from the bottom to the top by a process of thermal conduction. In other words, there is no macroscopic displacement of fluid, but rather an increase in the thermal agitation of the particles which, from one neighbourhood to another, move to the colder surface. Layers of liquid at different temperatures acquire different densities and hence different masses; under the effect of gravity, there is thus a force which pushes the higher (colder) layers toward the bottom and the lower (warmer) layers toward the top. This force is dependent on the difference in temperature between the bottom and the top of the liquid. The force is in competition with the viscosity of the liquid, which itself inhibits movement. This is why, when the temperature difference is small, the liquid itself does not move and only thermal conduction takes place. But if the temperature passes a certain threshold, then the liquid suddenly begins to move at the macroscopic level, giving the appearance of convection currents. What is interesting is that these currents are not random, but organize themselves into quite particular structures which break the symmetry of the liquid. However, the symmetry does not disappear altogether, and its typical dimensions are several orders of magnitude greater than those of the forces applying at the molecular level. At first, just above a "critical" temperature threshold, parallel rectangular stripes are formed (see Figure 3.1). Two adjacent stripes circulate the liquid in opposite directions to each other. The initial symmetry is minimally broken: since it has two even dimensions, the system stays symmetric along the dimension which is parallel to the stripes. If the temperature is raised further, then stripes at right angles to the first stripes appear spontaneously. The liquid is organized into square convection cells. If the temperature is raised still further, then polygonal forms appear, which can sometimes cover the whole surface with regular hexagons. Figures 3.2 and 3.3 show these different states. If the temperature is raised very high, the regular pattern becomes chaotic and turbulent. If the heating is stopped, so as to equalize the temperature at the top and the bottom, the convection patterns disappear and the liquid returns to its equilibrium state, with a uniform temperature and no macroscopic displacement of molecules. It is interesting to remark that these successive, spontaneous steps are the result of a continuous, linear increase in temperature: the dynamic of the system is therefore nonlinear. Here we have an example of a system in which the formation of patterns requires that the system be pushed far from its equilibrium by a continuous flow, in and out, of energy (thermal agitation in this case). Such systems are known as dissipative systems (Prigogine and Nicolis, 1977). However, self-organization is not concerned only with systems pushed out of equilibrium, but can equally take place when a system evolves toward its equilibrium state.

3.1.2 Ferro-magnetization: Structure formation at equilibrium

Another example of a natural system with the self-organizing property is that of iron plates. An iron plate is an assembly of atoms each of which is a sort of microscopic magnet. Each atom can have two possible magnetic orientations, termed -1 or +1. The state of each atom depends on and evolves as a function of two parameters: the states of its neighbours, whose majority orientation it tends to adopt, and temperature, which makes it randomly change state all the more often when it is raised (and which thus has no effect when it is zero). First of all, note that the behaviour of a piece of iron, a macroscopic arrangement of its atoms, is interesting at zero temperature. Whatever the initial state of the atoms, the system self-organizes in such a way that after a certain time, all the atoms are in the same state, which can be +1 or -1, and remain in this state of equilibrium. That is, even if initially each atom is in a random state, a kind of global consensus is formed according to which even two quite

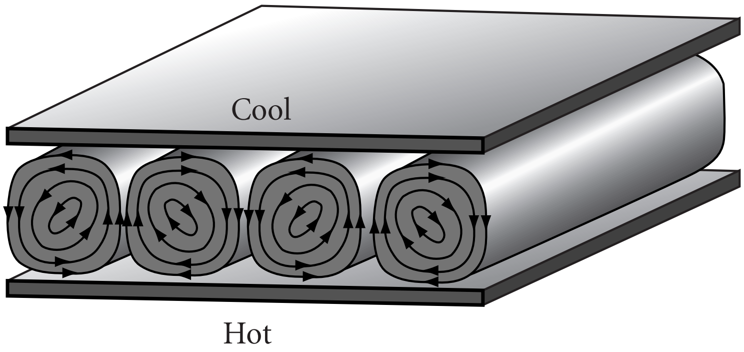


Figure 3.1: If a thin layer of oil is heated on a stove, then given a certain minimum temperature difference between the top and the bottom of the liquid, there is self-organization of convection currents in parallel stripes (adapted from Ball, 2001).

distant atoms end up with the same orientation. The two equilibrium states, "all atoms are +1" or "all atoms are -1" are called the attractors of the dynamical system formed by the set of atoms. If the orienting of each atom is carried out asynchronously and randomly, which is a good approximation to reality, it is not predictable which equilibrium state will be reached: this depends on the particular history of each system. The outcome is most uncertain when the atoms are in a random state, so that there are equally many in the -1 and +1 states. This is a globally symmetric state of the system, as neither orientation is favoured. This initial state is an equilibrium since just as many atoms will switch from +1 to -1 as in the opposite direction, but it is an unstable equilibrium. In fact the random updating of the states of the atoms causes fluctuations which make the ratio of one state to the other vary around 1. Then, for example, the more +1s there are, the greater is the probability that atoms with this state will convert others to the same state as themselves. This can 'snowball': it is what is called a positive feedback loop. At a certain moment, one of the random fluctuations in the ratio of states is amplified by a positive feedback loop. This is how one particular magnetic orientation is "chosen" by all the atoms and magnetizes the piece of iron at a macroscopic level. Symmetry is thus broken.

Now if the temperature exceeds zero and randomly changes the states of the atoms more often as it gets higher, there are three possible situations. First of all, if the temperature is low, then its effect only slows down the convergence of the piece of iron toward a state where all the atoms share the same magnetic orientation. Conversely, if it is very high, it becomes the dominant factor among the forces affecting the local magnetic interactions between atoms. Here, no order appears and the state of each atom evolves randomly over time. The piece of iron is demagnetized. What is more interesting is an intermediate situation, corresponding to a very narrow temperature band: large regions appear in the piece of iron with complex but well-defined forms, composed internally of atoms which are mostly in the same state. It is a state between order and disorder, corresponding to "complexity at the edge of chaos" (Kauffman, 1996). By changing the temperature, one can see phase changes: from a completely ordered state, after a certain critical threshold, a state with complex patterns is reached, after which total disorder soon appears. Figures 3.4 and 3.5 represent

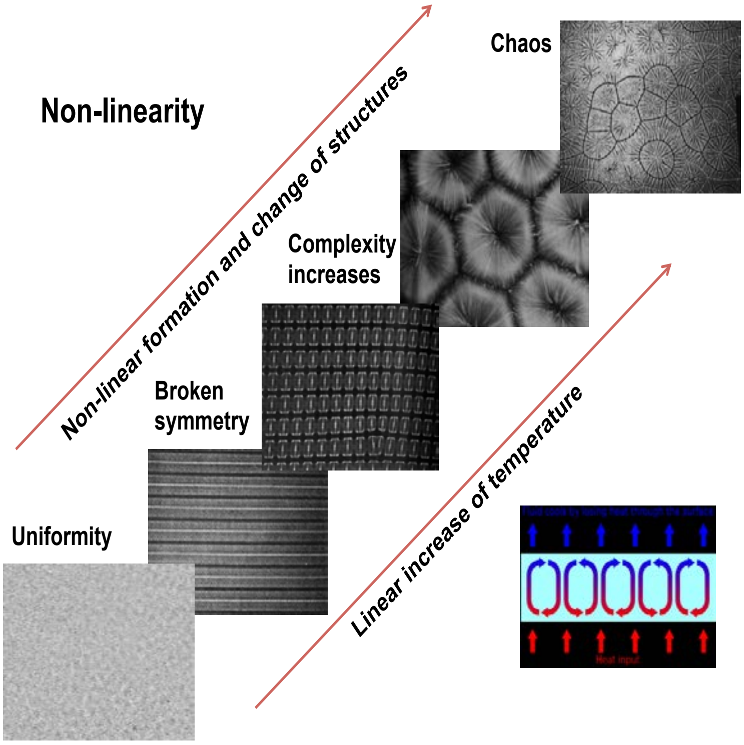

Non-linearity
Non-linear formation and change of structures
Chaos
Complexity increases
Broken symmetry
Uniformity
Linear increase of temperature
Fluid cools by losing heat through the surface
Heat input

Figure 3.2: Representation of convection currents in Bénard liquids when the temperature is raised: at first parallel stripes are formed, then there are square cells, and for higher temperatures polygons appear. If the temperature is raised even further, the regular patterns dissolve into turbulence (adapted from Tritton, 1998; and Velarde).

these phase transitions in a two-dimensional model of ferromagnetic plates.

These two examples of self-organizing systems have some points in common which we will see again in the artificial systems described in the following chapters. They are both characterized initially by symmetry at the macroscopic level which is then disrupted. Nevertheless the final self-organized state is still characterized by other symmetries which make it an "organized" system; it is possible to predict the overall form of this global state qualitatively but not quantitatively because it depends on the history of the system subject to random fluctuations. There is competition between forces pushing the system in different directions. The system has a control parameter whose values determine several types of behaviours or "phases", and continuous linear variation along this parameter is mapped to rapid nonlinear transitions between the different phases.

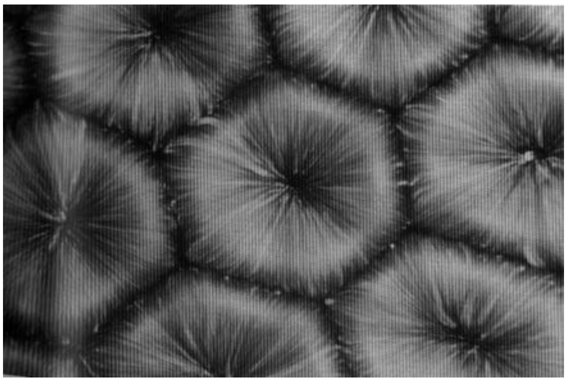


Figure 3.3: At a certain temperature, the Bénard cells tesselate the surface with regular hexagons (Photo: Manuel Velarde, Universidad Complutense, Madrid).

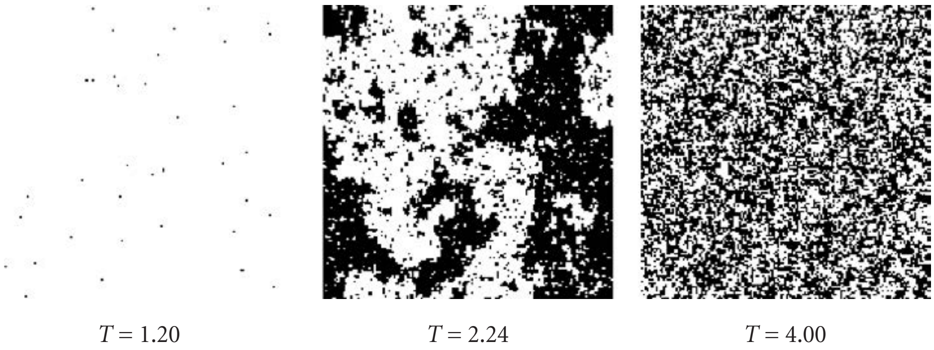


Figure 3.4: Representation of the states of atoms in a two-dimensional ferromagnetic structure model. The points are black or white according to whether the atoms they represent are in state +1 or -1. The left-hand square shows a typical configuration starting from an initial random state with low temperature (the atoms are almost all in the same state); the right-hand square shows a typical configuration when the temperature is raised (the atoms are in random states); the middle square shows a typical configuration in an intermediate temperature band (the atoms form regions with complex shapes within which all are in the same state.

3.2 Self-organization, natural selection and individual development

The examples of the previous section were chosen deliberately from inorganic systems to show that the property of self-organization can be found in systems whose mechanisms are very different from that of natural selection. However, self-organization applies to living systems too. It is a concept widely used in several branches of biology. It is particularly central to theories which explain the capacity of insect societies to build nests or hives, to

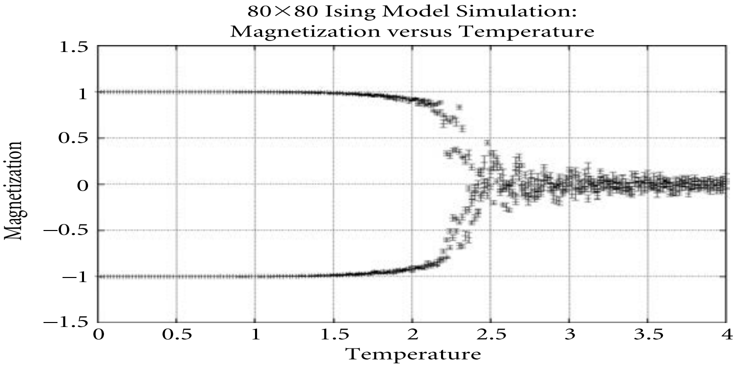


Figure 3.5: Representation of the magnetization of a two-dimensional ferromagnetic model, after settling from a random state, and as a function of temperature. At low temperatures, the metal self-organizes and all atoms adopt the same magnetic orientation. Two orientations are possible, corresponding to two opposing magnetizations, as can be seen in the figure. At high temperatures, the final state also consists of atoms in random states, and yet globally there are not more atoms oriented in one direction than in the other: the iron fragment is not magnetized. Between the magnetized and non-magnetized states, it can be seen that the transition is rapid and nonlinear. This diagram is also a way of showing the phenomenon of bifurcation at a branching point.

hunt in groups or to explore in a decentralized and effective way the food resources of their environment (Camazine et al., 2001). In developmental biology, largely as a consequence of the visionary work in mathematical and computer modelling of Alan Turing in the 1950s (Turing, 1952), it is used, for example, to explain the formation of coloured patterns on the skins of animals like butterflies, zebras, jaguars or ladybirds. Self-organization also characterizes the way shoals of fish move in a synchronized way to avoid predators (Ball, 2001).

It seems possible, then, that there are shape- and pattern-forming mechanisms in biological systems which are orthogonal to natural selection, just because they have the property of self-organization. Now natural selection is at the heart of almost all the arguments in biology when it comes to explaining the presence of a shape, a pattern or a structure in an organism. What, then, is the relationship between the the theory of natural selection and self-organization?

Some researchers have suggested that self-organization casts doubt on the centrality of natural selection in explaining the evolution of living organisms. Waldrop writes (Waldrop, 1990):

> “Complex dynamical systems can sometimes go spontaneously from randomness to order; is this a driving force in evolution? ... Have we missed something about evolution – some key principle that has shaped the development of life in ways quite different from natural selection, genetic drift, and all the other mechanisms biologists have evoked over the years? ... Yes! And the missing element... is spontaneous self-organization: the tendency of complex dynamical systems to fall into an ordered state without any selection pressure whatsoever.”

However, rather than seeing self-organization as a concept which minimizes the role of natural selection by suggesting competing form-creating mechanisms, it is more accurate to see it on the one hand as belonging to a somewhat different level of explanation and above all on the other hand as describing mechanisms which actually increase the power of natural selection by an order of magnitude. Mechanisms with the self-organizing property are completely compatible with the the mechanism of natural selection in explaining the evolution of life forms.

3.2.1 Natural selection and the limits of neo-Darwinism

To see the matter clearly, it is first necessary to recall the mechanism of natural selection that underlies Darwin's theory of evolution. It is a mechanism characterizing a system composed of individuals each having particular traits, shapes or structures. In addition, the individuals in this system are capable of replication. This replication must occasionally produce individuals which are not exact copies of their ancestors, but are slight variants. These variations are the source of diversity among individuals. Finally, each individual has a greater or lesser capacity for replication, according to its surrounding environment. Moreover, most often the environment is such that not all individuals can survive. Thus, differential replication of individuals in an environment, which does not allow everyone to survive, gives rise to "selection" of those who are most capable of replicating themselves. The combination of the processes of variation and selection means that, over the generations, the structures or traits of individuals that best help them to reproduce themselves are preserved and improved upon.

Now there is one crucial point on which the theory of natural selection as formulated by Darwin is neutral: it is the way in which variation is generated, and more generally the ways in which the individuals with their shapes, traits and structures are produced. Neo-Darwinism, which can be seen as a 20th century extension of Darwin's theory, provides an explanation: the biological structures of individuals are specified by their genes, which control how an organism is constructed, and mutations and crossover (in the case of sexual reproduction) are the engines producing variation in the properties or traits of these structures. This explanation would be sufficient if the relationships between genes and traits or shapes of the organism were simple, direct and linear. In this case, exploration of the space of phenotypes (which determine, along with the environment, the relative effectiveness of the genes at replicating) could simply involve studying changes in genetic sequences by mutations and crossover. Now the mechanisms of mutation which bring about these changes within the space of genotypes are rather straightforward and of little amplitude (most mutations only affect a very small part of the genome when replication succeeds). What this means is that under the hypothesis that phenotypic and genotypic space have the same structure and can be mapped approximately linearly, the space of possible biological forms can be searched quasi-continuously, by successive small modifications of pre-existing forms. Fortunately for the appearance of complex life-forms, this is not the case. In fact, although this mechanism of small successive variations in form is notably effective for fine tuning the structures of organisms, it would make the search for forms as complex as those of organisms such as mammals equivalent to the search for a needle in a haystack (Keefe et Szostak, 2001).

3.2.2 Epigenetic self-organization

It is here that the concept of self-organization comes to the rescue of this naive search mechanism in the space of phenotypic forms. In fact the relation between genes and the forms of organisms is complex and strongly nonlinear. Organisms are constructed starting from a stem cell containing a whole genome. This stem cell, together with all its biophysical structures, can be seen as a dynamic system parameterized by its genome, influenced by perturbations imposed by the environment, and which progressively engenders a complete organism as it develops. This processes is known as ontogenetic development, or sometimes epigenesis when full consideration is given to the role of the organism's interactions with its environment (Gottlieb, 1991; Smith and Thelen, 1993). This dynamic system is crucially a self-organizing system with the same type of properties as the self-organizing systems described in the previous section. The genome can be seen as a set of parameters analogous to temperature and the viscosity of liquids in Rayleigh-Bénard systems, and the environment as analogous to noise (but evidently highly structured noise!). Thus the development of an organism from a stem cell has strong similarities to the the self-organized formation of Bénard cells: shapes, structures and patterns appear at the global level, and are qualitatively different from those implementing functioning at the local level, that is, different from the patterns characterizing the structure of the stem cell and its genome. The hexagonal pattern which can appear as a result of a simple difference in temperature in a homogeneous liquid gives an idea of the way in which a simple sequence of nucleotides, enclosed in a system of molecules which transforms them automatically into proteins, can generate a bipedal organism endowed with two eyes and ears and an immensely complex brain.

3.2.3 Attractors, non-determinism and saltationist evolution

As with Rayleigh-Bénard systems or ferromagnetic plates, dynamic systems defined by the cells and their genomes are characterized by a landscape of attractors: there are large regions in the parameter space within which the dynamic system systematically adopts behaviour which is more or less the same. For Bénard systems, there is a range of temperatures giving rise to parallel stripes which is wide enough to locate easily. For ferromagnetic plates the range of temperature in which the system settles to global magnetic coherence is also very wide. For living organisms it is not only possible to generate self-organizing structures with complex global properties, but in addition these structures are generated by genomes belonging to broad sub-spaces of genome space, called basins of attraction (Kauffman, 1993). The structuring of genome space into basins of attraction by this kind of dynamic system facilitates the evolutionary search of the space of forms so that it does not resemble a search for a needle in a haystack. As in ferromagnetic systems, the structured noise imposed by the environment on the development of the dynamic system can lead it to follow developmental pathways that are different, starting out from more or less identical initial conditions and with the same mechanisms. For pieces of iron at low temperatures, this corresponds to magnetization in one direction or another. For a living organism, this corresponds to its possible shapes; this is how it happens that even monozygotic twins can show quite important morphological differences.

This insight, of which Waddington (Waddington, 1946) was one of the precursors, also illustrates why the relationship between genes and the forms of organisms is not only complex and nonlinear, but also non-deterministic. As in Bénard systems where search of the parameter space of temperature can sometimes lead to fast and qualitative changes in the behaviour of the system (for example the change from parallel stripes to square cells), which have been called phase-transitions, the search within genome space can also lead to fast qualitative changes. This can correspond to the numerous observations that have been made of rapid form-changes in evolution, as witnessed by the fossils studied by anthropologists, and which are the basis of the theory of punctuated equilibrium proposed by Eldredge and Gould (Eldredge and Gould, 1972).

The self-organizing properties of the dynamic system composed by the cells and their DNA thus adds crucial structuring to the phenotypic space as the individual develops by constraining this space and so making the discovery and natural selection of complex, robust forms much easier.

On the one hand, these properties enable a genome to generate complex, highly organized forms without the need for precise specification of each detail in the genome (in the same way as Bénard's polygonal shapes are not specified precisely, or encoded in a plan, in the properties of the liquid's molecules). This extends all the way from the formation of regular patches on the skin of the leopard to the formation of sensorimotor and behavioural capabilities such as walking on two legs, as Thelen and Smith's theory of infant development based on dynamic systems (Smith and Thelen, 1993) has so ably demonstrated.

On the other hand, these self-organizing properties structure the landscape of possible forms into basins of attraction within which they resemble each other greatly (here is where gradual evolution happens, involving fine tuning of existing structures), and between which there can be substantial differences among forms (transitions from one basin to another are what provide abrupt and powerful innovations in evolution). To give a simple picture, self-organization provides a catalogue of complex forms distributed over a landscape of valleys in which and between which natural selection moves and makes its choices: self-organization proposes, natural election disposes[2].

3.2.4 The structure of evolutionary explanations

This view of the relationship between natural selection and the concept of self-organization not only allows a unification of the concerns of many researchers who often only work on one of the two aspects, but also shows up the need for scientific explanations of the form of organisms which are more complete than those that are often put forward. In fact, one often sees studies explaining the presence of a shape or trait in an organism in terms of the reproductive advantage that it confers. This reproductive advantage is sometimes transformed into a survival advantage, or even more abstractly into an advantage in realizing some function, but such arguments are just alternative ways of talking about reproductive advantage. For example, bipedalism in humans might be explained by the fact that in the savannah environment walking on two legs helps them to survey their environment better, for predators and food alike, and thus to survive more easily and logically to reproduce more effectively. Another example, much closer to the questions which concern us in this book: following for example Lindblöm (Lindblöm, 1992), we might try to explain vowels systems, made up of elements sufficiently distinct from each other, by the fact that they

[2] Obviously this is only an image to facilitate understanding, because with its movements natural selection actually enables new mechanisms, themselves self-organized, to appear, and these in turn structure the space of forms within which it moves. Thus, natural selection participates in the formation of these mechanisms which help it to move effectively in the space of forms. Vice versa, the mechanism of natural selection certainly appeared in the history of life due to the self-organized behaviour of systems which were as yet completely unconnected to natural selection. Natural selection and self-organizing mechanisms thus help each other reciprocally in a sort of spiral which enables complexity to increase during the course of evolution.

allow information to be passed from one individual to another with minimum risk of mixing up sounds and not understanding, and thus maximizing communicative capacity.

This type of explanation can contain arguments that are correct and essential, while nevertheless remaining insufficiently complete to be satisfactory. These arguments belong, even where this is not acknowledged, to a classical neo-Darwinian vision which makes the simplification set out above of the relation between genotypic and phenotypic space: they say nothing about the way in which natural selection was able to find such a solution. Now according to the view presented in the preceding paragraphs, it is actually this aspect which can be crucial to understanding the origin of a shape or trait. We might imagine a team of Martian researchers landing on earth and asking how humans came to cross the oceans with airplanes. An initial response, which would be that of classical neo-Darwinism, might be: "Because that is the fastest and most effective way of crossing the oceans". This response is accurate, even necessary, but incomplete and unsatisfactory. The neo-Darwinian Martian might add: "and airplanes have structures which were discovered by a cultural transposition of natural selection; For a long time humans tried out many structures, first at random, and then kept the best and made small random variations, replacing a bolt here with a bolt there, which they then selected, and so on, until they hit upon working aircraft.". This would plainly fall a long way short of a satisfactory understanding of the cultural and nonlinear history of the invention of airplanes. Maybe this explanation is somewhat valid as an explanation of how engineers today adjust the exact shapes of wings to increase speed or reduce fuel consumption (by the method of trial and error and hill-climbing with simulation programs), but it is far from revealing the aeronautical revolution between the end of the 19th century and the beginning of the 20th[3].

So it seems that an explanation for the origin of numerous forms in biology requires much more than simply establishing that they enhance the reproduction of the organisms that have them. One needs to identify how the form was generated, and in particular to understand the structure of the relationship between genotypic space and phenotypic space constraining the operation of natural selection. This requires understanding how the structure is formed at the ontogenetic scale of individual development to try to identify the bases from which it could self-organize. In practice, this means looking for structures in the organism itself and in its physical and social environment that are considerably simpler than those that one is trying to explain, and showing how their interactions could give rise to the global structure. The formation of a certain number of macroscopic patterns on the skins of animals, like the stripes of the zebra or the regular round patches of the leopard, can be explained as the almost inevitable attractor outcomes of microscopic molecular interactions between the chemical components present in their epidermis, analogously to hexagonal Rayleigh-Bénard cells, as shown by Turing in the 1950s (Turing, 1952; Ball, 2001). In this way the problem of the origins of these patterns is considerably clarified: evolution did not have to search through all the mathematically imaginable patterns, coding them particle by particle, but only had to find how to produce some chemical molecules whose interaction contrives to make stripes (and it seems that the combinations of chemical elements, just as the numbers of patterns that they can produce on the skins of zebras, for example, is effectively very limited, see Ball, 2001).

Next, it is necessary to understand how variation in the genomes of organisms, and so in the parameters of the organisms seen as dynamic systems, causes changes in form, and

[3] Moreover a parallel is found here on the cultural level between gradual and abrupt changes in shapes and structures during the course of evolution, due precisely to the nonlinearity of self-organizing phenomena which work in tandem with natural (or cultural) selection.

in particular at what speed and of what type. D'Arcy Thompson was one of the pioneers of this work, fundamental to understanding the origin of the forms of living organisms. His seminal book, *On Growth and Form* (1917), is probably one of the most important in the history of biology, standing beside those of Darwin. In particular, he studied the impact of parameters in the growth of structures like the shells of molluscs. These are built by self-organizing processes of cellular division, whose parameters are speed and orientation. D'Arcy Thompson showed how with a constant mechanism, the simple numerical variation of these parameters, which from a modern point of view we can easily imagine to be fairly directly controlled by the genes, can lead to surprising nonlinear and very diverse variations in the generated forms. Figure 3.6 gives some examples of these kinds of parameterized variations of forms in molluscs. He repeated this work for many animal species, and in particular with fishes, where he shows the same phenomenon: figure 3.7 shows the variety of forms which can be obtained with a constant growth mechanism simply by varying the parameters of speed and orientation. These studies considerably clarify the way in which the space of phenotypic forms can be searched and constrained, thus facilitating the natural selection of efficient, complex forms.

D'Arcy Thompson has another example which illustrates very well how self-organizational phenomena can facilitate the discovery and natural selection of efficient, complex structures by natural selection[4]. This is the case of the hexagonal cells formed on the walls of bee-hives[5]. This form of cell is remarkable because the hexagonal shape is optimal: hexagons use less wax than any other possible shape to cover the same surface area. There are two ways of accounting for this form. The first takes a classical neo-Darwinian point of view and considers no mechanism other than classical natural selection. The bees would have tried out a whole range of possible shapes, starting with random shapes and selecting those whose construction used up less energy, varying them little by little, and selecting again, and so on, until finally coming across the hexagonal shape. This amounts really to a search for a needle in a haystack, taking into account on the one hand the huge number of possible cell-shapes, and on the other that this view presupposes identity between the genotypic space and the space of cell-shapes; in other words the search is unconstrained. Fortunately for the bees, their search is helped by a providential phenomenon of self-organization. D'Arcy Thompson noted that if the cells are taken to be of roughly the same size and of a relatively smooth shape, and if the temperature generated by the bees makes the wax walls flexible enough, then cells packed close next to each other will behave more or less like drops of water in the same situation if surrounded by a viscous fluid. Now, the laws of physics work so that each water-drop in such an assembly of drops will take on a hexagonal shape, as shown in Figure 3.8. So the bees don't need to work out how to design a regular hexagonal tesselated pattern, which would require abilities worthy of a young mathematician armed with compasses and rulers, but, much more simply, they have to work out how to make cells which are roughly the same size and not too twisted, packed up close to each other. Physics does the rest. And thus, in explaining the origin of hexagonal shapes of wax cells

[4] Of course, in his time D'Arcy Thompson did not describe this example in the same way as we do, because the concepts of self-organization and dynamic systems had not yet been invented. Nevertheless, a reading of "On Growth and Form", which describes many mechanisms that we could now call "self-organizing", without having been conceived as such at the time of writing, suggests that his intuition was moving in the direction in which we interpret his work here.

[5] I take the position in this book that structures built in one way or another by organisms, like the wax walls made by bees, are themselves aspects of the characteristic form of the organisms. This point of view corresponds to the idea put forward by Dawkins in "The Extended Phenotype". In the same way, the speech code will be considered as belonging to the characteristic form of humans.

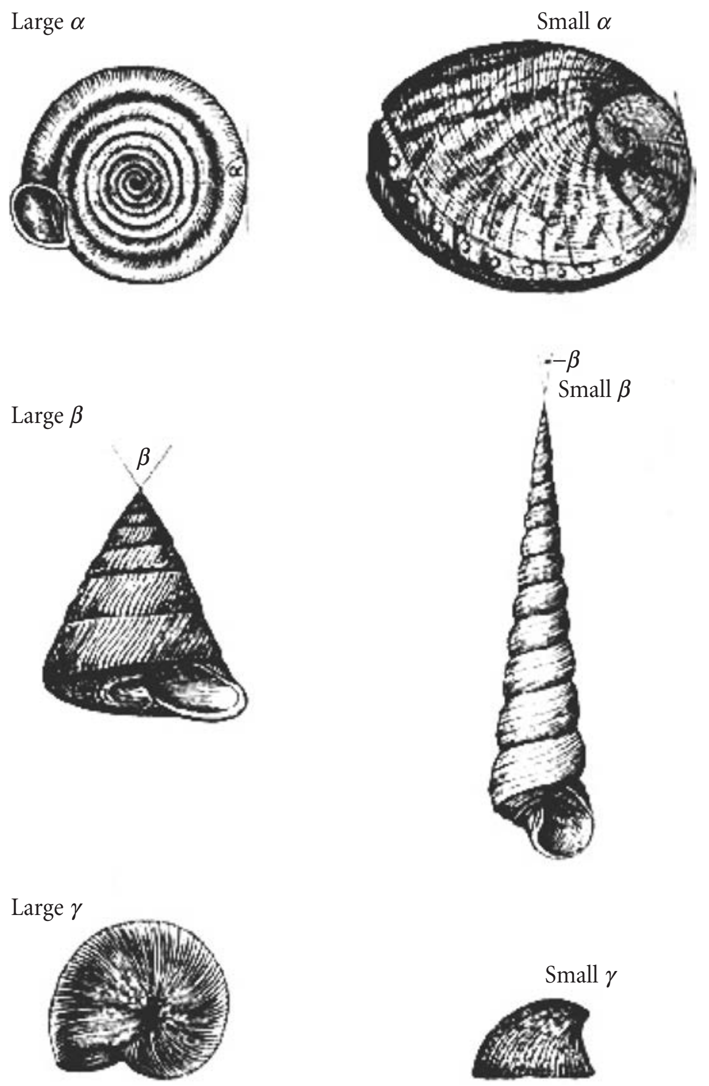


Figure 3.6: D'Arcy Thompson showed that with a constant growth mechanism, simple numerical changes in the parameters of cell division, such as speed and orientation, can result in the generation of very diverse forms in molluscs, corresponding to different species.

in beehives, the role of self-organization of physical structure is just as important as the metabolic advantage that this structure confers on the bees[6].

[6] This does not mean that nowadays honeybees do not have a precise innate hard-wired neural structure which allows them to build precisely hexagonal shapes, as has been suggested in further studies such as those of von Frisch (von Frisch, 1974). The argument of D'Arcy Thompson nevertheless shows how the self-

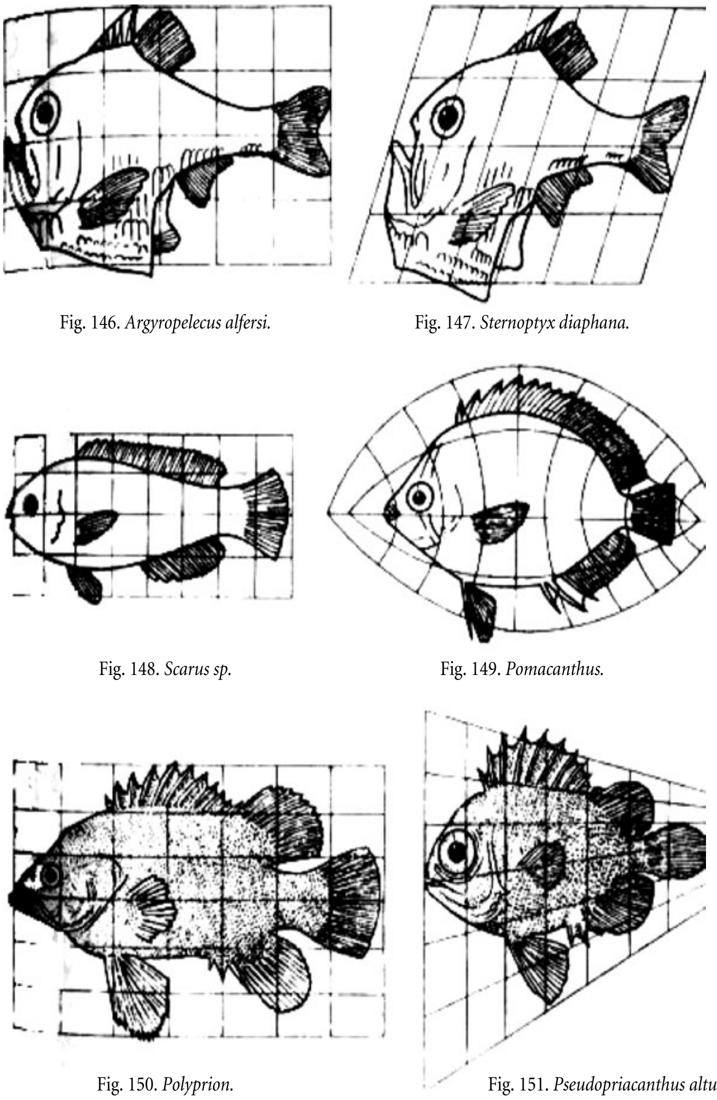


Fig. 146. *Argyropelecus alfersi.*

Fig. 147. *Sternoptyx diaphana.*

Fig. 148. *Scarus sp.*

Fig. 149. *Pomacanthus.*

Fig. 150. *Polyprion.*

Fig. 151. *Pseudopriacanthus altus.*

Figure 3.7: D'Arcy Thompson extended his work on molluscs to all kinds of living species, such as fish for example.

organization of heated packed wax cells could initially have helped the bees to 'find' these optimal hexagonal forms. Later on, evolutionary mechanisms such as the Baldwin effect (Baldwin, 1896) might have resulted in schemata being incorporated in the genome for building the hexagons directly.

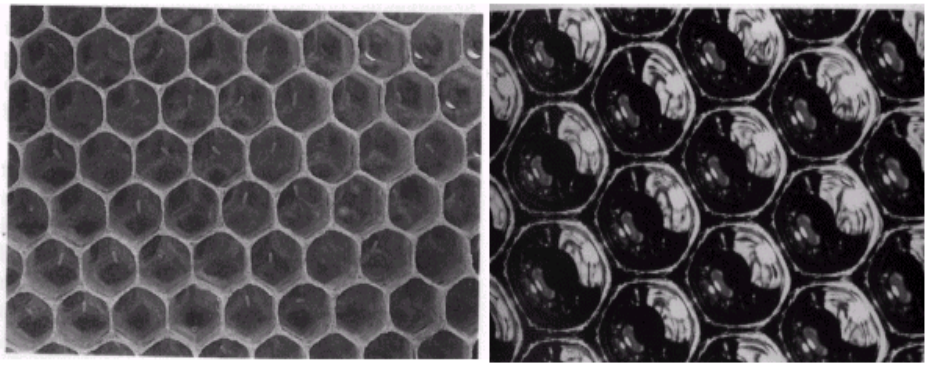


Figure 3.8: The figure on the left shows the regular hexagonal tesselation in the walls built by bees in their hives. The right-hand figure shows the shape taken by drops of water when they are packed together: it is exactly the same shape as seen in the walls of beehives. (Photos: Scott Camazine, Pennsylvania State University (left); B.R. Miller (right)).

3.2.5 Collateral effects

We have just seen that sometimes the explanation of a shape or a trait in an organism requires much more than establishing its usefulness in promoting successful reproduction. Indeed, it can even happen that such usefulness is not even directly involved in an explanation of the origin of a form. This is again possible because of the existence of self-organization. The interaction of several structures appearing in an organism for Darwinian reasons (each structure helping to replicate its genes more effectively), can lead to the self-organizing formation of a new structure which may have no usefulness at all for the organism. One of the most striking examples is that of zebra-like or leopard-like striped or spotted patterns in certain species of molluscs (Ball, 2001, pp. 89, and see Figure 3.9). Although one can easily imagine a function for these patterns in zebras and leopards (camouflage), it is hard to imagine one for these molluscs because they live buried under the sand in the depths of the oceans where no light penetrates, and where their visual patterns are not perceived by any living creatures. These patterns are formed by growth processes in the shells. They correspond to the pigmentation produced during the continuous gradual calcification of the cells at the edges as the shell grows by a process of cell division with a dynamic very similar to that of Turing's model (Turing, 1952) and to the well-known Beloutzov-Zabotinsky reaction, a classic example of a self-organization phenomenon (Ball, 2001).

3.2.6 Architectural constraints

In fact, we need not appeal to self-organization to give evidence of forms whose origin is not directly linked to the advantage they provide for the reproduction of the organism. Some structures can be side effects of the formation of other structures which themselves are useful in the reproduction of the organism. These side effects arise from architectural constraints, a topic developed in depth by Stephen Jay Gould (Gould, 2006). We will use a technological example to illustrate the concepts of side effect and architectural constraint. The operation of electric light bulbs, whose purpose is to provide light, has traditionally involved passing a

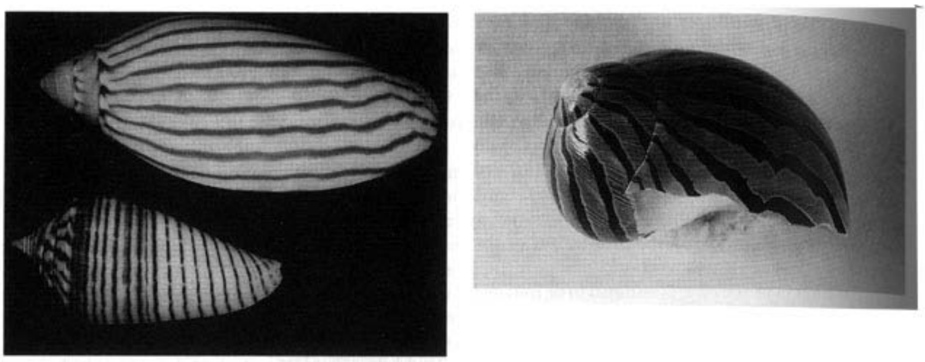


Figure 3.9: Certain species of mollusc living in darkness at the bottom of the ocean have striped patterns on their shells. These patterns result from the pigmentation produced during the continuous gradual calcification of cells at the edges of shells as these grow by a process of cell division with a dynamic very similar to that of the well-known Beloutzov-Zabotinsky reaction. These self-organized forms have no adaptive value for the molluscs (Photo: Hans MeinHardt).

current through a metal wire. Now this has a systematic consequence which can sometimes be inconvenient for users of the light bulb: heat is emitted. The generation of heat was consequently for many years one of the characteristic traits of electric light bulbs. It is the same with oil lamps. LED-based technology has recently allowed this production of heat to be reduced, but not entirely removed. Human engineering cannot make lamps for everyday use that do not also produce heat. Likewise, it is highly probable that evolutionary engineering is sometimes obliged to tolerate the cost of inconvenient side effects for the benefit of other useful structures from which they result. Gould gives the example of the tibial sesamoid bone (part of the tibia) of the panda which tends to hinder its walking, but results from a growth regime which allows the radial sesamoid (its equivalent on the forelimb) to serve as an opposing thumb for handling bamboo shoots (Gould, 1982). In fact the coupling of the growth of these two bones has brought it about that the adaptation of the radial bone on the forelimb for grasping has also produced a morphological change in the tibia. The first change is positive for the panda, while the second is an inconvenient side effect (but which is worth the cost because this arrangement has been selected).

3.2.7 Exaptation

It is clear that any explanation of such "collateral" structures also calls for a precise explanation of why they are the consequences of the formation of other structures which must be identified, and whose origin must in turn be explained. These examples of collateral effects are mostly useful for avoiding getting trapped when trying to explain certain characteristics of organisms. In fact, for the sake of simplification, science will frequently isolate particular traits which are highly characteristic of an organism and then explain them afterwards. Herein lies a risk: that of looking for a utilitarian reason for the existence a trait where there is none, other than in relation to the organism in its entirety. This structure or trait that one is seeking to explain may be a collateral effect of another structure that one has not considered. It is especially important to be aware of this risk where the structure concerned has a useful function for the organism in the eyes of the scientist studying it, but where the function is in fact more recent than the structure. This is the phenomenon of exaptation (Gould and Vrba, 1982): some property N was taken from an earlier state (hence "ex") to be used (hence "apt") in a new role. To illustrate the situation, we will use the example of very long suspension bridges. Imagine that Martian scientists come across one of these. They notice that the pillars of the bridge are extremely high. Why so high? They also see that each pillar carries a collection of antennae, serving as radio relays for telecommunications. They conclude that this must be the reason for the height of the pillars: they enable the relaying of radio waves over long distances. In this they are of course completely mistaken: the pillars were built without anyone knowing that they would be used to house radio antennae. They are very high simply because it was desired to minimize the number of pillars along the bridge. Besides, in order for each pillar to be able to support the immense weight of the long roadways, the cables supporting this horizontal part need great strength as well as great resistance. And the closer they are to the vertical, the more effectively they support the roadway. And the higher the pillars, the closer to the vertical are the cables. This is a pure example of an architectural constraint. So the Martian scientist's error comes from his isolating the pillar from the rest of the structure, and from the pillars having, by chance, found a supplementary use after their construction. In biology, Gould and Vrba give the example of birds which initially evolved feathers for regulating their body temperature, and recruited them later for flight (Gould and Vrba, 1982). They also give the example of certain species of snail which have a space inside their shell in which they "hatch" their eggs. There are also species of snail which have this space but do not use it, and these latter species appeared in evolution before those which use the space. This space is in fact a result of the process of construction of the shell complying with architectural constraints similar to those which require the pillars of suspension bridges to be very high. This space in the snails' shells is an architectural side effect with no particular function initially, that was only later on recruited for use as a shelter for eggs.

3.3 Towards a systemic understanding of the origins of speech

It has been shown in the previous sections that the genesis of a form is a complex process, which could involve several causal factors. This means that in seeking an explanation for the origin of a living form it is necessary to provide several types of answers, representing complementary visions of the same phenomenon from different points of view.

A first type of answer concerns the utility of a form, structure or trait in terms of the reproductive effectiveness of the organisms which possess it. This is classical neo-Darwinian argumentation, and is often fundamental to an evolutionary explanation. It sometimes takes a short cut by replacing reproductive effectiveness with survival value, or often even with usefulness for some given function. In addition to the dangers of this type of explanation set out above, this kind of short cut, which can be useful, can also sometimes be dangerous, because, as (Dawkins, 1982) has shown, what counts in the mechanism of natural selection is reproductive or replicative effectiveness: it does not always follow that a trait enabling individual organisms to survive better enables them to to reproduce better (and moreover some organisms "commit suicide" in order to perpetuate their genes, like certain species of insect which die to provide food for their offspring, see Dawkins, 1982). Another example concerns communication, and more particularly the sharing of information, which in certain ecological conditions can adversely affect the reproductive capacity of an organism as has been shown by Gintis and his colleagues (Gintis et al., 2001), and has been studied by Dessalles (Dessalles, 2007).

A purely neo-Darwinian argumentation will often conclude its explanation by proposing that the optimal form was formed by the action of the optimization mechanism constituted by natural selection, without providing any more detail regarding this process of formation. However, we have seen that natural selection is not a complete mechanism in the sense that it does not specify the way in which individuals are formed or, above all, the way in which variation arises. Also, if this form of explanation is filled out with a "naive" version of the search of phenotypic space of the kind that neo-Darwinism often puts forward, then it is of limited power and many of the evolutionary problems it is meant to address start to resemble the search for a needle in a haystack. This is why a second type of answer is needed to back up the original answer in terms of utility and optimization: this second answer should explain how natural selection was able to find the solution, and in particular how it was able to be guided by the self-organization of the systems on which it operates, and by architectural constraints on the structures which it builds.

In practice, this second kind of answer can consist in identifying simpler structures than those which one is trying to explain, corresponding to a phenotypic space which is easier for natural selection to explore, and whose self-organizing dynamic generates spontaneously, during the development of the individual, the global structure in that we are seeking to explain. For example, if one tries to explain the hexagonal shape of cells in beehives, the first type of answer consists in explaining that hexagons are the shapes that require least energy on the part of the bees (optimality), while the second type of answer consists in explaining that these hexagonal shapes appear spontaneously once the bees pack together cells which are not too twisted in shape and of roughly the same size, and warm them. Thus the space of shapes explored by the bees is significantly simplified, and the chance of happening upon a genome leading to the construction of such cells is much greater.

It is exactly this second type of argumentation that provides the drive for this book. We will show that speech codes and their complex properties as described in the previous chapter can be generated by self-organization starting from developmental and interactional structures that are much simpler. In particular, these structures that are simpler, but whose dynamic and systemic interactions are crucial, will therefore constitute the assumptions behind the models presented here and their role will be twofold. First, it will be evident that their complexity, of a completely different order from that of the speech code, makes their discovery and natural selection much more understandable than if one starts from a classical explanation that ascribes the origin of speech solely to the advantage that speech conveys to the function of linguistic communication. Secondly, the generic character of these structures will enable us to explain in Chapter 9 how they could have appeared while having no causal relation with the use of language: in particular, we will propose that they might be collateral effects of other mechanisms whose function is not directly linked to language, such as imitation or mechanisms for spontaneous exploration of the body through curiosity and intrinsic motivations. In short, this will enable me to suggest that the speech codes that we use nowadays could be exaptations, whose first versions could have been the outcome of the self-organization of structures whose origin and functions were distinct from those of language.

Before presenting this line of reasoning and the computer-based experimentation that underlies it, in the next chapter we will provide an overview of the different theories proposed in the literature in response to the questions about the origin of human speech that were posed in Chapter 2. We will interpret them within the general theoretical context of the origin of life forms that this chapter has just presented.

Chapter 4

Theories of how human speech originated

Some of the questions described in Chapter 2 regarding the origins of human speech are at the heart of much scientific research. Different approaches have been proposed, coming from various scientific cultures, and each with its own methodology. This chapter presents an overview of the most representative theoretical and experimental approaches. We will also see that a number of fundamental questions remain largely unexplored.

4.1 Reductionism and nativism

One major approach to this question by the scientific community can be qualified as "reductionist". It tries to reduce the properties of speech to some of its parts. This approach consists in trying to find physiological or neural structures whose characteristics suffice to deduce from them the characteristics of speech.

"Cognitive nativism", which has been argued for by, among others, Pinker and Bloom (Pinker and Bloom, 1990) holds that the brain has a specific neural disposition for language, and in particular for speech (the "Language Acquisition Device"), which knows at birth the properties of systems of speech sounds. This knowledge, according to cognitive nativists, must be pre-programmed in the genome. A limitation of this approach is that its proponents have remained rather imprecise on what exactly it means for a brain or an individual to know innately the properties of speech. In other words, this is a hypothesis that has not been naturalized, which is to say that the link to biological matter, the issue of its implementation, has not been addressed in any meaningful way.

Other researchers defend an approach that could be called morpho-perceptual nativism. They focus their attention on the physics of the vocal tract and on the electromechanical properties of the cochlea. This approach holds that the sound categories appearing in human languages reflect nonlinearities of the system that maps sounds and percepts to articulatory trajectories. Two theories propose different ways of exploiting these nonlinearities.

First, there is the quantal theory of speech, proposed by Stevens (Stevens, 1972). Stevens observes that there are certain articulatory configurations for which small changes produce small acoustic changes, and other articulatory configurations for which small changes produce large acoustic changes. The phonemes used by languages are, then, located in zones of stability, and unstable zones are avoided.

Then there is the "Distinctive Region Model" developed by Carré and Mrayati. This approach, which uses arguments from information theory (Shannon, 1948), proposes on the contrary that speech prefers to use zones in articulatory space for which small changes produce large acoustic modifications. Both Stevens and Carré carried out simulations with models of the vocal tract and made rather good predictions on possible places of articulation, even though their theories are based on assumptions that to some extent contradict each other.

Even though it is quite clear that the properties of the articulatory and auditory apparatus influence the form of speech sounds, any approach based purely on morphology and perception will have its limitations. For a start, strong and obvious nonlinearities are not found in all regions of articulatory space, especially where the production of vowels is concerned. Moreover, as we saw in Chapter 2, a certain number of perceptual nonlinearities are completely cultural and not perceptible to speakers of other languages: Japanese speakers cannot hear the difference between the "l" of "lead" and the "r" of "read" in English. Therefore such nonlinearities do not explain the great diversity of sounds across human languages (Maddieson, 1984), and are of no help in understanding how a given language "chooses" its phonemes.

The theories of Stevens, Carré and their colleagues attack the question of why there is such-and-such a phoneme rather than some other, but do not get to grips with the more basic questions of why there are phonemes at all and how a system of phonemes can become the norm within a population. In short, they do not deal with the fundamental problem of phonemic coding (discreteness and combinatoriality). Among reductionist approaches, that of Studdert-Kennedy and Goldstein (Studdert-Kennedy and Goldstein, 2002) addresses one aspect of this question. This is the organization of utterances into parallel independent gestural tracks, making possible the delivery of information fast enough for humans to convey complex messages effectively (Studdert-Kennedy, 2005). They note that the vocal tract is composed of independent articulatory organs, such as the tongue, the lips and the velum. This implies on the one hand that there is a discrete aspect to the physiology of speech, and on the other hand, since there is only a small number of such organs, that there is a systematic re-use in complex utterances, at least from the point of view of which organs move. Studdert-Kennedy is undoubtedly right on this point. However, other aspects of phonemic coding, and in particular discreteness, remain to be explained. In fact, as they have noted elsewhere (Studdert-Kennedy and Goldstein, 2002), each organ or set of organs can be used to make a constriction in the continuous space of places and manners of articulation. How is this space discretized? Goldstein has proposed an answer to this question that is not reductionist, but a mixture of self-organization and functionalism (Goldstein, 2003), and that will be examined later in this chapter. In the following chapters we will also propose a solution to this second aspect of phonemic coding, although it will not be covering what might be called the "organization of vocalizations into independent gestural tracks".

Generally speaking, these reductionist approaches study the universal properties of speech, and in particular phonemic coding and the common features in phoneme inventories in the languages of the world. They do not deal with the diversity of speech sounds, nor the cultural formation of specific systems shared by language communities. A fortiori, they propose no solution to the problem of the formation of the first conventional codes at a time when they did not already exist. Such was not, however, their goal.

Besides, more generally, this kind of research does not really attempt to explain the origin of the speech code, but rather attempts to discover some it its physiological, morphological and neural correlates. Consequently, it does not really address the problem of causal explanation, but rather approaches language from the perspective of a natural historian by grounding the speech code in its biological substance. Nevertheless, when it is reviewed in the theoretical framework presented in the previous chapter, this kind of grounding enables us to clarify the problem, precisely by “naturalizing” it. Studdert-Kennedy, for example, points out an essential biological aspect of the combinatoriality of the contemporary speech code, namely the independent control of the different organs which can modify the shape of the vocal tract. This allows him to formulate a hypothesis according to which this independent control is a result of co-opting the control structures of the facial muscles, in the functional context of imitation (Studdert-Kennedy, 1998), involving a transfer of information from the visual to the auditory modality. We will not pursue this theory since it concerns an aspect of phonemic coding which will not be dealt with in this study, namely the organization of utterances into independent gestural tracks.[1]

4.2 Functionalism: the speech code is optimized for communication

The functionalist approach tries to explain the properties of speech sounds by relating them to their function. It uses the cover-all notion of function, which is why it is here referred as “functionalist”, although it could also be qualified as a classical neo-Darwinian approach.

The function of the speech code which is typically invoked to explain its form (and implicitly its formation according to a naive version of natural selection) is “communication”. The speech code provides a repertoire of forms which should be as efficient as possible, in order for the individuals who use it to understand each other. This efficiency is evaluated by many criteria. The criterion which is always appealed to is perceptual distinctiveness. That is to say that sounds should be distinct enough from each other for them not to be confused and for communication to take place. The other criteria are often tied to the costs of production, like the energy needed to articulate sounds, or perceptual salience. A repertoire of sounds is thus a set of forms which is quasi-optimal for communication and simultaneously minimizes the costs in terms of energy. This approach differs from the reductionism of the previous section because, instead of looking at phonemes individually, it considers the system they belong to, studied as a whole according to its structural properties. Also, it is truly in the business of explaining the origin of the modern speech code, giving answers in terms of “optimality”.

Some researchers have produced precise models of this idea and have explored the consequences with the aid of computer simulations. Lindblöm and Liljencrants were pioneers (Liljencrants and Lindblöm, 1972; Lindblöm, 1992). Their model concerns the prediction of vowel systems in human languages. Given a number n of vowels, they define the energy in a system of n vowels by $E_n = \sum_{i=0}^{n} \sum_{j=0,j\neq i}^{n} 1/r_{i,j}^2$ where $r_{i,j}$ is the perceptual distance between two vowels. Each vowel in this simulation is a point in a space defined by the first two formants[2]. The possible points are articulatorily confined to a triangle, the vowel triangle (see Chapter 7). This energy is used to measure the perceptual distinctiveness of the whole system. If the vowels are very similar to each other, then the $r_{i,j}$ are small and

[1] As far as phonemic coding in general is concerned, Studdert-Kennedy’s theory is compatible with and complementary to the theory that we will put forward in this study; that is, the one theory does not rely on the other, and one fills out the other in explaining the origin of phonemic coding.

[2] Formants are the frequencies where there is a peak in the energy spectrum.

the energy is high. If they are distant from each other, the energy is low. Using numerical optimization techniques enabled to see which were the vowel systems which had least energy, that is, were minimal. This lead a certain number of resemblances with the most frequent vowel systems of human languages, as far as the systems with fewer than six vowels were concerned. The predictions of the model were then improved by adding a term modelling the articulatory cost.

However, Lindblöm's results were not very life-like for systems with more than six vowels, and gave a large number of high peripheral vowels between [i] and [u], compared to the languages of the world. A second model, incorporating a new criterion, perceptual salience, was then developed within the framework of the Dispersion-Focalization Theory of Schwartz, Boë, Vallée and Abry (Schwartz et al., 1997). This perceptual salience characterizes the relative nearness of the formants of vowels: the nearer they are, the more the energy in one region of the spectrum is reinforced, thus giving the vowel a focal quality. The authors of this theory propose that this property is registered by the brain. This makes it possible to improve on Lindblöm's predictions.

This research method was taken up by Redford, Chen and Miikkulainen (Redford et al., 2001) working on syllable structure. Given a repertoire of phonemes, they looked to see which were the syllable systems, sequences of these phonemes, which represented the best compromise between such criteria as minimization of word-length in the lexicon, minimization of number of syllable types in the repertoire, perceptual distinctiveness between adjacent phonemes, and maximization of the difference in jaw opening between adjacent phonemes. They were then able to predict some phonotactic regularities in the world's languages, like the preference for syllables of type CV (Consonant-Vowel) over syllables with consonant clusters, or the iterative principle of syllable construction (simple syllables are more frequent than complex syllables in the same repertoire, in an organized distribution).

These works, which tried to explain certain properties of the speech code in terms of their quasi-optimality for communication, have led to significant scientific progress. However, they should not lead us to forget that there are traps such as those described in the previous chapter. For a start, the definition of such optimality is far from clear, as is shown by the different criteria used by different authors: evaluation of perceptual distinctiveness depends on ones model of speech perception; some works take energy costs into account, but our knowledge of the speech production system is scant, so that these costs are inevitably modelled very crudely; other works add psychological constraints, such as a preference for salient sounds, that is a preference for sounds whose formants are sufficiently separated from each other. Not only is our knowledge of these different criteria very rough, but it is also hard to know how to weight them in relation to each other. For this reason, while it is interesting to show that one can devise definitions of criteria and combinations of criteria that yield qualitative predictions about the form of the speech code, the quantitative predictions must be treated with greater caution. Moreover, we have seen in the previous chapter that forms could sometimes be explained without necessary recourse to their utility in a given function. It will be shown in the following chapters that this can be the case for certain properties of phoneme inventories: the models we will discuss enable us to predict the same properties of vowel systems as those predicted by Lindblöm, even though no measure of cost or optimality for communication will be used to guide the evolution of the system.

4.3 Synthetic and systemic approach: computer and robotic models

As argued in the previous chapter, any functionalist explanation of the type given by Lindblöm or by Schwartz and his colleagues should be accompanied by an explanation of the mechanisms of morphogenesis of the optimal solution for communication, embodied in the speech code. In fact, while Lindblöm's models show that the most frequent phoneme systems of human languages can be predicted on the basis of optimizing a certain number of criteria, they do not explain how this optimization could have taken place in practice in the biological, cognitive and social substrate. Furthermore, as with reductionist theories, he does not undertake to explain how a community of speakers can come to share a particular sound system.

4.3.1 Models of the formation of languages

Accordingly, a new approach, based on the use of computer models and robotics, started to be explored around the middle of the 1990s, to look at the question of the morphogenesis of languages and of language in general. Luc Steels, one the pioneers of this new approach, talks about 'synthetic methodology" (Steels, 1997; Steels, 2001). This consists, in practice, of creating computer models of groups of artificial individuals, sometimes in the form of robots, equipped with perceptual, motor and cognitive models that enable them to learn during the course of their interactions with one another.

For example, Luc Steels and his colleagues have developed the "naming game" model, in which a group of artificial individuals progressively negotiate a lexicon, that is to say the choice of associations between words and semantic representations (Steels, 1997;2012; 2016). At the outset each individual has a repertoire of semantic representations, but no words to designate them. During the course of interactions they create associations between words and meanings, which are marked with a score enabling the force of these associations to be evaluated and those that are too weak to be eliminated. Agents interact in pairs in a random order. During an interaction, one of the two chooses a referent within the environment (corresponding to a semantic representation), and pronounces the word that has the strongest association with it in its memory (if no associated word is present, an arbitrary symbol is generated). The other individual observes this association, increases the corresponding score in its memory and decreases the score of all other associations containing either this word or this meaning. A competition between word-meaning associations thus takes place among the population of individuals, with a dynamic that we might call cultural Darwinism (Kaplan, 2001; Oudeyer and Kaplan, 2007). In this competition the mechanism that bolsters the scores of the most frequently used associations and eliminates the weaker ones gives rise to a positive feedback loop: in a systematic fashion the repertoires of word-meaning associations self-organize, and the group converges towards a repertoire that is shared by all individuals and that allows them to communicate without ambiguity. A conventionalized lexicon has thus appeared spontaneously during the course of the pairwise interactions of individuals, without any norm having been imposed or pre-programmed. Different groups of individuals will moreover converge towards different conventions, which suggests an explanation for the diversity to be found among languages.

The "naming game" computer model has thus had a major role in showing how simple, local mechanisms of linguistic interactions could allow the formation of linguistic conventions at the global level of a society of individuals. Prior to the development of this kind of model, no explanation existed for what kind of mechanism could drive the formation of language. Such models subsequently gave rise to detailed mathematical analysis, based in particular on the use of theoretical tools in statistical physics (Baronchelli, Loreto and Steels, 2008). At the same time, many other computer models came along to extend the naming game for modeling a variety of more complex phenomena in the evolution of languages.

Some models have focused on the origin of semantic categories linked to the words of a given language. Systems of semantic categories of human languages, that is to say ontologies, can differ widely from one culture to another, and are therefore to some extent cultural structures (Hagège, 2006). For example, the conceptualization of colours, of time and of space can vary according the linguistic environment (Levinson, 2003). Some languages make no distinction between blue and green, while others such as Guugu Yimithirr (Gumperz and Levinson, 1996) do not encode spatial concepts that are relative ("in front of", "behind", "opposite") and only conceptualize the position of objects in relation to an absolute framework ("north of", "east of", etc).

How then can a group of individuals agree on the use of one ontology rather than another? How can a system of categories be built collectively by these individuals? The "guessing game" model shows how this is possible (Steels, 2003; Steels, 2012). In this model, robots that have the ability to perceive their environment visually are endowed with mechanisms for creating concepts in the form of mathematical equations that are then used for identifying referents in their context and that can be associated with words as in the "naming game". At the start, individuals possess neither words nor concepts. Repertoires grow out of interactions with other individuals and from the mechanisms for creating new concepts when they are required. For example, if they do not currently possess a concept enabling them to identify the referent to which their attention is drawn during an interaction, and within a given context, then they will generate a new mathematical equation that can do so. Now, each time this happens there will be a large number of conceptual representations (equations) that would be suitable, and so one is chosen at random. A competition resembling that of the naming game thus takes place between concepts, words and associations. Those that are used the most frequently in interactions where communication is successful are strengthened, while others are eliminated. After a certain time a complete lexical system, shared by all the individuals in a group, is formed, consisting of the same words, the same concepts and the same associations. This system is different for each group. For example, in the "Talking Heads" experiment, one of the most comprehensive scientific experiments ever performed to model language evolution, a population of robots was able to evolve over a period of several months, corresponding to tens of millions of interactions between dozens of individuals, and converging towards a lexical system of several hundred words and concepts. Interestingly, the concepts that were formed were adapted to the properties of the robots' perceptual environment: a world of simple shapes and different colours laid out on a white background (see Figure 4.1). In this way, rather "sophisticated" concepts relating to colours (e.g. "light red", "dark red") and to relative positions (e.g. "below", "to the left") were formed, while concepts relating to shapes, which did not vary greatly, remained few in number. The Talking Heads shows how the physical properties of the external world, in interaction with cognitive bias, can influence the semantic categories that are formed, while their diversity tends to be the result of historical contingencies and of arbitrariness in the process of social interaction.

Using a similar approach and similar methods, other models have been developed to study how syntactic structures are formed and become more complex (Kirby and Hurford, 2002; Christiansen and Chater, 2016), how grammatical structures and categories evolve

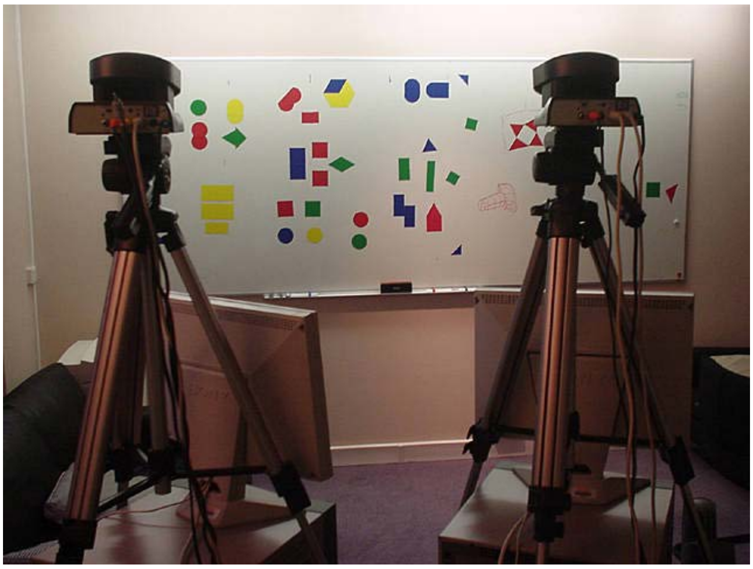


Figure 4.1: The Talking Heads experiment, designed by Luc Steels et his team at the Sony Computer Science Laboratory (Steels, 1999, 2003). A population of robots interacts to create and progressively negotiate a system consisting of its own particular semantic concepts, words and word-meaning associations that enables individuals to communicate effectively within their environment.

(Steels, 2012), and how diversity comes about and is linked to social structures (Coupé, 2003). There are models that have focused on the formation of specific semantic categories concerning, for example, space (Spranger and Steels, 2012), time (Dessalles and Ghadakpour, 2004), and perspective (Steels and Loetzsch, 2008).

4.3.2 Models of the morphogenesis of systems of vocalizations

The computer and robotic models of the formation of languages described in the previous section all presuppose that from the outset individuals share the same system of symbolic forms for building words. Thus, in practice, the "words" used in these models are composed of phonetic symbols that are already common to a whole group. Another family of models, using the same methodology, address the formation of systems of vocalizations among groups of individuals, that is to say the origin of phonological systems.

Origin of vowel systems

A pioneering model was developed by Berrah, Glotin, Laboissière, Bessière and Boë in 1996 (Berrah et al., 1996; Berrah and Laboissière, 1999), in which a population of artificial individuals generated a system of vowels over a number of pairwise interactions. Each individual started out with a repertoire of a few arbitrary vowels. During an interaction between two individuals, a vowel possessed by one of the two was selected at random. Following this the nearest vowel in the repertoire of the other individual was tweaked to bring it closer to the selected vowel, and the other vowels made to be further away, thus simulating a linguistic communication pressure. In parallel, a mechanism evaluating a compromise between perceptual distinctiveness and articulatory cost selected the best adapted individuals in a generational evolution model. In these simulations, using a mechanism similar to those of statistical physics, individuals converged after a certain time towards a shared system of vowels.

De Boer subsequently presented a related model that he used to study in detail the properties of vowel systems as they are formed (de Boer, 2001). In this simulation, which used artificial individuals, the same mechanism explains both the acquisition of vowels and their emergence: this mechanism is imitation. Thus, he undertakes in addition to answer the question: “How are vowel systems learned?”.

The individuals in de Boer’s simulation interact according to rules of a language game which is called the imitation game. Each individual has a vowel synthesizer, so that, given a point in articulatory space (tongue position, height or manner of articulation, lip-rounding) it can produce a vowel. Vowels are points in a space defined by the first two formants. Each individual also has a repertoire of prototypes which are associations between a point in articulatory space and the perceptual representation of the corresponding (acoustic) vowel. This repertoire is initially empty. It expands either by random invention or in the course of the learning which takes place during an episode of interaction. In one interactive episode between two individuals, one of them, the speaker, chooses a vowel from its repertoire and pronounces it to the other individual, the hearer. The hearer then looks to see which prototype in its repertoire is closest to what it has just heard, and pronounces it in imitation of the other individual. Then the speaker categorizes this sound by looking in its own repertoire to find the closest prototype. If it is the same as it used in its own initial utterance, it judges the imitation a success and makes this known to the other individual by telling it that it was good. Otherwise, it tells it that it was bad. Each prototype in a repertoire has a score which is used to reinforce the vowels leading to successful imitation, and to eliminate the other vowels. In a case of poor imitation, depending on the score of the prototype used by the hearer, either this is modified so as to resemble more closely the sound made by the speaker, or a new sound is created, as close as possible to that of the speaker.

It can be seen from the description of this game that it requires a certain number of complex abilities on the part of the individuals. For a start, they have to be able to follow the conventionalized rules of a game with successive turn-taking and asymmetric roles (speaker and hearer). Then, they need to be able to voluntarily copy the sound productions of other individuals, and be able to evaluate such copies. Finally, as speakers, they have to recognize when someone tries to imitate them intentionally, and give a feedback signal to the hearer to tell it whether or not it has succeeded. The hearer must be able to understand this feedback, to understand whether its imitation was successful as seen by the other individual.

The level of complexity necessary for the formation of shared vowel systems in this model is characteristic of a population of individuals which already have complex ways of interacting socially, and in particular already have a primitive system of communication (which enables them to know, for example, who is the speaker and who is the hearer, and what signal means “imitation succeeded” or “imitation failed”). The imitation game itself is a conventional system (the rules of the game), and the individuals communicate when they play it. This in effect necessitates the intentional transfer of information from one individual to another, and thus requires a system of shared forms making such information transfer possible. The vowel systems here do not emerge from an entirely non-linguistic situation. De Boer’s model concerns modelling how sounds are formed and how they change within a context where there are already some elementary linguistic interactions. This model shows how changes in systems of vowels can occur during the cultural history of a group of individuals. This results from the stochastic nature the model and the properties of the learning mechanisms in successive generations of individuals.

A generalized mathematical model of the mechanisms addressed in the models of Berrah et al. and of de Boer has recently been proposed by Moulin-Frier, Schwartz, Diard and Bessière (Moulin-Frier et al., 2008; 2015; Moulin-Frier, 2011). This model, expressed in the framework of Bayesian modelling, allows the essential elements of these mechanisms to be abstracted and makes explicit the respective roles of the motor and perceptual systems, that is to say the role of the body. It has also allowed the mechanisms involved in the formation of systems of sounds to be integrated into complete linguistic interactions, thus linking it to the "naming game" model described above: the artificial individuals in the Moulin-Frier et al. model negotiate simultaneously a repertoire of sounds and the associations of these sounds with the external referents about which they communicate. The model has also widened the works described above to the question of the formation of consonants and to the production of syllables, by including phenomena of co-articulation. Within a mathematical context, it can therefore express a variety of different theoretical perspectives on human speech – such as Liberman and Mattingly's motor theory (Liberman et Mattingly, 1985), or the perceptual theory of Diehl and his colleagues (Diehl et al., 2004) – as special cases of a more general sensorimotor theory. This means that computer experiments can be performed allowing some combination of these different theoretical perspectives (Moulin-Frier et al., 2012).

However, the models we have just described do not engage with the chicken-and-egg question of how the first repertoire of shared forms could have appeared in population of individuals without prior conventionalized modes of linguistic interaction. In particular the question of why individuals seek to imitate each other in the models of Berrah et al. or de Boer (where it is programmed in) is left unanswered. The same goes for the origin of language games such as the deictic games providing the conventionalized protocols of interaction that are simply presupposed by these two models. Nevertheless, along with those of Moulin-Frier et al., Kaplan, Steels and Kirby, they represent a crucial step forward in showing how conventional shared linguistic structures on the scale of a whole population of individuals can emerge through self-organization starting from local interactions that a are structured for language communication (that is to say the creation of a system of forms allowing them to distinguish between external referents). They are a naturalization of the cultural phenomenon of speech formation and evolution.

Origins of combinatoriality

One very important question about the origins of phonological systems is the question of combinatoriality. All languages include a repertoire of basic sounds that are used to form syllables (this is not the case for writing systems, where syllabaries are to be found in which the graphical symbols representing syllables are holistic and not re-used between different syllables). In all languages the vocal continuum is discretized: only a few key configurations are used in any given language, even though the vocal tract is capable of producing a continuum of sounds. And in all languages these basic sounds are re-used in a systematic way.

Models have consequently been developed to take into account the syllabic and temporal dimensions of vocalizations. Moulin-Frier's model, and also a model that we developed (Oudeyer, 2001c; Oudeyer, 2005a), feature language games where individuals negotiate systems of syllables. In the first case the language game is deictic, and in the second case it is an imitation game. In the model presented in (Oudeyer, 2005a), individuals play the same imitation game as in de Boer's model, but their repertoire consists of syllables. At the beginning of the simulation, they are given a shared repertoire of phonemes (imagined as the result of an earlier game, such as one of de Boer's simulations). The scores of the individuals' prototypes no longer depend only on success or failure in imitation, but also on the energy necessarily taken to produce them. It was possible to predict two regular properties of syllabic structure in human languages with this model. The first is a preference for CV syllables, followed by CVC, then CCV and CVV, then V and VC, then CCVC and CCVV, and finally other syllable types. The Sonority Hierarchy is also predicted by this model: syllables tend to begin with a phoneme with a high degree of obstruction to airflow, then to allow the obstruction to diminish to a certain point, and then to increase the obstruction again up until the final phoneme of the syllable.

This model can also shed some light on the learnability of syllable systems (Oudeyer, 2005a). While individuals have no problem learning a syllable system already constructed by a similar population of individuals[3], this does not happen given a randomly invented syllable system: if an arbitrary system of syllables is artificially incorporated into the memories of a population of individuals, any other individual who does not already know will generally be unable to learn it. This is due to the fact that the individuals' learning procedures, like all learning procedures (Duda et al., 2000), are biased, in that they are better adapted to learning certain families of structures than others (despite the fact that it is a general learning procedure that can be used for very different tasks, like learning to classify flowers according to their shape, or chemical substances according to measurements by electronic instruments). The fact that there are biases implies that certain types of data are easier to learn from than others. Thus, in this model groups of individuals culturally select syllable systems that they are capable of learning efficiently. In other words, syllable systems adapt to the ecological niche provided by the brains and morpho-perceptual organs of these individuals. This finding, which has also been shown to apply to the learnability of syntax (Zuidema, 2002), reverses the perspective taken by researchers in the nativist cognitive tradition presented earlier, who propose a different scenario to explain the fact that children learn a language and its grammar so easily, and especially the sound patterns, despite the poverty of the stimulus that they receive (Gold, 1967). The nativist argument points to the impossibility for a generic learning device with no a priori knowledge of language to learn any arbitrary language it is presented with. The (logically invalid) conclusion drawn from this is that the brain must know innately how language is organized, especially in its sound patterns, in order to be able to learn it. Thus, the brain itself would have adapted to language during the course of biological evolution in order to be able to learn it (Pinker and Bloom, 1990). The computer models discussed above suggest another explanation: it is the process itself by which languages are formed that develops and selects only those languages that individuals are capable of learning. It is quite possible that languages adapt to the general cognitive constraints of their speakers, rather than the other way around.

Even though they address the problems of the grammar of sounds, the models of Moulin-Frier et al. and Oudeyer make certain assumptions about the bases of this grammar: individuals are endowed from the start with a shared and discrete repertoire of phonemes (Oudeyer, 2005a), or are constrained to use a small number of phonemes, which is equivalent to programming combinatoriality in at the outset (Moulin-Frier, 2011). They therefore provide no real explanation of how discreteness and combinatoriality arose in the first place.

[3] The individual is quite simply made to interact with individuals who already "speak" the syllable system.

This is a central concern of the models we will present in the following chapters. However, one might reasonably have imagined that combinatoriality would arise during some of these simulations, for example in cases where individuals have a wide repertoire of phonemes at their disposal. Lindblöm (Lindblöm, 1992), along with Zuidema and de Boer (Zuidema and de Boer, 2009) suggested that combinatorial systems could be optimal compromises between perceptual distinctiveness and ease of articulation. Thus, just like the regularities in vowel systems or the phonotactic regularities emerging from de Boer's (de Boer, 2001) and Oudeyer's (Oudeyer, 2001d) models, one might have expected that among the phonemes put at the disposition of the individuals, certain phonemes would be chosen and systematically re-used in constructing syllables. However, this was not observed. A more precise study into this matter was carried out, varying the numbers of phonemes given to the individuals as well as the degrees of freedom in their vocal organs (a model of the vocal tract in the form of an acoustic tube) and their perceptual organs (a model of the cochlea). When the number of phonemes given to individuals was small in proportion to the number of syllables that their repertoire could contain, then necessarily they were systematically re-used, but this is equivalent to pre-programming combinatoriality, and leaves open the question of where these phonemes, a radical discretization of the articulatory continuum, come from. When the number of phonemes was increased, for all configurations of the vocal tract and the perceptual apparatus of any reasonable dimensions, the syllable systems which emerged (consisting of several hundreds of elements), were never combinatorial. Thus, individuals in the simulation were able to construct very large shared syllable systems and communicate efficiently without there being any need for combinatoriality.

Browman and Goldstein (Browman and Goldstein, 2003; Browman and Goldstein, 2000) proposed a different model addressing the question of the origin of discreteness (they refer to the emergence of discrete gestures). They constructed a simulation in which two individuals can each produce two gestures, parameterized along a dimension of constriction whose values are taken from a one-dimensional continuum (typically, this space is the space of place of articulation). The individuals interact following the rules of a game called an "attunement game" (this could be paraphrased as a "game in which each individual adjusts to the adjustments of the others"). In one episode of the game, the two individuals produce their two gestures, each one using a value of the parameter taken from the continuum with a certain probability. At the beginning of the simulation, this probability is the same for all values: in other words, all values on the continuum are equally usable. Next, each individual reconstructs the parameter used by the other individual for its first gesture, and compares it with the parameter which it has used itself. If the two values correspond, within a certain tolerance interval, then two things happen: the probability of using this value of the parameter for the first gesture is increased, and the probability of using the same value for the second gesture is decreased. This simulates the idea that the two individuals are each trying to produce their gestures differently (so that they can be differentiated and contrasted), and at the same time in a way similar to the other individual (so that a shared conventionalized usage is established). At the end of the simulation the individuals converge on a state where they use a single value for each of their gestures; thus the space has been discretized, and the pairs of values of the individuals are the same in each simulation, but different from one simulation to the next. Browman and Goldstein carried out simulations both using and not using nonlinearities in the function from sounds to articulatory parameters (this is implemented by modelling the noise added when an individual reconstructs the parameters used by the other individual). When nonlinearities were not used, the set of parameters used across all simulations covers the space uniformly. When nonlinearities were used, then certain parameters in regions of stability were statistically preferred.

As in the models described above, here the individuals interact in accordance with a coordinated structure: they obey the rules of the game. Indeed, every time, they have to produce their gestures together during one interactional episode. Thus, as in the imitation game, a pressure to differentiate sounds is built in, as is a pressure to copy the parameters used by the other individual. This implies a presupposition that the individuals already live in a community in which a complex communication system exists. However, this was certainly not among the question that this model sought to address. So it still remains to be seen how the discreteness of speech, which seems to be crucial for the birth of language (Studdert-Kennedy and Goldstein, 2002), could have appeared without assuming that a complex communication system was already in place. More precisely, how could a discrete speech system have appeared without an explicit pressure to contrast sounds? This is one of the problems will be examined in the following chapters.

One assumption made in Goldstein's model is that the individuals directly exchange the articulatory targets which they use in producing gestures. However, human vocalizations are continuous trajectories, firstly in acoustic space, and secondly in the articulatory space of relations between organs. Thus what a human perceives in the vocalization of another is not articulatory targets which were used to specify gestures, but rather the acoustic realization of these gestures which is a continuous trajectory between the starting state and the target. And because several targets are aimed at in sequence, vocalizations do not stop when one target is reached, but continue on their path toward the next target. To retrieve the targets from the continuous trajectory is very difficult, and moreover a task which speech recognition engineers have not conquered. Possibly the human brain is equipped with an innate capacity to do this, by detecting events in the flux of sounds which correspond to the targets, but this is strongly speculative.

4.4 Mechanisms that give rise to language

We have seen that operational models developed in the literature show how the cultural phenomenon of language formation can be naturalized, by demonstrating that languages can emerge by self-organization out of the decentralized interaction of individuals. The capacities for interaction and communication possessed by the individuals in these simulations are already quasi-linguistic, with, for example, the capacity to play language games, conventional interactions which are themselves norms structurally equivalent to primitive linguistic norms. These capacities are so complex that we are still a long way from understanding how natural or cultural selection could have formed them.

In the following chapters we will ally with the research tradition that builds artificial systems to support explanations of the mechanisms of speech morphogenesis. However, the structures constituting the initial biological baggage of the individuals will be much less complex[4]. In particular, the individuals will not have the capacity to interact in structured ways, they will not have access at the outset to any conventional norm, and no measures of optimality or communicative effectiveness will be used to guide the evolution of systems of vocalizations. In this way, starting from relatively simple mechanisms, the speech codes which will be generated through self-organization during the interaction of these structures

[4] It is crucial to note that we are speaking here of 'evolutionary' complexity, which is unrelated to the complexity of the program implementing the artificial system, that depends on the degree of detail one wishes to model.

will be characterized by the following properties: discreteness and combinatoriality will be self-organized; the repertoires of emerging units will be shared by all the individuals in a group and variable from one group to the next; and the way phonemes are put together will be organized according to the rules of syntax. In addition, the use of morphological constraints will lead us to make the same predictions about vowel systems as those made in the models described above, but with assumptions that are less complex from an evolutionary point of view.

These less complex structures which, then, will constitute the assumptions of these systems, are interesting in two ways:

- First, these structures can be considered within the framework of a classical neo-Darwinian theory of the origin of language. Within an evolutionary scenario assuming a pressure for linguistic communication, the models presented in the next chapters will make it more understandable how natural selection could have arrived at the biological bases of speech. In effect, we will show how these bases can be relatively simple by comparison with the speech systems that they generate. We will thus propose an explanation analogous to D'Arcy Thompson's explanation for the formation of hexagonal cells in beehives: the biological structures that natural selection arrives at for the bees are not those of a mathematician armed with a compass and rules and exact plans for dividing up the surface regularly, but simpler structures allowing for cells of roughly similar size, not too misshapen, to be piled together.
- Secondly, the generic nature of these structures will allow us to explain in Chapter 9 how combinatorial systems of speech could have appeared spontaneously as collateral effects of the evolution of capacities that are more generic and anterior to language, such as imitation and intrinsic mechanisms that incite an organism to explore its own body, and in particular its own vocal system, out of pure curiosity (Oudeyer, 2006b; Oudeyer and Smith, 2016). This will lead us to suggest that the speech codes that we use nowadays might be exaptations, the first versions of which might have been the outcome of self-organization of structures whose origin was unconnected to language.

In both cases the modelspresented in the next chapters[5] will attempt to illuminate the issue of the origin of the language faculty (rather than the origins of specific languages), by showing how the evolution of one its fundamental prerequisites, the speech code and its biological bases, could have been crucially facilitated by phenomena of self-organization.

Before presenting these models in detail, we will focus more broadly on the scientific justification for an approach which builds artificial, computer-based and robotic systems in order to better understand how living creatures function.

[5] These models will also be robotic to the extent that they simulate the physical properties of the vocal tract and of the ear

Chapter 5

Epistemology of Artificial Systems as Research Tools

The mechanisms that could have given rise to language, and in particular to speech, are necessarily complex and cannot be explained only by theories that are expressed in words. Theories whose formulation is verbal, as is often the case in the social sciences and sometimes in the life sciences, remain approximate about their assumptions because they use natural language. Moreover, their verbal nature makes it impossible to prove that their premises lead to their conclusions, because these theories necessarily involve complex, dynamic systems whose behaviour is difficult to predict exactly using intuition alone. Building artificial systems, whether they be mathematical, or computer or robotic simulations, is a way of complementing a purely verbal approach and evaluating the logical coherence of a given theory; it is also a way of generating and formulating new theories. This has been called "the methodology of the artificial" (Steels, 2001; Pfeifer, Lungarella et Iida, 2007). This approach, widespread in biology (Langton, 1995; O'Reilly, 2006), has been used more and more in evolutionary linguistics in recent years, as we saw in the previous chapter (Steels, 2012). We looked at examples of artificial systems showing the formation of shared systems of vowels, consonants, syllables, lexicons, semantic categories, and syntax. We will now examine the epistemological foundations of this approach.

5.1 What is the logic of science?

Science has been perceived in a variety of different ways by different thinkers, including Bachelard, Bernard, Chalmers, Feyerabend, Kuhn and Popper (Bachelard, 1865; Bernard, 1945; Chalmers, 1990; Feyerabend, 1979; Kuhn, 1970; Popper, 1984). We adopt here the constructivist point of view as propounded by Glaserfeld (Glasersfeld, 2001).

Several different types of activity can be distinguished as belonging to the sciences and practiced by scientists:

- Performing experiments, observing reality, analyzing data to detect regularities
- Examining existing theories in order to derive predictions about reality from them, and devising experiments to falsify or confirm a theory.

- Determining whether what is observed corresponds to what the different theories predict.
- Examining the internal coherence of theories and the compatibility or incompatibility between theories; trying to build bridges between different theories.
- Constructing new theories. There are several ways of arriving at a new theory. One of them is a process of "abduction", a term invented by Peirce (Peirce, 1931). This consists, given the state of a system, in figuring out an initial state as well as a mechanism which could have led to the observed state. It is thus a matter of finding a set of premises which leads to a certain conclusion. It is in some way the inverse of 'deduction' (Peirce, 1931). For example, if one sees a pair of shoes sticking out from behind some curtains, one can generate the hypothesis that there is someone behind the curtains. This is abduction. Obviously, the theories and hypotheses generated could be false: maybe the shoes are empty. Consequently, once a theory has been formulated, it is necessary that researchers engage in activities 2), 3) and 4) to evaluate the correspondence between the theory and reality (but a theory can still be useful even if it corresponds poorly with reality as explained below).
- Very importantly, formulating new questions or reformulating old questions so as to guide minds towards uncharted aspects of a problem, and sometimes open the way to surprising discoveries.

In this book we use the logic of abduction to formulate new theories, formulated in the language of mathematics, computer science and robotics. Computer simulations are used to evaluate the internal coherence of these new theories, and their predictions can be tested through comparisons with what is observed in reality. The use of abduction is almost inevitable, given the problems attacked here. In effect, we discuss a quest for a theory about the origins of certain aspects of speech; but the state that humans were in when speech arose has left few traces. In this way, the construction and testing of computer models can be seen as a novel scientific language in which old questions about speech, along with new ones, can be formulated, and the multidimensional structure of this complex system adequately expressed.

We will also be using the term "artificial system" to qualify the models presented in this book. This is a reminder that the aim of these models is not to mimic reality realistically, but rather to help us to understand reality with the aid of a functional abstraction of fundamental mechanisms. Another term we will use to qualify them is "theory". We believe that artificial systems are themselves structures of explanation, of a non-verbal kind, which can be essential for understanding the world. Thus artificial systems, even though they are often inspired by verbal theories, are not simply formal translations of these theories. Formal languages, whose properties are different from those of natural languages, are able to express and make explicit mechanisms that cannot be described precisely using natural language (and the reverse might also be true). For this reason, we believe that an explication of the origin of natural human languages cannot be couched entirely in a natural language, and that for this purposes other languages, such as formal languages, are required.

5.2 What is the point of constructing artificial systems?

Let me now address some criticisms touching on the usefulness of this kind of formal approach when trying to explain structures arising in living organisms, including languages and speech. First of all, how can we link abstract entities in a computer program with the real world? How do we verify the hypotheses embodied in the program? Is it even justifiable to attack a problem like the origin of speech, since after all there is a high likelihood that we will never be able to verify our theories, given that so few traces remain?

In short, what is the point of an artificial system? How can it be verified? Answering these questions requires to first set out our view of science as a highly constructivist activity (Glasersfeld, 2001). Science can be seen as an activity that seeks to form representations (often in the form of real or abstract machines) which help us to understand the world we live in. In other words, theories are representations of the world that help us to put some order into our view of it. To be sure, there are constraints on how we construct these representations, with the result that certain representations such as those to be found in the Bible are not scientific theories. An important constraint is that there must be no "miracles". Furthermore, these representations must form a coherent whole, and it should be possible to pass from one to the other via logical connections. There may be gaps in the picture they paint, and in fact a large part of scientific activity is guided by the goal of filling these gaps. These representations must also be compatible with one's observations, which themselves depend on the theoretical context in which they are made. From time to time, new theories are seriously at odds with the set of representations accepted by scientists of one era, but scientists end up replacing the old narratives with new ones, because they judge them to be more useful in understanding the world (this is what Kuhn calls a paradigm change, see Kuhn, 1970). An example is the shift from Newtonian physics to the theory of relativity. This vision is summarized by Einstein:

> "Physical concepts are free creations of the human mind, and are not, however it may seem, uniquely determined by the external world. In our endeavor to understand reality we are somewhat like a man trying to understand the mechanism of a closed watch. He sees the face and the moving hands, even hears its ticking, but he has no way of opening the case. If he is ingenious he may form some picture of a mechanism which could be responsible for all the things he observes, but he may never be quite sure his picture is the only one which could explain his observations. He will never be able to compare his picture with the real mechanism and he cannot even imagine the possibility or the meaning of such a comparison" (Einstein and Infeld, 1967[1938], p31).

Einstein is thus able to enounce what amounts to a central principle:

> "The object of all science, whether natural science or psychology, is to co-ordinate our experiences and to bring them into a logical order" (Einstein, 1955[1922]).

On this basis, we defend here the idea that a theory can be useful if its connection with reality is tenuous or even if observations show it to rest on false premises. For a start, there are phenomena that we understand so poorly that it is already useful to be able to imagine which families of mechanisms might explain them. Let us return to the example of ice crystals whose morphology has been the subject of careful study by Nakaya (Nakaya, 1954) and Libbrecht (Libbrecht, 2004). In physics there are are well established theories on the behaviour of water molecules. In addition, we know quite well the shape of ice crystals, and even under what conditions of temperature, pressure and humidity these shapes appear. However there are numerous aspects that we still do not understand concerning the mechanisms that cause the physics of water molecules to give way to the

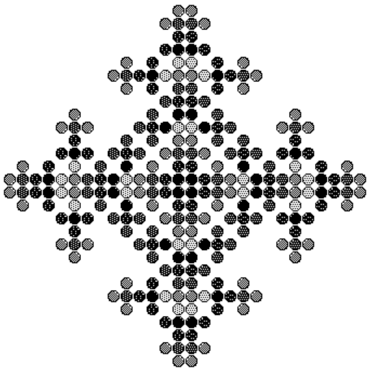


Figure 5.1: This figure was obtained after 13 iterations of the rule “if a neighbour is on, switch on too”, starting from an initial state with four central cells switched on and all the others off. The different colours correspond to different steps in the iteration (the black cells correspond to stages 0, 5 and 10). This shows that rules of local interaction can lead to the growth of shapes resembling ice crystals, thanks to self-organization. For more information, see http://www.lps.ens.fr/˜weisbuch.

physics of crystalline structures. In addition to purely mathematical analytical models (Langer, 1980), researchers have set up computer simulations that have led to significant progress in our understanding of these mechanisms (Wolfram, 2002). They used cellular automata. These are grids in which cell can be in an ‘on’ or an ‘off’ state. Their state evolves as a function of their neighbouring cells and according to rules of the type: if one neighbour is on, then switch on too. Wolfram (Wolfram, 2002), for example, presents rules of this type such that starting from a group of four cells switched on at the centre of the grid (the seed molecule that starts the crystalline growth), a shape similar to that of ice crystals appears spontaneously by growth and self-organization, as shown in Figure 5.1. At the level of micro-structures, this cellular automaton is very different from real water molecules. However, this simulation suggests a particularly stimulating vision of the formation of ice crystals: after having seen it, it is easy to imagine that only the interaction of water molecules, with their known properties, even if they have nothing to do with the symmetry of crystals, can lead spontaneously to their formation without the need to appeal to other forces. Thus one aspect of the properties of ice crystals is no longer a miracle. At the same time, thanks to other computer simulations modelling the physics of water models, the formation of specific crystalline structures can be studied quantitatively rather than qualitatively (Kobayashi, 1998).

Next, a theory, especially a formal or computational theory, can be useful in evaluating the coherence of other theories or logical compatibility among several theories. A theory which is at some distance from reality or unverifiable can still bring order to the set of existing theories. In fact, many of the representations which scientists construct, especially in the social sciences, are verbal. Because they make only approximate assumptions or depend on intuition to derive their conclusions, they may contain errors, or on the other hand not be convincing enough for part of the scientific community. We will give examples of simulations in computer science which allowed the subject to advance in these two kinds of case (for numerous other examples see (Oudeyer, 2010).

The first example concerns imitation and representation of the self by babies. It has been suggested that when babies imitate the arm movement of another person with a movement of their own arm, this shows that they must have some sort of 'other-awareness'. This has led some researchers to deduce that a baby possesses a rudimentary form of Theory of Mind, and a fortiori that it possesses primitive representations of other beings (Guillaume, 1925). However, a group of scientists comprising roboticians, psychologists and neuroscientists have presented a robotic model in which a robot copies an arm movement of someone else, even though it has no concept or representation of 'other' (Andry et al., 2001). The copying is an effect of what they call perceptual aliasing: the robot thinks that the arm that it sees is its own, and corrects the error between this perception and the state of its motors as a consequence of a form of cognitive homeostasis. This shows that the observation of an apparently complex imitation behaviour does not require a Theory of Mind. To be sure, this does not show whether or not the baby has a Theory of Mind or not when it acts in such a way, but this experiment constrains the intuitive conclusions that one may reach on observing its behaviour. The way in which the robot copies could be very different from the way in which the baby copies, but this robotic simulation acts as a very powerful cautionary tale. This kind of model is interesting because we know exactly what is in the robot or the program, and this opens the way to suggestive comparisons based on behavioural analogies between the robot and living beings (many researchers also use comparisons between animals, which has its advantages, but also the disadvantage that we are a long way from understanding the cognitive and behavioural mechanisms of animals!).

A second example of a computer simulation concerns the case of the Baldwin Effect. In the 19th century Baldwin (Baldwin, 1896) and Morgan (Morgan, 1896) developed a theory that genetic evolution could be modified by learning. More exactly, they proposed that certain acquired behaviours could become innate (genetically hardwired) with the passage of generations and at the level of the whole population. This theory was mainly verbal and was ignored by the scientific community until recently (French and Messinger, 1994). This was not surprising because this theory is quite close to Lamarckism, which says that evolution consists in the inheritance of acquired characteristics, an idea that was long discredited (until recent discoveries in epigenetics showed that the mechanisms of heredity are more complex than previously thought (Richards, 2006)). However, Hinton and Nowlan (Hinton and Nowlan, 1987) presented a computational model proving that the concept of the Baldwin Effect is compatible with Darwinian theory. Their model consists of a population of sequences of 0 and 1 and 'wildcards', postulated to be the genes of a population of individuals. Each gene codes for a trait which can have the value 0 or 1. This value can be innately specified in the genome or acquired (by using the 'wild cards'). The individuals are evaluated according to a selection function: if all the individual's traits, after using its wild cards, are 1, then the individual has the maximal adaptive value; otherwise it has zero adaptive value. This adaptive value is used to determine which individuals reproduce the most. The simulation showed that when 'wild cards', modelling learning, are allowed, the population converges much faster on the situation in which everyone has all traits set to 1, than when learning is not allowed. Moreover, at the end, all the wild card alleles disappear and all the traits are set to 1 innately. This model is very distant from real biology. The assumptions behind it are even false (for example it is assumed that there is no structural difference between the genotype and the phenotype). However, it has changed evolutionary biology: the idea of the Baldwin Effect is now considered as plausible by biologists. Whether the phenomenon actually occurs in nature is still an open question, but science has taken a step forward.

To summarize these different examples, abstract models (mathematical or computer based) can be a long way from reality but nevertheless useful for drawing and constraining the contours of the research space of theories. They can point to new ways of understanding phenomena that were previously mysteries.

This is the epistemological framework in which the models presented in the next chapters were designed. We do not intend to propose definitive explanations of how speech originated. The complexity of these questions means they remain largely beyond the reach of a research domain that is still in its infancy. The aim is exploratory: we will try to formulate and suggest families of mechanisms that are plausible from an operational point of view and capable of generating some of the speech structures that we are seeking to explain. The relationship between these models and human systems is not a relationship of identity, but a relationship of analogy. We will attempt to show that this kind of model is useful for spelling out the contours of the space of possible explanations that have so far been proposed, as well as generating new types of explanation. The principal aim of this work is to participate in a new way of organizing the theoretical speculation about the origin of speech, in which the concept of self-organization has a central role.

Chapter 6

A Model of Self-Organization of Systems of Vocalizations

This chapter describes a simple version of an artificial system for studying certain aspects of the morphogenesis of systems of vocalizations. It consists of a computer model simulating a group of individuals that are capable of perceiving, producing and learning vocalizations. Inasmuch as this system includes models (albeit simple and simulated) of physical properties of the vocal tract and the ear – in other words, of the body – it is also a robotic model. These individuals move randomly within a virtual environment, producing vocalizations, that is to say movements of their vocal tract, in a kind of baby-talk or babbling that is statistically influenced by the vocalizations uttered by their neighbours. Through a progressive self-organization all the individuals will come to share the same system of vocalizations where a small number of elementary sounds are systematically re-used for constructing and perceiving vocalizations that are analogous to syllables. The system thus becomes discrete and combinatorial.

6.1 Neural Networks, unsupervised learning and vocal mechanisms

To construct the individuals' speech production and perception mechanisms, we will adopt the point of view of articulatory phonology (Browman et Goldstein, 1986), and more generally of a whole section of the research community in mammalian motor control (Kandel et al., 2001), which defends the idea that gestures (or more generally relations between organs) are the central elementary representations out of which the vocal movement is constructed. Here, to be precise, a 'gesture' denotes a command specifying an articulatory target, itself defined by a relation between several organs (like place of articulation or the distance between the lips)[1]. The following simplification is made here in relation to the description of vocalizations made by articulatory phonology: instead of having vocalizations defined by several parallel tracks of gestures, there will only be a single track. This means that a vocalization will be a sequence of gestures strung together, but there will be no other

[1] However, the results presented here are also compatible with the idea that articulatory targets are configurations defined by the individual positions of the organs, or else defined by a set of formants, for example, in perceptual space

gestures to carry out in parallel. In addition, for the purpose of visualizing the results, only one, two or three dimensions will be used to define the articulatory targets (which could be seen as place and manner of articulation as well as lip rounding, for example). One vocalization will consist of a sequence of two, three or four gestures. These gestures, which are motor commands, are executed by a lower-level control system which makes the organs move continuously to achieve articulatory targets according to a specific timing.

In the version of the system presented in this chapter, individuals will be assumed to be capable of retrieving the trajectory of relations between organs corresponding to a vocalization that they hear[2]. In this chapter individuals are also assumed to be capable of finding the muscular activations which will move the organs in such a way that the relations between organs specified corresponding to a given sound are realized. Various works in the literature present mechanisms explaining how these translations from one space to another can be learned and realized (Bailly et al., 1997; Howard and Messum, 2011; Moulin-Frier and Oudeyer, 2012). An example will be given in the following chapter. For the moment, we simply assume that individuals can do this. This will allow us to use just a single representation in this chapter, namely that of relations between organs. Thus, acoustic or muscular representations will not be used, which will give us a more intuitive visualization and understanding. The term ‘sound’ will be used to refer to a vocalization, but only the trajectory of relations between organs will be manipulated (thus we allow ourselves in this section a slight extension of the term ‘sound’). What an individual perceives directly in the vocalization of another individual is the continuous trajectory of relations between organs (but not the gestures which produced it).

The next sections describe in detail the foundations, that is to say the pre-suppositions, of the mechanisms comprising the internal architecture of the individuals.

Neural units

The artificial brain of each individual is composed of neural units, in the notation n_i (see figure 6.1). A neural unit is a box which receives input signals, corresponding to measurements, and integrates them to calculate a level of activation. Functionally, the neural unit models certain operations performed by biological neurons. The integration is typically achieved by first comparing the vector of all the inputs with an internal vector, noted v_i and called the ‘preferred vector’, peculiar to each neural unit[3]. Then, the result of this comparison is filtered by what is called an ‘activation function’, which calculates the activation, or response, of the neuron.

The activation function here is a Gaussian function, whose width is a parameter of the simulation. Gaussian activation functions ensure that there is one input which produces the greatest response in the neuron: this is the input with the same value as v_i. This is why it is called the preferred vector. When inputs are distant from the preferred vector in the space of inputs, the level of activation decreases according to a Gaussian function. When the width of the Gaussian is broad, this implies that it is not very specific, i.e. that there are many different inputs to which the unit responds significantly.

[2] The expression ‘trajectory of relations between organs’ is used instead of ‘sequence of gestures’ because the individuals are not assumed to be initially capable of retrieving the articulatory targets from which the continuous trajectory was generated (articulatory targets will be points like others on this trajectory). In fact, any point on this trajectory could be a target used in specifying the production of the vocalization. This means that initially the individuals are not capable of detecting the ‘sound events’ which might correspond to articulatory targets.

[3] This internal vector corresponds to the weighting of neurons often used in the neural network literature.

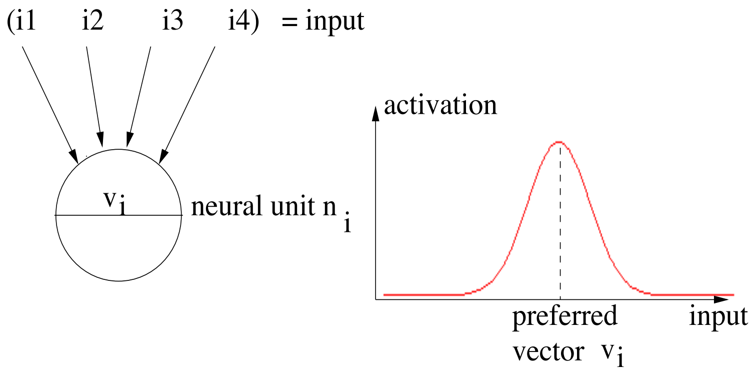


Figure 6.1: A neural unit and its Gaussian activation function. The input is a space of dimention D. To increase clarity, the projection of the activation function on only one of its input dimension is shown.

The preferred vectors of neural units change as and when new inputs are perceived. This is how learning happens, and is described below under the heading “Plasticity”. The width of the Gaussians does not change when stimuli are perceived; that is fixed. Thus the centres of the activation functions evolve over time, but not their width [4].

[

Assumption 2: perceptuo-motor correspondences]Translation between perceptual space and organ relation space

As said above, in this chapter an individual is assumed to be capable, given a continuous acoustic signal trajectory, of retrieving the corresponding trajectory of relations between organs [5]. However, individuals are not initially able to retrieve the articulatory targets which were used to construct the trajectory of relations between organs. Given a sequence of articulatory targets in organ relation space, individuals are also assumed to be capable of finding a trajectory in the space of muscular activations which will move the organs in such a way that the targets are achieved[6]

This will allow us to use only the level of representation dealing with relations between organs and thus give a more intuitive visualization and understanding. If the three repre-

[4] If $tune_{i,t}$ denotes the activation function of n_i at time t, s a stimulus vector, and v_i the preferred vector of n_i, then the form of the function is:

$$tune_{i,t}(s) = \frac{1}{\sqrt{2\pi}\sigma} * e^{-\frac{1}{2}|v_i - s|^2/\sigma^2}$$

The notation $\mid v_1 - v_2 \mid$ denotes the norm of the difference between vectors v_1 and v_2. The parameter σ determines the width of the Gaussian, and so if it is large the neurons are broadly tuned. A value of $\sigma^2 = 0.001$, as used below, means that the neural units respond significantly to about 10 percent of the space of inputs.

[5] The next chapter describes how the type of translation that we assume here might be learned by the individual.

[6] The function mapping organ relation space to the space of muscular activations is not an isomorphism: a target in organ relation space can often be realized by several organs or combinations of organs. We assume that at least one possibility can be found. How the choice between several possibilities can be made is described in the literature (Bailly et al., 1997).

sentations had been used, as schematized in Figure 6.2, there would have had three neural networks, each composed of neural units coding a space. There would be a perceptual neural network, composed of neurons receiving their inputs from the activation of the cochlea. These inputs could be formants, for example. Next, there would be neural units in a network of the relations between organs, connected to neurons in the perceptual network. The third neural network would be composed of neural units coding for muscular activations, and connected to the neurons in the network for relations between organs. After that, the activation of neurons in the muscular neural network would be used to control the speech organs, producing sounds. Technically, assuming that individuals are capable of passing from one representation to another means that the connections between neural maps are such that when a sound activates the cochlea and the neurons in the perceptual network, then the neurons in the network of relations between organs which are the most highly activated have the preferred vector corresponding to the relation between organs which produced the sound. In addition, the neurons in the network of muscular activations which are most activated have the preferred vector corresponding to the muscular configuration which produced this relation between organs. Several papers have shown how these connections could be learned during the course of infant babbling (Bailly et al., 1997; Oudeyer, 2003b).

Because only the organ relation space is used, the neural units take inputs directly coded in terms of relations between organs. Producing a vocalization, programmed by a sequence of activations of several n_i as explained below, is therefore solely a matter of forming a continuous trajectory in organ relation space. The individuals only exchange trajectories. Figure 6.3 shows this simplification schematically.

There is also a module called "inhibition/activation" which can send a GO signal allowing a vocalization to be produced when the neural units are activated. This is to say that in the absence of this signal, the commands specified by the activation of the neural units are in effect not carried out. This means that the activation of neurons n_i by an external sound does not directly provoke the reproduction of this sound. Copying requires the GO signal. In the system, individuals never use this GO signal when they have just heard a sound, but only at random instants, with the result that they do not copy what they have heard, and thus in practice we do not see individuals imitating the vocalization of another individual immediately after hearing it. However, as we will see, the neural system adapts in such a way that the vocalizations of individuals are influenced by those that they hear.

Perception and plasticity

The last section explained that what an individual perceives in the vocalization of another individual is a continuous trajectory in organ relation space. Now we will explain what is done with this trajectory and how this changes the preferred vectors of neural units.

At the outset, the individuals are not able to detect high-level events in the continuous trajectory, which would allow them to find what points correspond to the articulatory targets used by the individual who produced the vocalization. They thus segment the trajectory into lots of little slices, corresponding to the temporal resolution of perception (if we had used the three representations, this would correspond to the temporal resolution of the cochlea). Next, each little slice is averaged[7], giving a value in organ relation space, which is then sent to the neural units. These neural units are then activated according to the formula given in the "Neural Units" section. Figure 6.4 shows this process schematically. The horizontal axis represents time, and the vertical axis represents organ relation space. Here organ relation

[7] If we are working in more than one dimension, each dimension is averaged.

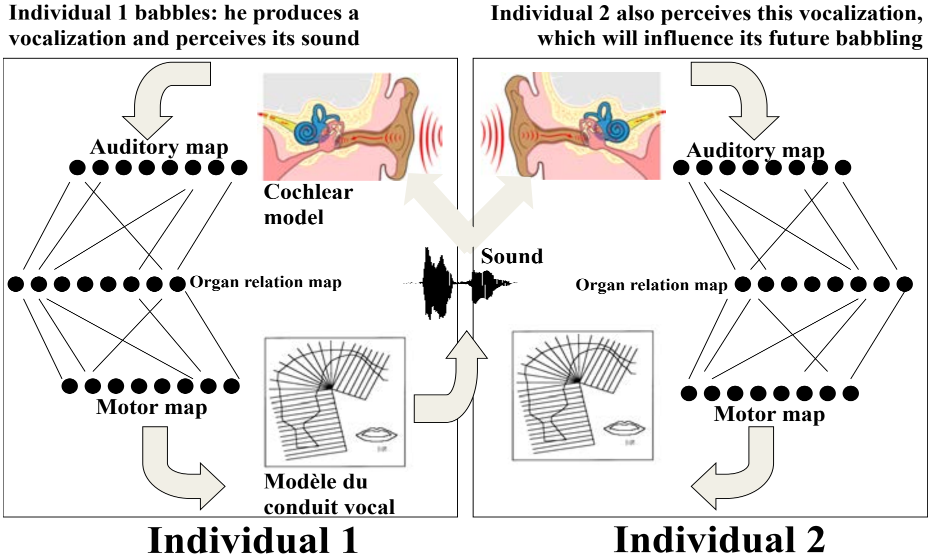

Individual 1 babbles: he produces a vocalization and perceives its sound
Individual 2 also perceives this vocalization, which will influence its future babbling
Auditory map
Cochlear model
Organ relation map
Sound
Organ relation map
Auditory map
Motor map
Modèle du conduit vocal
Motor map
Individual 1
Individual 2

Figure 6.2: A general architecture of a full system involving the three representational spaces. Individuals include models of the ear, the vocal tract and its control and the neural system connecting them, and are capable of learning. They move around randomly within their environment while babbling, neural activity occurring in an arbitrary fashion. This give rises to vocalizations that are perceived by the individuals themselves and by their neighbours, who make no distinction between the vocalizations of others and their own. The perception of vocalizations affects the neural system, which in turn makes the production of future babblings be shifted towards the distribution of vocalizations that have been heard. In this chapter we make the simplifying assumption that correspondences between the auditive and motor representations of vocalizations are already known to individuals, and so only an intermodal representation is used (organ relation map). The next chapter shows how such correspondences may be learned by individuals during the course of their babblings and as a shared combinatorial system self-organizes within their group.

space is one-dimensional. The continuous line represents a vocalization perceived by the individual. The little segments into which it is divided represent the averages of the little slices extracted by the temporal filter. These average values of these little segments are the inputs given sequentially to the neural units. The neural units of this individual are represented on the vertical axis: each point corresponds to a preferred vector. In effect, the values of these points are within organ relation space. Each of the neural units is activated by each of the averages.

When they are activated, the neural units are modified. This means that their preferred vectors are changed. This change is a sensitization to the stimulus for those neurons which responded significantly. This implies that if the same stimulus is given as input immediately afterwards, the response of the neural units will be a little bit stronger. Each such change is very slight, and weighted by the activations of each neural unit (a value between 0 and 1). Units which are very active change more than units which are not. Figure 6.5 illustrates the changing of preferred vectors. This figure represents the activation function of a neural unit before it receives any input, and after it has received and processed one: the preferred vector,

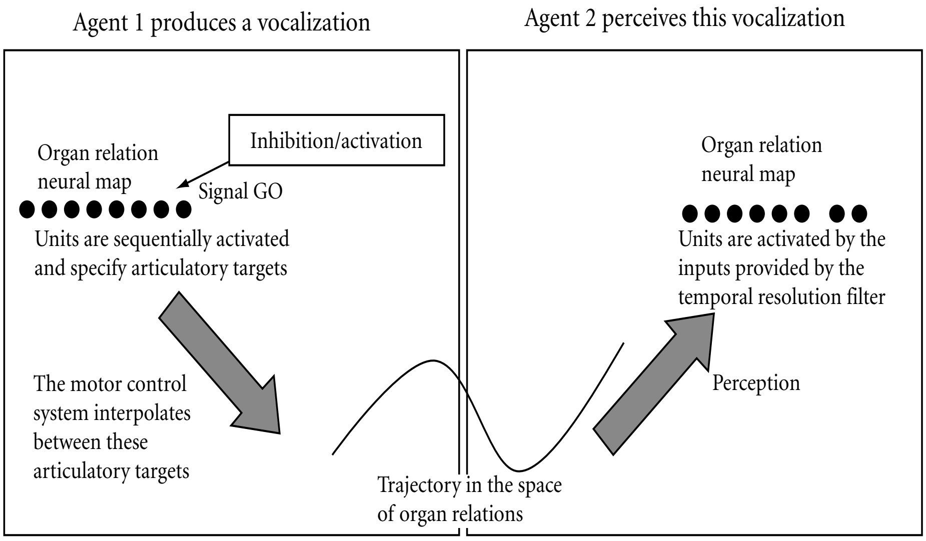

Agent 1 produces a vocalization
Agent 2 perceives this vocalization
Inhibition/activation
Organ relation neural map
Signal GO
Units are sequentially activated and specify articulatory targets
The motor control system interpolates between these articulatory targets
Trajectory in the space of organ relations
Organ relation neural map
Units are activated by the inputs provided by the temporal resolution filter
Perception

Figure 6.3: In this chapter one uses only the representation level of relations between organs, because individuals are assumed to be capable of passing from perceptual space to organ relation space, and from there to the space of muscular activations.

that is the centre of the Gaussian, has been shifted. We see that from a geometrical point of view the preferred vector is shifted in the direction of the stimulus, and the amplitude of this shift depends on the activation of the neuron[8].

Production

The production of a vocalization consists in choosing a sequence of articulatory targets and moving the vocal tract so as to achieve them. These articulatory targets are specified by relations between organs, which are one-dimensional in the first simulations which will be carried out. To choose these articulatory targets, an individual sequentially and randomly activates units in its neural network, and at the same time sends the GO signal described above. This activation is a command implementing the concept of gesture in this work. A target is specified by the preferred vector of the activated neural unit. Then there is a control system which executes these commands by making the relations between organs move continuously and sequentially toward the target[9].

[8] The mathematical formula of the new activation function is

$$tune_{i,t+1}(s) = \frac{1}{\sqrt{2\pi}\sigma} * e^{|v_{i,t+1}-s|^2/\sigma^2}$$

where s is the stimulus, and $v_{i,t+1}$ the preferred vector of neuron n_i after the processing of s:

$$v_{i,t+1} = v_{i,t} + 0.001 * tune_{i,t}(s) * (s - v_{i,t})$$

[9] Note that this way of producing vocalizations already contains elements of discreteness. This model assumes that vocalizations are specified by a sequence of targets. This is in fact in agreement with the

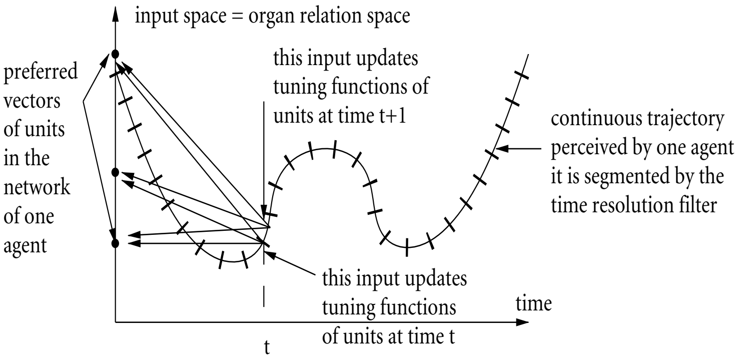

input space = organ relation space
preferred vectors of units in the network of one agent
this input updates tuning functions of units at time t+1
continuous trajectory perceived by one agent; it is segmented by the time resolution filter
this input updates tuning functions of units at time t
time
t

Figure 6.4: Each individual obtains a continuous trajectory (in organ relation space) when it perceives the vocalization of another individual. It then uses a temporal resolution filter which segments this trajectory into a sequence of very short parts. For each of these parts, the average is calculated, and the result is a stimulus which is sent to the neural network, whose units are then activated. After reception of each stimulus and activation of the units, these are updated.

If three representations had been used, the control system would activate the muscles in such a way that the organs would be moved towards configurations satisfying the specifications of relations between organs. Here, the control system directly generates a continuous trajectory in organ relation space which passes through the targets. This is realized by polynomial interpolation. Figure 6.6 shows this process of production schematically. In this figure the horizontal axis represents time and the vertical axis represents organ relation space. The preferred vectors of the neural units of an individual's network are represented also on the vertical axis. Five of these units are activated sequentially, defining five articulatory targets. Then the control system (polynomial interpolation) generates a continuous trajectory passing by all these points. This trajectory is the vocalization which will be perceived by the individuals who hear it.

One crucial point is that the neural units n_i are used both in the process of perception and in the process of production. Consequently the distribution of articulatory targets used in production is the same as that of the preferred vectors, which themselves change as a function of the vocalizations heard in the environment. This implies that if an individual hears certain sounds more often than others, it will also tend to produce them more often than other sounds (a "sound" here refers to one of the little subdivisions or slices of a vocalization created by the temporal filter, as described above). We may remark that individuals do not attempt to reproduce vocalizations directly after hearing them produced

literature on mammalian motor control (Kandel et al., 2001), which describes it as organized at two levels: a level of discrete commands (our articulatory targets), and a level concerned with their execution. Thus the element of discreteness seen in discrete commands should not be a trait which it is necessary to explain in the context of research on the origins of speech, since it is already present in the motor control architecture of mammals. However, the model does not assume that the articulatory targets are initially organized: the set of commands used to define the targets is a continuum, and there is no re-use of articulatory targets from one vocalization to another; discreteness and a form of primitive combinatoriality will emerge from the simulations. The model also does not assume that there is initially any discreteness at the level of perception, in the sense that initially the individuals are not capable of perceiving “events” in the flux of sound (however, at the end of a simulation it will be possible to identify the categories of articulatory targets used to produce the vocalization).

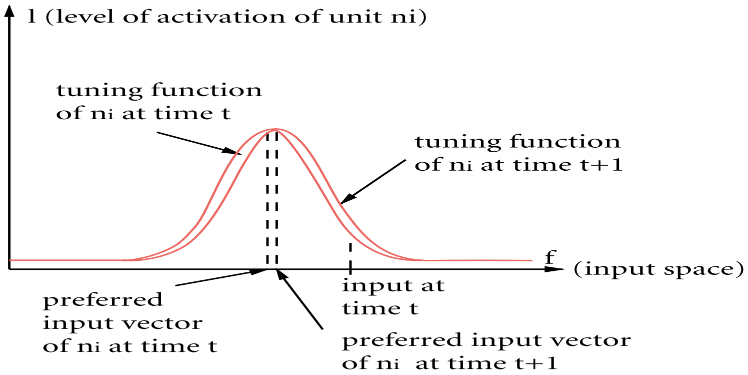

l (level of activation of unit ni)
tuning function of ni at time t
tuning function of ni at time t+1
f (input space)
input at time t
preferred input vector of ni at time t
preferred input vector of ni at time t+1

Figure 6.5: The updating of each unit when activated by a stimulus is such that the preferred vector is changed so that the unit responds a little more if the same stimulus is presented again immediately afterwards. This is a sensitization of the units, which is stronger when the neurons are highly activated, and weaker when they are less activated.

by another individual. The influence that hearing the vocalizations of others has on an individual's babbling is a collateral, statistical effect of the increase in sensitivity of the neurons, a quite generic mechanism of the low-level neural dynamics (Kandel et al., 2001).

Initial configuration of the neural units

The preferred vectors of the neural units are by default initially random with a uniform distribution in the basic version of the model that we are considering. A uniform distribution signifies that there are preferred vectors throughout the whole space at the same density everywhere. This means that at first the individuals produce vocalizations composed of articulatory targets distributed uniformly through the space. This in turn implies that at the outset the whole continuum of possible gestures is used (so there is no discreteness), and because there are many neurons portioned out in the whole space, the re-use of articulatory targets is very rare and due to chance (so there is no combinatoriality). Also, the overall activation of the neural map is of about the same amplitude whatever the stimulus.

This assumption will be modified in a "biased" version of the model, to be presented later, in which the initial distribution of the preferred vectors will no longer be uniform. A biased distribution is not symmetric: initially certain regions of organ relation space will contain more preferred vectors than other regions. Such a bias makes it possible to address constraints resulting from the function mapping articulatory configurations to sounds. A uniform distribution comes down to assuming that this function is linear and symmetric. A biased distribution takes into account the possible non-linearities.

In fact, looking at the human vocal tract, we see that there are relations between organs where a small change produces a small change in the sound, but there are also relations between organs where a small change produces a large change in the sound. If an architecture like that in Figure 6.2 had been used, in which the neural units of the network of relations between organs are activated by neurons of the perceptual network, this would mean that there are certain sounds which significantly activate neural units in the network of relations between organs whose preferred vectors are in a stretched-out region of organ relation space,

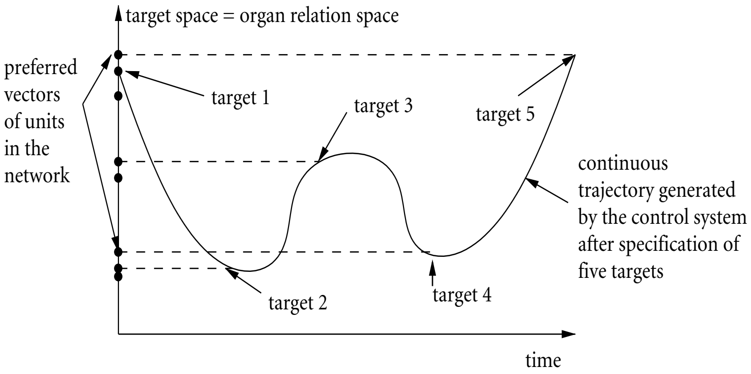

target space = organ relation space
preferred vectors of units in the network
target 1
target 3
target 5
continuous trajectory generated by the control system after specification of five targets
target 2
target 4
time

Figure 6.6: When an individual produces a vocalization, several articulatory targets are specified by random sequential activation of neural units. The preferred vectors of these neural units define the relations between organs to be reached at given times. These activations are commands called gestures in articulatory phonology. Next, a control system constructs a continuous trajectory which passes through all the articulatory targets.

and other sounds which only activate the neurons whose preferred vectors are in a confined region of organ relation space. This makes the learning rules of the neurons have different outcomes in different parts of the space. In some regions of the space the preferred vectors will change faster than in other regions. This leads to a non-uniformity in the distribution of preferred vectors, with more neurons in the regions where small articulatory changes yield small changes in sound, and fewer neurons in regions where small articulatory changes give large changes in sound. In the next chapter, where a part of the architecture in Figure 6.2 will be implemented, this bias will be introduced by using a realistic articulatory synthesizer throughout the simulation. For the present, and to aid comprehension, this bias is introduced from the start by warping directly the initial distributions of preferred vectors.

Tweaking the initial distribution, in particular using a uniform distribution, allows us to examine what outcomes are due or not due to the presence of non-linearities in the function mapping from sounds to articulatory configurations. In particular, we will show that neither discreteness nor combinatoriality require non-linearities in order to be explained, a rather original conclusion given the existing theories in the literature (see Chapter 4).

Interactions between individuals

The individuals move around in their world in a random fashion. At random moments, they produce a vocalization, which is heard by the nearby individuals, and also by themselves. The choice of how many individuals hear the vocalization produced by one of them does not affect the results: these are about the same whether one, two, three or more individuals are made to hear the vocalization. For the sake of generality, the simulations reported below only use one individual. From an algorithmic point of view, this is equivalent to randomly choosing two individuals from the population, and making one produce a vocalization which is heard and processed by both of them.

The individuals are therefore not playing a language game in the sense used in the literature (Hurford et al., 1998; Steels, 1997), and in particular they are not playing the

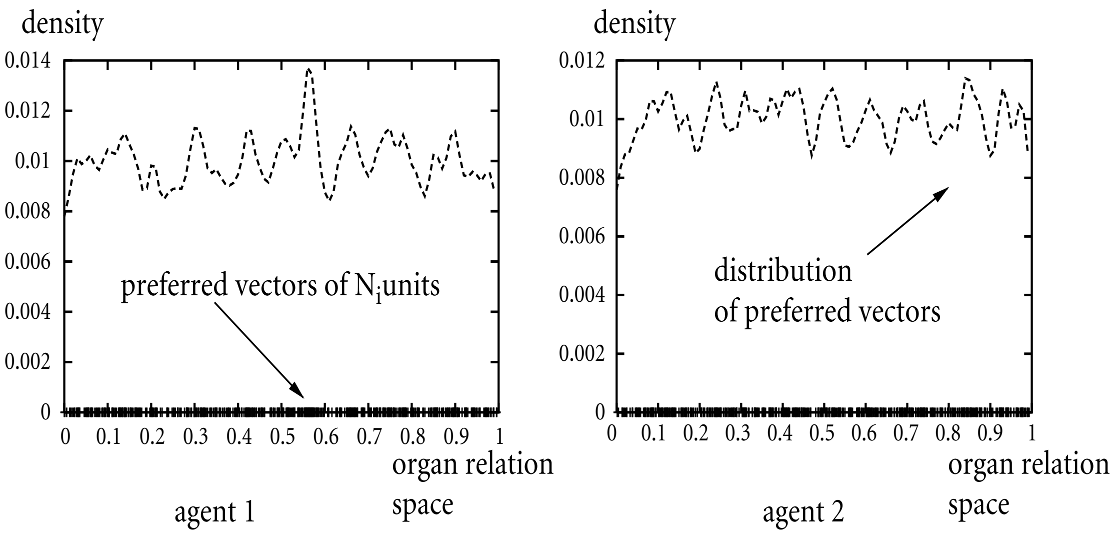


Figure 6.7: Initial distribution of preferred vectors of two individuals. The organ relation space, on the horizontal axis, is one-dimensional here. The individuals produce vocalizations specified by articulatory targets spread across the whole continuum (the vocalizations are holistic).

imitation game used some other models of the origin of speech (de Boer, 2001; Oudeyer, 2001b; Oudeyer, 2002c). Their interactions are not structured; there are no roles or coordination. They do not distinguish between their own vocalizations and those of others, and do not even possess representations of “others”. There is no linguistic communication, that it to say individuals never emit a signal with the intention of transferring some information which will modify the state of another individual.

6.2 Dynamics

6.2.1 Linear vocal apparatus

We will now describe what happens to a population of individuals who implement these assumptions. Organ relation space will here be one-dimensional, and the initial distribution will be uniform, which models the absence of morpho-perceptual constraints on the production and perception of sounds[10].

Figure 6.7 illustrates the distribution of preferred vectors of two individuals at the start of the simulation. The horizontal axis represents organ relation space (for example, a place of articulation or lip-rounding) with normalized values between 0 and 1, and the points shown inside it are the preferred vectors of the neural units of one individual. The vertical axis represents the density of preferred vectors, which makes it clearer how they are spread around, especially where many points are on top of each other. They are distributed approximately uniformly. As the learning rule of the neural units makes the individuals tend

[10] Here $\sigma = 0.001$ and there are 150 neural units and 10 individuals.

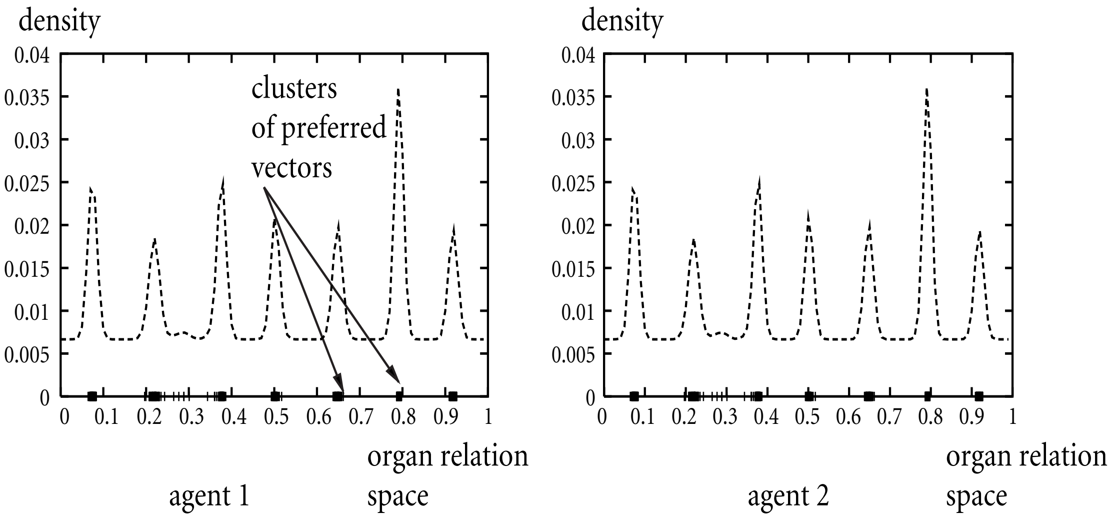


Figure 6.8: The distribution of preferred vectors of neural maps of the same two individuals as those in figure 6.7, after 2000 vocalizations: they are multi-modal, which means that the articulatory targets used are taken from among a small number of clusters, and moreover these modes are the same in both individuals (the speech code is shared and discrete). Due the the fact that there are few modes, they will automatically be systematically re-used to construct vocalizations that sequence several random selected targets (so the code is combinatorial).

to produce the same distribution of sounds as they hear around them, and since all the individuals produce roughly the same distribution of sounds initially, the initial situation is an equilibrium and is symmetric. It is a situation in which each neural map is in an initial state analogous to the initial state of the ferromagnetic plates described in Chapter 3 (with the difference that here we have a population of neural maps that interact with each other). Because of the stochasticity in the mechanism, there will be fluctuations. Studying the evolution of the distributions shows that the initial equilibrium is unstable: fluctuations change the state of the system. Figure 6.8 shows the distribution of preferred vectors of the same two individuals 2000 vocalizations later. It can be seen that now there are clusters (i.e. clearly discernible groups) and these clusters are the same for both individuals. The new distribution of their preferred vectors is multi-modal; symmetry has been broken. This means that the articulatory targets that they use to produce vocalizations are now taken from among one of the clusters (or *modes*, as they may also be called). The continuum of possible targets has been broken, and production of vocalizations is now discrete. In addition, the number of clusters appearing is small, which automatically results in the articulatory targets being systematically re-used to produce vocalizations, which have become combinatorial. All the individuals share the same speech code in the same simulation. By contrast, in two different simulations, the position and number of modes is different, and this is true even when the parameters used in the simulation remain the same. This is due to the inherent stochastic nature of the process. Figure 6.9 illustrates this diversity.

Evolution stabilizes during the simulation. To show this, the degree of cluster-formation

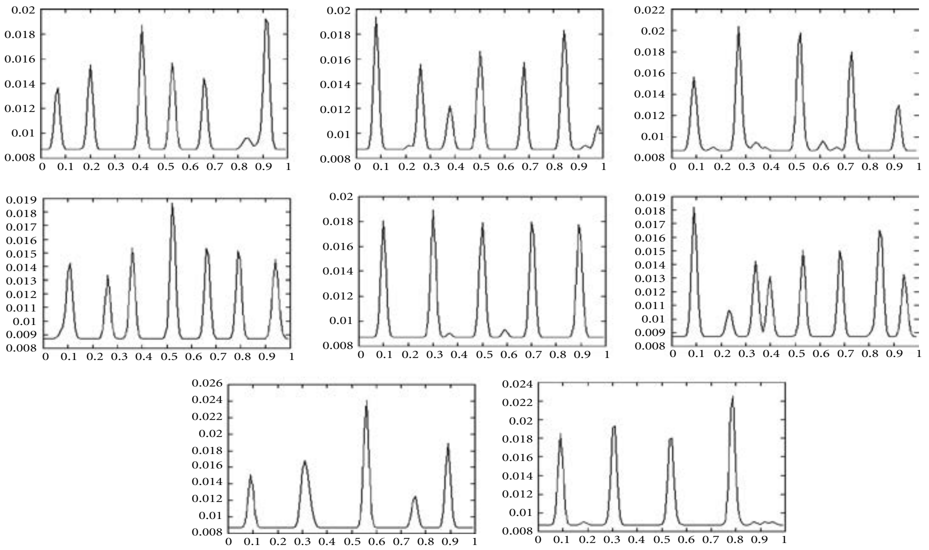


Figure 6.9: Several examples of final distributions of preferred vectors in different simulations. All these results are obtained using the same parameters. The stochasticity of the system makes possible the generation of phoneme inventories of different sizes, and their spatial arrangements are also different.

and the similarity among the distributions of preferred vectors of individuals were calculated at each time-step. This was done using the average entropy of the distributions and the Kullback-Leibler distance between two distributions (Duda et al., 2000)[11]. Figures 6.10 et

[11] At first a model of the distributions of preferred vectors in each neural map is made. The “fuzzy binning” technique (Duda et al., 2001) is used. This consists of approximating the distribution locally at a certain number of points spread out in the space to be modelled. Here, we take 100 points regularly spaced between 0 and 1. For each of these points v, an approximation of the local density of points is calculated with the formula:

$$p_v = \frac{1}{n}\sum_{i=1}^{150}\frac{1}{2\pi\sigma}e^{-\frac{||v-v_i||}{\sigma^2}}$$

where the v_i are the preferred vectors of the neurons of the neural map. σ is set so that the Gaussians have a width equivalent to 1/100. Once the distributions of the preferred vectors of the maps of all individuals have been modelled, the entropy of each one can be calculated (Duda et al., 2001). Entropy allows indirect evaluation of the degree of cluster-formation, or organization, of the points in a distribution. Entropy is maximal for a completely uniform distribution, and minimal for a distribution of points or vectors which all have the same value (when there is a single point cluster). Entropy is defined by the formula:

$$entropy = -\sum_{i=1}^{100} p_v ln(p_v)$$

Next, the average of all the entropies of all the distributions (one for each individual) is calculated, giving an evaluation of the average degree of cluster-formation across the maps of all the individuals. This is thus a measure of the degree of phonemic coding in the population of individuals. To evaluate the degree of similarity between two distributions p and q of the preferred vectors of each individual, the Kullback-Leibler distance function is used, defined as follows:

$$distance(p,q) = \frac{1}{2}\sum_v q_v log(\frac{q_v}{p_v}) + p_v log\frac{p_v}{q_v}$$

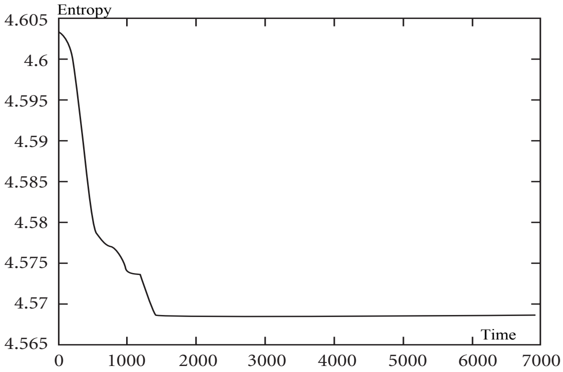


Figure 6.10: To evaluate the temporal evolution of distributions of preferred vectors of neural units, their average entropy was calculated at each time-step. It can be seen, in this example of a simulation run, that it decreases (this corresponds to the formation of clusters or modes), and then stabilizes (this corresponds to a converged state with several modes).

6.11 show how these two measures evolve over the course of a simulation involving 10 individuals. On the one hand, entropy decreases, and then stabilizes, which shows crystallization, or cluster-formation. On the other hand, the average distance between distributions of two individuals does not increase (initially, they already have the same uniform distribution!), and even decreases, showing that the modes that emerge are the same for all individuals.

The reason why there is crystallization is that because of the natural stochasticity of the mechanism, from time to time certain sounds are produced more often than others by the population of individuals (here again, a "sound" refers to a small slice of a vocalization). This creates deviations from the uniform distribution, which are sometimes amplified by the learning mechanism in a positive feedback loop. Then symmetry is broken. We see the same typical ingredients of self-organization phenomena described in Chapter 3.

However, to be precise, what we are here calling the state of crystallization in which several modes emerge is not yet the equilibrium state of the system. In fact, if it were possible to let the simulation run for an an extremely long time, in all cases, whatever the parameters, the outcome would be a single cluster, a single mode. This is because the adaptation rule for preferred vectors is the only driving pressure, at each time-step pushing the preferred vectors toward the vector corresponding to the stimulus. As the stimuli are generated from these distributions of preferred vectors, they are always inside the zone defined by all the preferred vectors of all the individuals in the community. And thus globally at each time step this zone shrinks (even if locally the neurons may distance themselves from each other because of the nonlinearity of the adaptive rule). But at the moment when different clusters are formed, the stimuli become concentrated right in the zones defined by the clusters. Imagine that there are two clusters C1 and C2. That means that stimuli are statistically very concentrated around the centres of C1 and C2. Now take a stimulus corresponding to the centre of C1. It will simultaneously make the preferred

In this way all the pairwise distances between the distributions of all the individuals is calculated, and the average is taken to determine at what point the individuals have (or do not have) a shared organization of their space of commands.

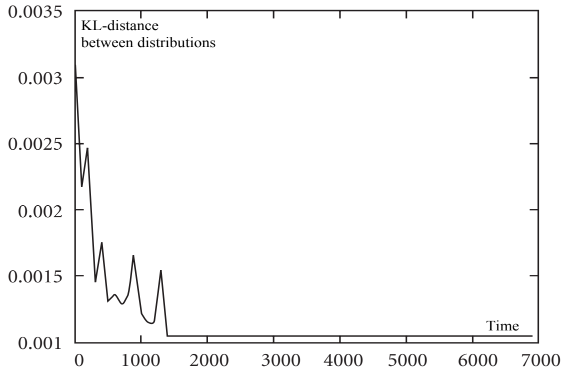


Figure 6.11: The average distance between distributions of preferred vectors of the neural units of each individual; we see that it stays the same, meaning that the modes of all the individuals are identical at the end of the simulation.

vectors of C1 and C2 move closer to the vector defining it. For those of C1, this will result in making them move even closer to the centre of C1, and will finish by hardly moving at all once they are there. For those of C2, there will be very little effect: the attractive force due to the Gaussian function is extremely weak at average and longer distances. If the default Gaussian function is used in our simulations[12], then if C1 and C2 are separated by a distance of 0.2, the displacement of the vectors of C2 toward C1 with each perception of a stimulus from C1 is $5.36 * 10^{-20}$. This means that it would need of the order of 10^{18} time-steps in the simulation to see the two clusters come completely together, which would take a simulation time on the computer longer than the age of the universe, and thus much greater than the lifetimes of the individuals in the model or indeed of any organism. This is why one can call the state in which several modes like C1 and C2 appear a "state of convergence". Thus clusters are formed which can quickly merge with each other, and when all the remaining clusters are far enough apart[13] we enter into a phase in which no more mergers occur for a very long time (this in fact never happens because the individuals do not live long enough). This phenomenon of self-organized pattern-formation during the passage of a dynamic system toward an equilibrium state is analogous to those discovered by Nicolis and Prigogine (Nicolis and Prigogine, 1977) in dissipative systems. Moreover, it may be noted that this kind of phenomenon was not studied or even conceptualized very early in the 20th century because the mathematical tools used by physicists only enabled them to calculate equilibrium states. It was computer simulations that made it possible to observe the behaviour of dynamic systems before they reached their equilibrium states and to discover that highly organized structures could be formed. In the same way, the study of the systems presented in this book necessitate computer simulations to investigate efficiently their dynamics.

Finally, if there is only one individual in a simulation, and it is allowed to produce and hear vocalizations, then its network of neural units will also self-organize into a state where several modes coexist. This means that there are two separable results: discreteness and

[12] that for which $\sigma^2 = 0.001$

[13] about 0.1 when $\sigma^2 = 0.001$

combinatoriality are explained by the coupling between production and perception with the neurons n_i, and can be obtained with just one individual; but when several individuals are brought together, then their repertoires of clusters, that is of commands and thus of gestures, converge (while if each developed its repertoire privately, this would be particular to each individual). Moreover, if one individual with a uniform neural map is put into a population of individuals which have already formed a speech code, this individual will learn that code. This means that the mechanism for learning a speech code is the same as that which enables a population to form one starting from nothing.

Robustness of the model

The artificial system has a certain number of parameters: the width of the neural tuning function (σ^2), the number of individuals, the number of neurons, and how many individuals hear a vocalization. Only σ^2 has a crucial influence on the dynamics. In fact, for example, the number of neurons changes nothing at all in the results when it is large enough, that is when it allows a sufficiently dense initial coverage of the space. Experimentally, the number has to be greater than 100 neurons to obtain the results we have presented. The influence of the number of individuals was also tested by doing simulations with between 1 and 50 individuals: convergence is obtained every time at the end of a number of interactions by each individual rising very little (between 150 and 500). The number of individuals hearing vocalizations pronounced by one of them also has very little influence. Thus, this section will focus on the study of the influence of σ^2.

This parameter was varied in a range of values from 0.000001 to 0.1 (using the values: 0.1, 0.05, 0.01, 0.005, 0.001, etc, 0.000001). Figure 6.12 shows some of the Gaussian functions associated with these parameters. It can be seen that all the relevant space is covered, that is that the possibilities go from the Gaussian with width equal to the width of the whole space down to the Gaussian with tiny width. For each of these parameter values, 10 simulations with 10 individuals were run, and the average entropy of the distributions of individuals was measured once convergence was reached (in the sense explained in the previous section). Entropy is a way of measuring the number of emerging modes (or the non-emergence of modes, when entropy is maximal). Figures 6.13 and 6.14 give the results. To represent σ^2, which varies over several powers of 10, a logarithmic scale (base 10) was used. For example, $\sigma^2 = 0.001 = 10^{-3}$ is represented by point 3 on the x axis in Figure 6.13. Three parts are distinguished, which could be called phases, as in the case of the ferromagnetic plaques discussed in chapter 3. The first phase involves all the values of σ^2 greater than 0.05: a single cluster forms. This is the maximal amount of order that can be obtained, corresponding to minimal entropy, and the greatest breaking of symmetry. At the other extremity of the space of values, when σ^2 is less than 10^{-5}, entropy is maximal: the initial symmetry corresponding to a random uniform distribution of preferred vectors is not broken. This state of maximal symmetry corresponds to a state of disorder. Figure 6.14 (bottom panel) gives an example of this. Between these two regions of possible values, there is a third phase corresponding to the formation of several well-defined clusters (between 2 and a dozen, beyond which there are no proper clusters and the individuals are not coordinated). It is in this region that the $\sigma^2 = 0.001$ default parameter value of the preceding sections is found. This organization of preferred vectors into well defined structures is a complex structure appearing at the boundary region between "order and chaos", as it is often simply labelled in the literature (Kauffman, 1996). The transitions between these three phases are analogous to those described for Bénard cells or ferromagnetic plaques in chapter 3. In

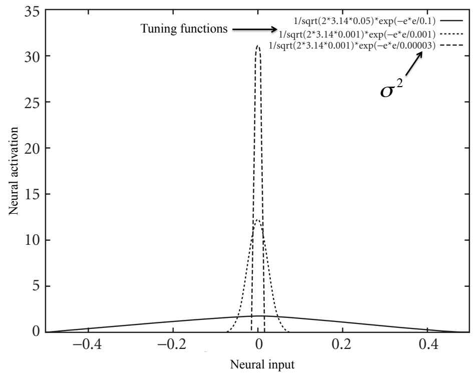


Figure 6.12: Some Gaussian functions corresponding to the different values of σ^2 used in this work.

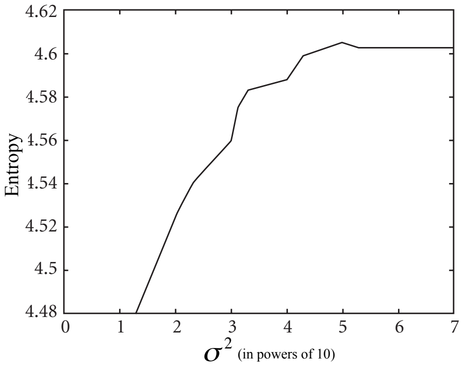


Figure 6.13: Variation in the average entropy of the distributions of preferred vectors produced by simulations using different values of σ^2. To represent σ^2 a scale of powers of 10 is used: for example $\sigma^2 = 0.001 = 10^{-3}$ is represented by point 3 on the x axis. Three phases can be seen: for low values, only one mode is formed; for high values the distribution stays random and uniform; for intermediate values several well-defined modes emerge.

addition, it should be noted that the region of parameter values in which repertoires shared between all the individuals are formed is very large: between 0.00001 and 0.01, that is a space covering several powers of 10! It can thus be said that the behaviour of the artificial system is very robust in the face of changes in parameter values.

6.2.2 Nonlinear vocal apparatus

In the preceding sections the initial distribution of preferred vectors was assumed to be roughly uniform. This meant that the function mapping a sound to articulatory configurations was linear, and thus took no account of constraints due to the physical nonlinearities of the vocal tract. This was interesting because it enabled us to show that no initial asymme-

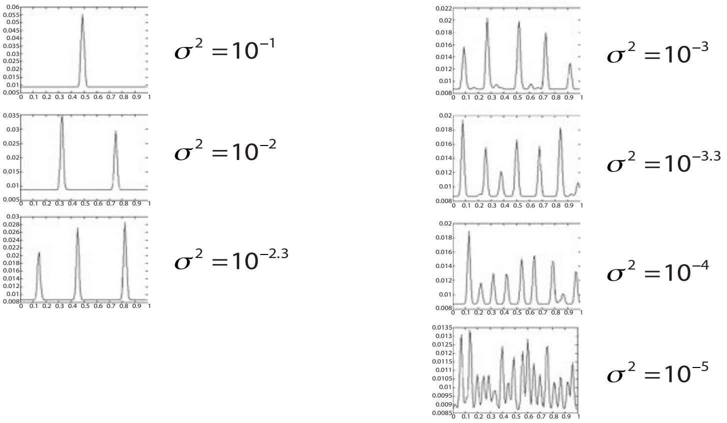


Figure 6.14: Examples of systems generated for different values of σ^2.

try is necessary to obtain discreteness (which is a property of asymmetry). In other words, this shows that there is no need to have discontinuities or non-linearities in the function mapping sounds to articulatory configurations to explain phonemic coding (this does not mean that the nonlinearities do not help, just that they are not necessary).

However, this function has a particular form in humans, which introduces a bias in speech sounds. We explained earlier that this bias could be modelled for the moment by manipulating the initial distribution of preferred vectors.

In this chapter we consider an abstract bias, which does not realistically reproduce the non-linearities of the human vocal tract, but whose generic nature will enable us to understand the consequences of the presence of non-linearities. We are still dealing here with the case of a one-dimensional organ relation space. The initial density of preferred vectors increases linearly between 0 and 1 (it was constant in the case of a uniform distribution). Figure 6.15 shows the initial distributions of preferred vectors of two individuals. It can

./figures/initbiased.pdf

Figure 6.15: Initial biased distribution of preferred vectors of neural units of two individuals. This models constraints due to the non-linearity of the function mapping sounds to articulatory configurations. It can be seen here that there are more preferred vectors in the second part of the space.

be seen that there are fewer neurons with preferred vectors close to 0 than neurons with preferred vectors close to 1. This naturally leads to a statistical preference for modes or clusters located in the second part of the space in relation to clusters situated in the first part of the space. Figure 6.16 shows the same two individuals' 2000 vocalizations later. The preference for modes in the second part of the space is however only statistical: it is possible for groups of individuals to develop a system with just as many modes in the first part of the space. Figure 6.17 gives some examples of the diversity of systems obtained. This phenomenon is crucial for understanding both the presence of statistical structural regularities in the phoneme inventories of human languages, and at the same time their great diversity. The following chapter will make a more detailed study of this aspect, using realistic constraints which will enable us to compare the results from the artificial system with the sounds of human languages.

6.3 Categorization and perceptual illusions

In the work so far, the individuals have had no mechanism for categorizing the articulatory targets that they use. That is, they were not able to collect similar targets in the same "bag". They had no way of seeing that when they activated two neurons from the same cluster (once clusters were formed), they were actually using this same cluster, and so the same phoneme. Thus in a certain way discreteness and combinatoriality were in the eye of the beholder, but the individuals had no knowledge of these properties. In the same way, when they perceived a vocalization and passed it through the temporal filter to decompose it into little slices

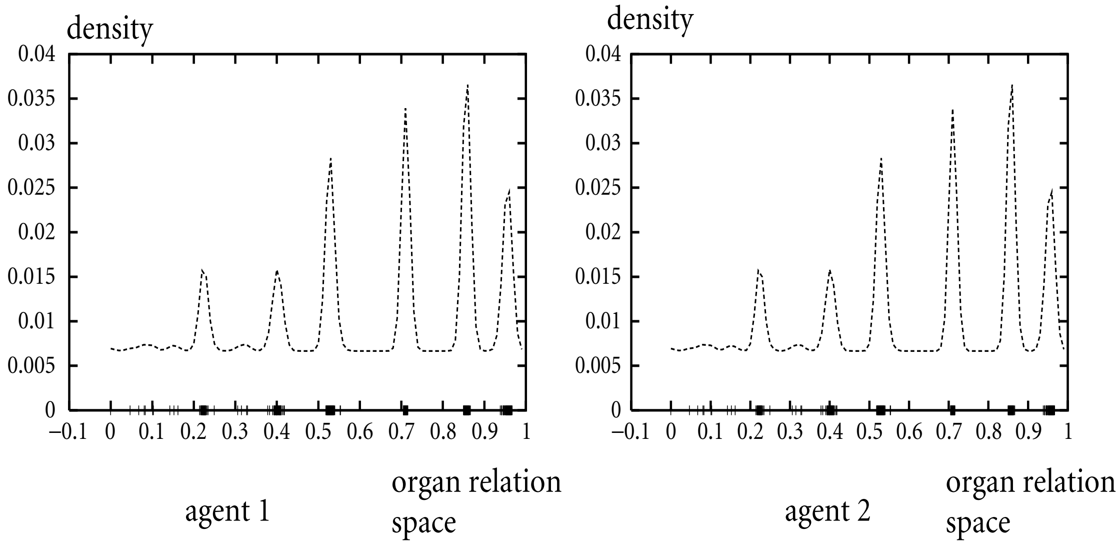


Figure 6.16: Distribution of the preferred vectors of the same two individuals after 2000 vocalizations. There is a preference for clusters, and therefore for articulatory targets, in the second part of the space.

thenceforth approximated by their averages, they had no way of organizing these sounds into different categories. This could have enabled them to retrieve the articulatory targets used to produce the vocalization. In this part of the work we will therefore extend the neural mechanisms so that the individuals are capable of categorization. We define here this capacity as follows: categorization comes down to getting the system into a stable state in which the activations of all the neural units remain fixed. Two stimuli will be categorized as the same if the state into which they move the system is the same, and different if this state is different. In the preceding sections, in a certain way, the individuals had neural maps which directly reached a stable state after the perception of a stimulus: one input activated all the neurons and this activation did not switch until a new input was provided. But in this way, two stimuli close to a cluster of preferred vectors, but slightly different, led to overall activity of the map similar, but not exactly identical. So we need to add a mechanism of relaxation that leads the neural units to precisely the same pattern of activations.

Neuroscientists studying the human brain have often used neural maps of this kind, that is to say variants of a model developed by Kohonen (Kohonen, 1982), to model cortical maps. Cortical maps, as their name might suggest, are capable of making models of their environment (there are auditory, visual, tactile maps, and so on), and their information is possibly used by other parts of the brain (Afflao and Graziano, 2006). These other parts of the brain use the stored information via a form of decoding that attempts to reconstruct the input stimulus which gave rise to the current set of neural activations. A discovery in neuroscience by Georgopoulos (Georgopoulos et al., 1988) makes it possible to set up a model of the way in which this decoding is carried out. The method uses the concept of "population vector". It involves the sum of all the preferred vectors of the units of the neural

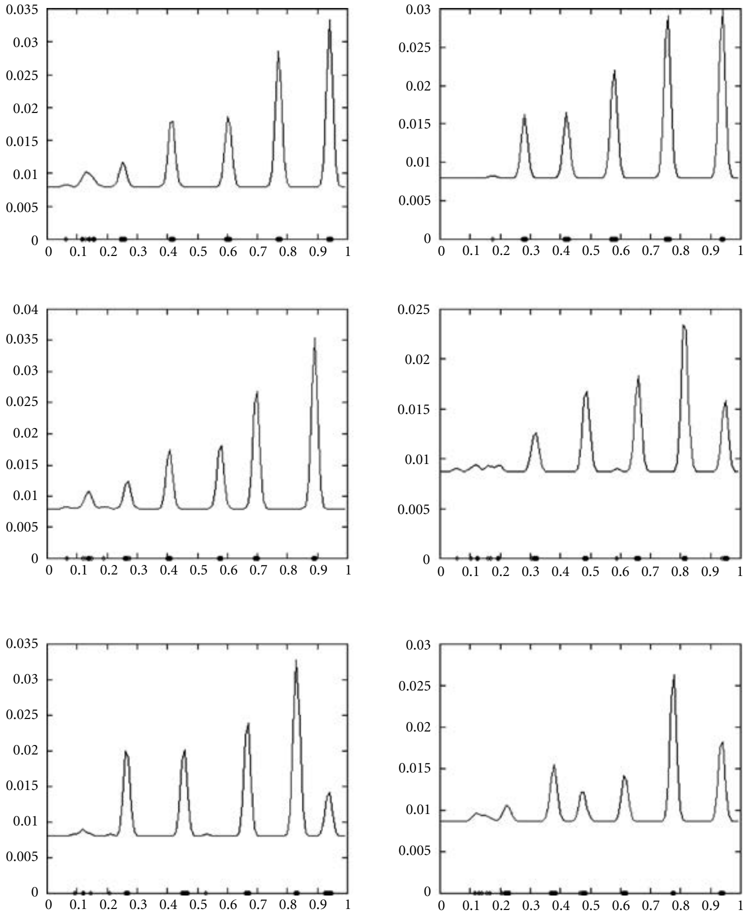


Figure 6.17: Some examples of systems obtained after 2000 vocalizations. The preference for clusters located in the second half of the space is only statistical: it sometimes happens that the first part of the space contains the majority of clusters.

map (weighted by their activation) and normalized by the sum of all activations[14]. When there are many neural units and their preferred vectors are uniformly distributed through the space, then this method reconstructs the input stimulus fairly precisely. However, if the distribution of preferred vectors is not uniform, some inexactness appears. Some researchers think that this inexactness is a defect of Georgopoulos' model, and have tried to put more precise formulas in place (Salinas et Abbot, 1994). However, this inexactness can also be

[14] If s is the stimulus, and N the number of neural units, then

$$pop(s) = \frac{\sum_{i=1}^{N} tune_{i,t} * v_i}{\sum_{i=1}^{N} tune_{i,t}}$$

is the population vector which recalculates s starting from the set of activations $tune_{i,t}$ of the neural units and of their preferred vectors v_i.

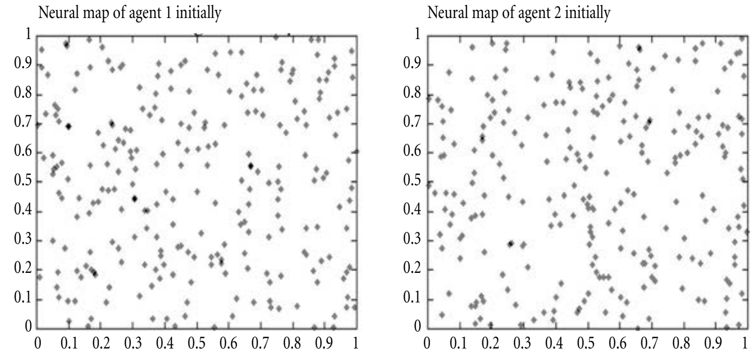


Figure 6.18: The preferred vectors of two individuals at the beginning of the simulation.

seen as allowing us to account for psychological phenomena like acoustic illusions. We will show that the perceptual magnet effect, studied by Kuhl and his colleagues (Kuhl et al., 1992) and described above in Chapter 2, can be explained by this inexactness. Then we will show how a simple feedback loop between the neural map and the decoding system gives us a categorization mechanism.

Let us briefly recall what the acoustic illusion known as the "perceptual magnet effect" consists of. Our perception of sounds is biased by our knowledge of the sounds of our own language. Kuhl's team showed that when people have to judge the similarity of pairs of vowels, they tend to perceive vowels as closer than they really are in an objective physical space when they are both close to the same prototype vowel in their language, and to perceive vowels as further away from each other than they really are in physical space when they are near to different prototypes. In brief, there is a sort of perceptual deformation that "attracts" sounds around each prototype in a language (from the point of view of a person's sensation of it). As a collateral effect, the differences between vowels of different categories grow. This is an instance of a wider psychological phenomenon called "categorical perception" and defined thus: "Categorical perception occurs when the continuous, variable and confusable stimulation that reaches the eyes and ears is sorted out by the mind into discrete, distinct categories whose members somehow come to resemble one another more than they resemble members of other categories" (Harnad, 1987). Evidently, as these illusions depend, in the case of sounds, on the sound prototypes in a given language, they are cultural phenomena.

The process of decoding with the population vector can be used to model the phenomenon of a hearer's sensation of a sound. It has already been used by Guenther and Gjaja (Guenther and Gjaja, 1996) to account for the perceptual magnet effect. In this model, the authors use a neural map similar to the one that we have presented, with the difference that they use a scalar product rather than a Gaussian function as the activation function of the neural units. The adaptation rule for preferred vectors of neurons is also similar to ours. But a large difference with the work presented here is that they get an individual to learn a sound system which already exists (in this case a vowel system). Their research does not centre on the origins of speech, and so they do not ask where the sound system they assume comes from.

To illustrate this process new simulations similar to those presented earlier are described

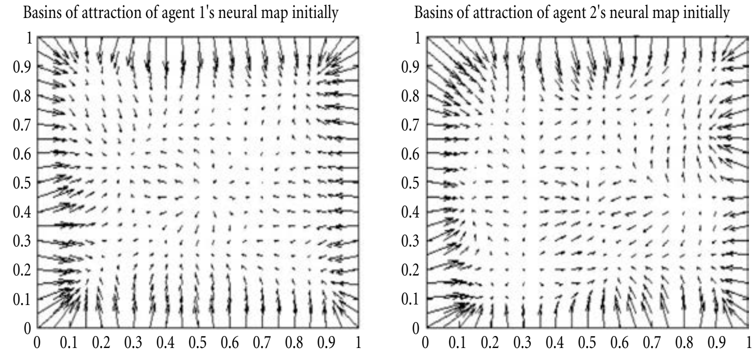


Figure 6.19: Representation of the way the individuals sense sounds at the beginning of the simulation: the start of each arrow corresponds to a stimulus activating the neural map, and the ends of these arrows correspond to the reconstructions of the stimuli after decoding of the activations of the neurons by the population vector. Here the perceptual warping, that is to say the difference between the stimulus and its internal reconstruction, is at first small.

here, but using a 2-dimensional organ relation space, which will enable us to visualize the perceptual magnet effect with the population vector decoding system. Figure 6.18 shows an example of the distributions of preferred vectors of the neural maps of two individuals at the beginning of a simulation. After 2000 vocalizations, the same two individuals are found to have neural maps like those presented in the upper panel of Figure 6.20. Clusters shared by the individuals are formed, but they are not very visible in the representation of preferred vectors because on the one hand most of the preferred vectors in the same cluster are represented by the same point (they have quasi-identical values) and on the other hand there are still a few isolated neurons themselves also represented by a single point. It is the lower panels, with arrows, that will let us see the distribution of preferred vectors that will be explained in the following paragraphs.

One can now evaluate the way in which these two individuals “sense” sounds at the beginning and end of the simulation. To do this, one can generate some artificial static sounds which serve as input stimuli. These stimuli are spread regularly throughout the space, according to a regular grid. For each of these stimuli, the activations of all the neural units are calculated, from which the population vector is calculated in its turn, giving a point in organ relation space, the result of the decoding. It is possible to represent the set of these stimuli and their decodings from activations of the neural maps using arrows: each arrow corresponds at its start to a stimulus and at its head to the point decoded by the population vector, that can be used to model the “sensation” of a sound experienced by an individual. Figure 6.19 thus represents the way in which the individuals “sense” sounds at the beginning of a simulation, and shows that the decoded points correspond fairly well to the stimuli (despite some tiny inexactitudes). Figure 6.20 represents the same two individuals’ way of sensing sounds 2000 vocalizations later (lower panel). The decoding is now no longer precise at all, and the perceptual warping – asymmetrically within the space of sounds – therefore substantial: the sensation of sounds by individuals is significantly different from the objective physical properties of the stimuli. After decoding the stimuli are shifted towards

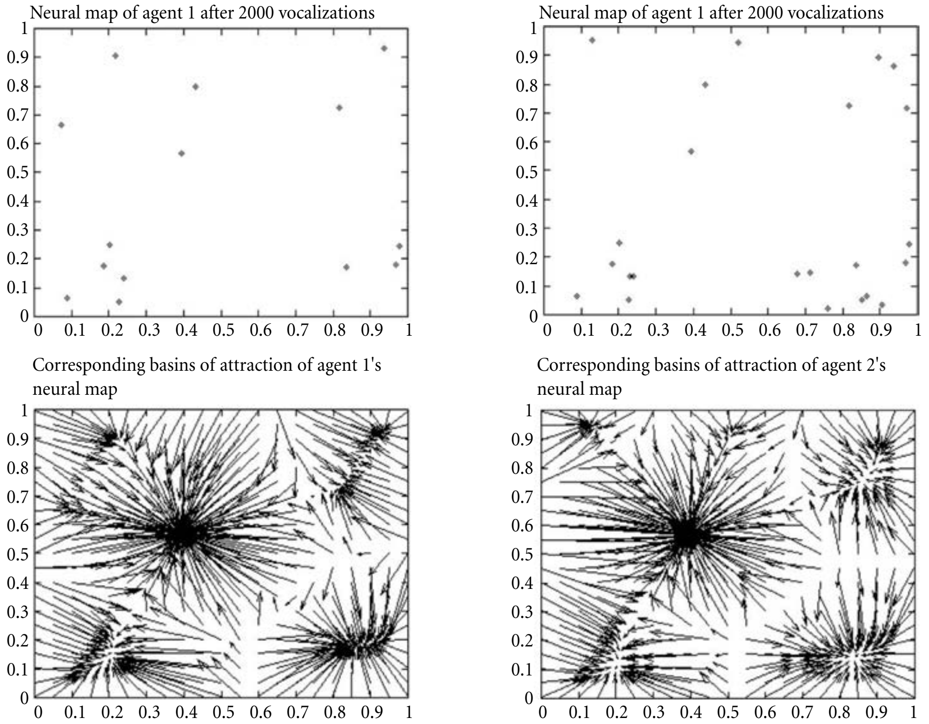


Figure 6.20: Example of the state of the neural maps of two individuals after 2000 vocalizations. The upper panels directly represent the preferred vectors. The lower panels represent the way in which the individuals sense sounds at this time. It can be seen that there is an organization of the space into regions in which stimuli are perceived as closer to the centres of these regions than they objectively are. These regions are shared by both individuals. They correspond to basins of attraction defined by categorization behaviour once a coding/decoding feedback loop is introduced into the neural map.

the closest cluster. This corresponds precisely to the perceptual magnet effect. In effect, the centres of clusters correspond to prototypes in their sound systems, and so the individuals behave in the same way as Kuhl's subjects (Kuhl et al., 1992).

If, moreover, if we treat Figure 6.20 as a surface viewed from above, whose slopes are represented by the arrows (the arrows pointing downwards), we see a landscape with valleys. If a marble is dropped in one of these valleys, it will roll to the bottom and stop at the same place, from whatever position it is dropped. If it is dropped in another valley, again it will roll downwards, but stop at a different valley bottom. This way of seeing Figure 6.20 is in fact the basis of an extension of the system enabling it to categorize sounds, in the sense set out above.

Since the point decoded by the population vector is specified in the same representational space as the input point, one can easily re-deploy this vector to the input layer of the neural map as a stimulus, but generated by the individual itself, as Figure 6.21 shows. This idea comes close to the re-entrance systems described by Edelman in his theory of human brain

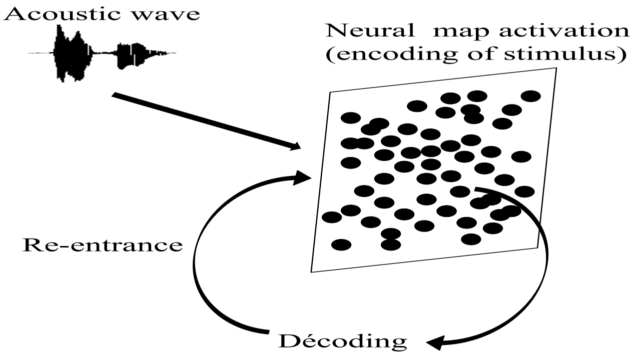


Figure 6.21: The categorization mechanism: after the stimulus has activated the neural map, the population vector is used to re-construct this stimulus and the result is fed back into the neural map. This process is repeated until the neural map activation is stabilized. The attractor corresponds to the recognized category of the stimulus.

functioning (Edelman, 1993). Indeed, Edelman describes the human brain not as a device in which information flows in one direction from the sensor to the control centres, but rather as a system in which the control system itself sends a lot of information toward the sensors, creating feedback loops. In the system, once the decoded point is re-deployed as an input, it reactivates all the neural units, and a new decoding is effected. Then the process is iterated.

This system is schematized in Figure 6.21. Geometrically, the sequence of positions of these successive points follows exactly the trajectory of the marble rolling down the valley: after several iterations the decoded point is the same as the point given as input to the neural map. The activations of the neural units stay the same with each iteration once this fixed point is reached. The system is then in a stable state, corresponding to a categorization of the stimulus given initially as input. To follow the trajectory of these successive points, simply start at the stimulus point and follow the arrows, for example in Figure 6.20. Each neural map defines a landscape of arrows, and thus of valleys and valley bottoms (in the language of dynamic systems these are called basins of attraction and attractors) which are peculiar to it. When the preferred vectors are distributed uniformly across the whole space, a certain number of valleys appear whose locations and shapes differ from one simulation to another. Figure 6.22 gives some examples. When clusters appear, then those which are big enough each define a valley and an attractor. Thanks to the smoothing property of the Gaussian activation function for neural units, the fact that there are neural units that do not belong to clusters (this results from the stochasticity of the mechanism), and which thus introduce a kind of “noise” into the landscape of neural units, does not modify the global landscape of the valleys of a set of clusters. Thus, for example, the individuals in Figure 6.20 do not have exactly the same neural maps, but share the same landscape of valleys, because they have the same clusters. They thus categorize sounds in the same way. This makes them able measure the discreteness of their speech system themselves. Each has a way of telling when two sounds are different and when two sounds are the same. They are now capable of segmenting the continuous trajectory of a vocalization into parts of which each little slice, the output of the temporal filter, is an input categorized in the same way.

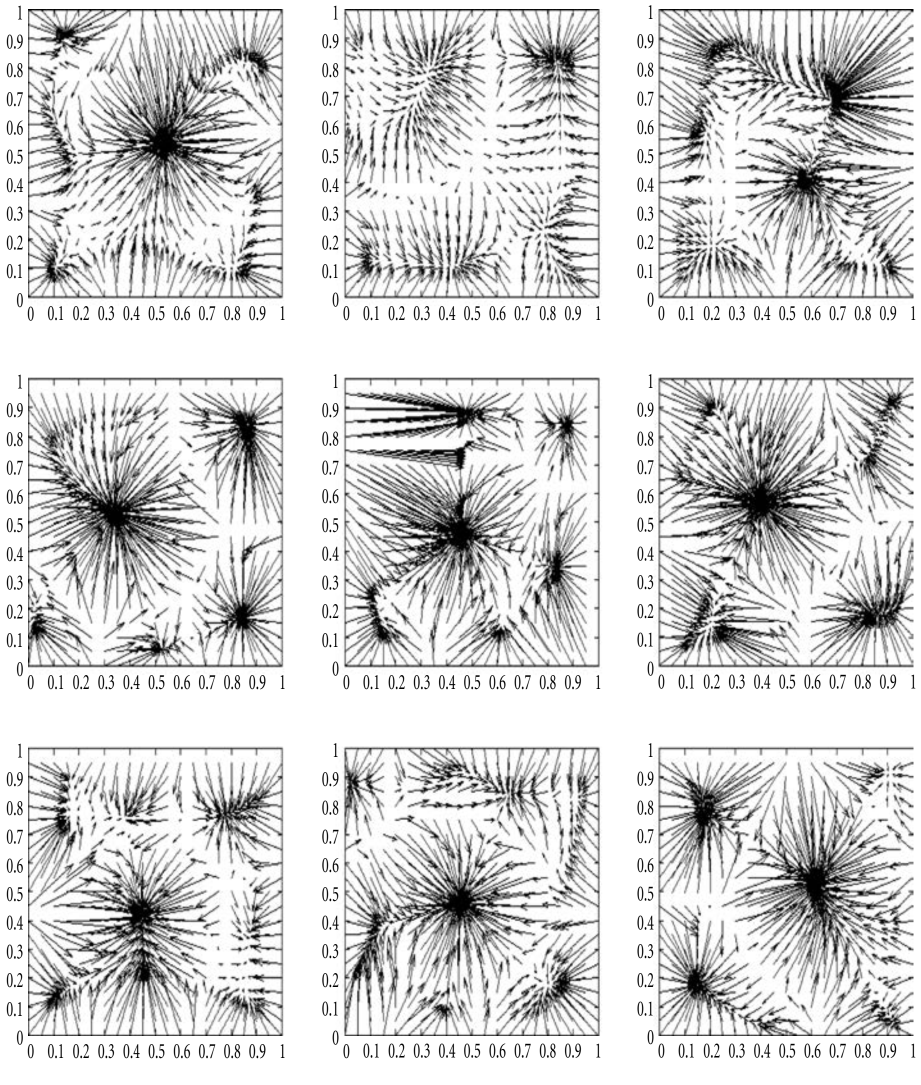


Figure 6.22: Examples of the systems of basins of attraction that can be obtained. It can be seen that their number, as well as their shapes and their locations, are varied.

Chapter 7

Learning Perceptuo-Motor Correspondences

In the previous chapter, it was assumed that individuals in the model knew the correspondences between the space of perceptual representations and the space of gestural and muscular representations. This allowed us to present a simple mechanism accounting for the self-organizational dynamics forming a discrete combinatorial speech code shared by a population of individuals. We have, however, suggested that the architectural components of the individuals would be generic and not specific to speech. It remains to show how the capacity to pass from one space to another can be realized by neural structures that are neither pre-wired nor specific to speech, in a generic learning process that is common to other faculties than speech. We will implement standard mechanisms of learning, based on the "Hebbian" reinforcement of neural connections, that have already been used for the learning of correspondences between hand and eye, between leg movement and body locomotion, and also between movements of the vocal tract and the sounds that are produced. However, here we will use these mechanisms in a novel context: rather than teaching an individual the perceptuo-motor correspondences in a system of vocalizations that exists already, the learning system will be influenced by the babblings of other individuals that initially have no structured repertoire of vocalizations. And a system of speech, shared and combinatorial, will nevertheless be seen to emerge in this situation.

To do this, we will modify these previous assumptions. While we assumed earlier that individuals were capable, given a sound stimulus, of retrieving the corresponding articulatory configuration, this assumption will no longer be made here. We also explained that there were two functions to learn: one function from perceptual space to organ relation space, and another mapping from relations between organs to the space of muscular activations. For the sake of simplicity, and because no effective and precise model is available of the physiological system that links relations between organs to muscular activations, we concentrate here on learning the function from sounds to relations between organs.

This means that two representations are used here, a perceptual representation and a representation of the relations between organs. The individuals no longer have one, but two neural maps, each coding a different space, as shown in Figure 7.1. The perceptual map is composed of neurons with properties identical to those described earlier, but these receive input values provided by a model of the ear. The motor network is composed of neurons taking as input activation values from the perceptual map. In addition, these motor

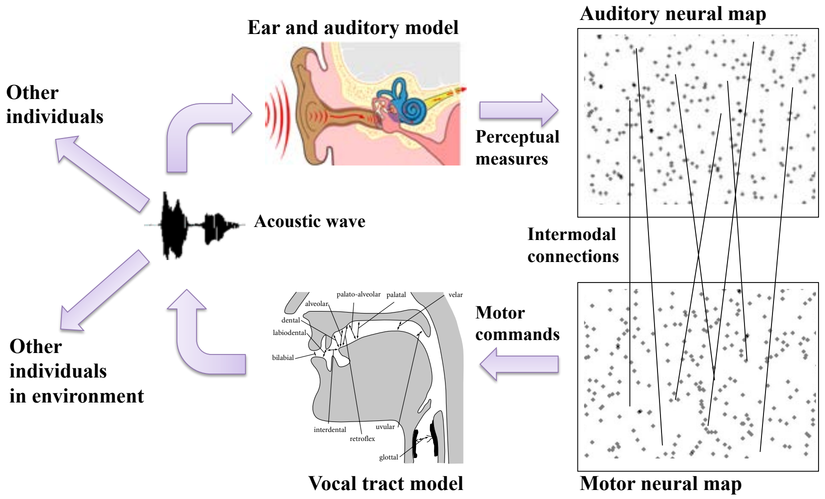


Figure 7.1: Architecture of individuals when learning the perceptuo-motor correspondences. They now are equipped with two neural networks: a perceptual network with input coming from a model of the cochlea, and a motor network whose output feeds into a model of the vocal tract. These two neural networks are interconnected via links that are arbitrary but plastic, allowing the individual to learn to pass from one representation to the other.

neurons have output connections destined to send commands to the vocal tract's control system. They have preferred vectors corresponding to these output signals: their value represents the relation between organs to be reached when they are activated at the same time as a “GO” signal is sent.

The two maps are fully connected; all the perceptual neurons are connected to all the motor neurons through connections which propagate activations from the perceptual map to the motor map. Each connection has a weight which initially has a random value close to zero. The preferred vectors of the perceptual neurons, representing the value that maximally activates each neuron, are also initially random, following a uniform distribution. The preferred vectors of the motor neurons, representing the relation between organs to be reached when they are activated, are also initially random and uniformly distributed.

To produce a vocalization, the mechanism is the same as before: neurons in the motor network are sequentially and randomly activated. The activation of a motor neuron fixes an articulatory target, which is a relation between organs to be reached. Then a control system operates to reach each articulatory target in sequence, thereby generating a continuous articulatory trajectory in organ relation space (this is still a polynomial interpolation). In contrast with the previous chapter, We here use operational models of the vocal tract and of the cochlea, which are fixed and the same for all individuals in a given simulation, and which map an acoustic representation to each configuration of the vocal tract. This acoustic representation will here be in terms of formants, as detailed below. The individuals in this version exchange acoustic trajectories, no longer directly exchanging trajectories in organ relation space, as in the earlier version.

The perception of a sound is also similar to the earlier version in algorithmic terms. The perceived acoustic trajectory is transformed into a perceptual trajectory by a model of the cochlea. This perceptual trajectory will here be composed in two dimensions: the first formant and the effective second formant, as explained below. This perceptual trajectory is then passed to the temporal filter, which cuts it into little slices, corresponding to the temporal resolution of the cochlea. Each little slice is averaged, giving a point value which serves as a stimulus to the individual's nervous system.

This stimulus first activates perceptual neurons. Their preferred vector is then modified just as in the earlier version: they are modified in such a way that they are activated a little more if the same stimulus is presented again immediately after, and this change is larger for more activated neurons[1].

Once the activation of the perceptual neurons has spread to the input connections of the motor neurons, two cases arise. The first case is where the motor neurons are already activated, because the vocalization was produced by the individual's own babblings, and here the connection weights are reinforced for connections between neurons whose activations are correlated, and weakened for connections between neurons whose activations are decorrelated (this is Hebbian learning)[2]. The second case is where the neurons in the motor map are not already activated. Here the activation of the neurons in the perceptual map (when a vocalization by another individual is perceived) is transmitted via the connections between the two maps and activates the motor neurons. Only a small amount of activation can spread via connections with low weights, whereas those with high rates transmit much more[3].

Once the motor neurons are activated, their preferred vectors are modified as follows. The preferred vector of the most activated neuron is taken as a reference, and the other preferred vectors are modified to bring them closer to this reference: the greater the activation of a neuron, the larger the modification. The same formula applies to the preferred vectors of perceptual neurons, the reference being equivalent to the perceptual stimulus.

When all the activities have been propagated and both the preferred vectors and the weights modified, the process of relaxation of the two neural maps begins so that individuals can categorize the vocalizations that they hear. This process is the same as that described above, applied in parallel to the two maps. The population vector is computed from an activation pattern of all the neurons in a map and then used as a new input, which produces a new activation pattern. The process continues until the activation pattern becomes stable.

[1] The mathematical formula for the new activation function is:

$$tune_{i,t+1}(s) = \frac{1}{\sqrt{2\pi}\sigma}.e^{|v_{i,t+1}-s|^2/\sigma^2}$$

$$v_{i,t+1} = v_{i,t} + 0.001.tune_{i,t}(s).(s - v_{i,t})$$

where s is the stimulus and $v_{i,t+1}$ the preferred vector of n_i after its activation by s.

[2] If i is a neuron in the perceptual map and j a neuron in the motor map, then the weight $w_{i,j}$ of the connection between these neurons changes according to the following formula:

$$\delta w_{i,j} = c_2.(tune_{i,s_i} - < tune_{i,s_i} >)(tune_{j,s_j} - < tune_{j,s_j} >)$$

where s_i and s_j are the inputs for neurons i and j, $< tune_{i,s_i} >$ the mean activation of i during a time interval T, and c_2 a small constant (here 0.01).

[3] The activation of the motor neurons is computed by the following formula:

$$tune_{i,t}(s) = \frac{1}{\sqrt{2\pi}\sigma} * e^{-\frac{1}{2}(v_i.s)^2/\sigma^2}$$

where s is the activity vector for the neurons in the perceptual map, and v_i the connection weight vector upstream of neuron i in the motor map.

This corresponds to a behaviour of categorization.

In this new architecture, the crucial feature of the coupling of production and perception is retained: the distribution of preferred vectors of perceptual neurons evolves in parallel with the distribution of preferred vectors of motor neurons. If at a given moment all the sounds associated with articulatory configurations coded by the preferred vectors of motor neurons are produced, and these sounds are transformed by the bias of the model of the vocal tract and of the ear into a perceptual representation, then the distribution of points obtained is roughly the same as the distribution of preferred vectors in the perceptual neural map. Conversely, if the distribution of sounds that an individual hears changes, this will affect both the distribution of preferred vectors in their perceptual map and in the motor map. This coupling between the motor map and the perceptual map has again an important dynamical consequence: the individuals will tend to produce more vocalizations composed of sounds that they have already heard. Said another way, when a vocalization is perceived by an individual, this increases the probability that the sounds that compose this vocalization will be re-used by the individual in its future vocalizations. Initially, as in the previous chapter, the preferred vectors are all random and uniformly distributed. This means that the vocalizations produced are specified by articulatory targets spread uniformly throughout the continuous space of possible targets. The space is thus not yet discretized; there is no phonemic coding. In addition, the initial situation is in an equilibrium state as all the individuals produce sounds composed of targets following the same distribution and adapt themselves to approximate the distribution of sounds that they hear. We will see, however, that here again random fluctuations will break this symmetry and push the individuals away from this equilibrium towards another organized state.

Using two neural maps not only illustrates how the articulatory function can be learned, but also and above all accounts more realistically than the previous chapter for the constraints due to the non-linearities of this articulatory function. Indeed, the earlier version of the model used an abstract articulatory synthesizer to generate the initial distribution of preferred vectors. Here an articulatory synthesizer will be used throughout the whole simulation and the bias will be directly applied by the synthesizer to the neural maps. Recall at first that there are articulatory configurations for which small changes produce small changes in the sound produced, as well as articulatory configurations for which small changes produce large acoustic changes. When the neurons in the motor network have random uniformly distributed preferred vectors, this distribution will be quickly biased: the non-linearities cause the adaptation of neurons to happen in a heterogenous way. For certain stimuli, many neurons will have their preferred vectors substantially modified, while for other stimuli, only a few neurons will have their preferred vectors modified substantially. This leads rapidly to non-uniformities in the distribution of the preferred vectors of the motor neurons, with more neurons in the regions where small articulatory changes produce small acoustic changes than in the regions where small articulatory changes produce large acoustic changes. Consequently, the distribution of the articulatory targets of vocalizations also becomes biased, and learning by the neurons in the perceptual network results in the preferred vectors of these neurons also being biased.

7.1 Articulatory synthesizer

This version of the system uses a realistic model of part of the function mapping sounds to relations between organs, and of the function mapping percepts to sounds. This model corresponds to the subsystem of the human vocal tract that enables us to produce vowels. It was developed by de Boer (de Boer, 2001).

The model uses a 3-dimensional articulatory space, each dimension representing a relation between organs: tongue position (place of articulation), tongue height (manner) and lip-rounding. The position of each articulator has values between 0 and 1, and a triplet of values $ar_i = (r, h, p)$ defines an articulatory configuration. From each point in this space, following de Boer's model, the first four formants corresponding to the vowel sound produced can be calculated: these are the frequencies of peaks in the energy spectrum (or poles in the function transforming articulatory configurations into acoustic waves). This calculation was modelled by a polynomial function generated by interpolation between points of a database (Vallee, 1994) representing vowels with their articulatory configurations and the associated formants (de Boer, 2001).

Next, we use a model of the cochlea summarizing the information that it sends to the brain when vowel-like sounds are heard. As in the previous chapter, vocalizations are still complex and dynamic. One uses again the temporal resolution filter which splits the continuous acoustic four-dimensional $(F_1, ..., F_4)$ trajectories into a sequence of points which are then fed into the cochlea model which then computes a perceptual representation. This model, used by Boë and his colleagues and by de Boer (Boë et al., 1995; de Boer, 2001), calculates a 2-dimensional representation from the first four formants. The first dimension is the first formant, and the second dimension is called the effective second formant and is a non-linear combination of formants F_2, F_3 and F_4 (de Boer, 2001)[4].

7.2 Dynamics

The simulations described here have the same structure as the earlier simulations: a dozen individuals wander randomly in a virtual space, and from time to time produce a vocalization, which is heard by themselves and by the nearest individual[5].

To visualize the state of an individual's neural system, the representation of that individual's perceptual neural maps can be used. In effect, these maps are of two dimensions (first formant and effective second formant) and contain after convergence the same information as the motor networks regarding distributions. An individual's perceptual network is represented in two ways: by showing all the preferred vectors and by showing the dynamic categorization landscape linked to the coding/decoding cycle by the population vector.

Figure 7.2 shows the perceptual maps of two individuals after 200 interactions. This allows us to visualize the bias due to the articulatory synthesizer. The unit of measurement is the Bark: the horizontal axis represents the first formant, and the vertical axis the effective second formant. The distribution of preferred vectors is no longer uniform. It is contained within a triangle (the vowel triangle), which is itself covered non-uniformly. The initial situation is an equilibrium, since all the individuals have the same uniform distribution of preferred vectors, but this distributional bias is now added to the natural fluctuations in the system to create non-uniformities which are amplified by a positive feedback loop. The individuals, as in the previous chapter, "crystallize" in a new situation where the preferred vectors of their neurons are grouped in clusters, which define phonemic categories. Figure

[4] The formants are also expressed in Barks, which are approximately a logarithmic transformation of measures in Hertz. These formulae model the fact that the human ear cannot discriminate narrow band frequency peaks in the high frequencies (Carlson et al., 1970).

[5] Each of their neural maps has 500 neurons, and $\sigma = 0.15$ (this is the width of the Gaussian defining their activation function, and is equivalent to 15 percent of the extent of each dimension).

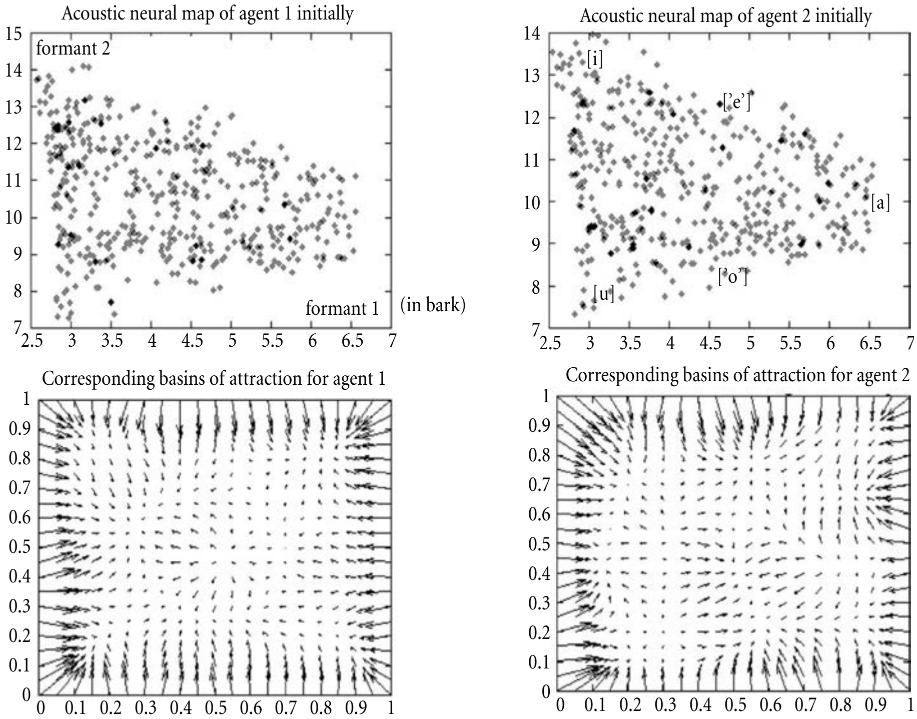


Figure 7.2: The perceptual neural maps of two individuals 200 vocalizations after the start. The upper panels show the preferred vectors and the lower panels show the basins of attraction that they define. It can be seen that the constraints due to the use of an articulatory synthesizer very quickly bias the initial distribution of preferred vectors: they are spread non-uniformly in a triangle (the vowel triangle).

7.3 shows the two individuals of the previous figure 2000 vocalizations later. Figure 7.7 shows the evolution of the similarity between the preferred vectors of the 10 individuals compared pairwise. This is roughly constant, which shows that they all have the same distributions of preferred vectors (as each other) over the course of the simulation, and in particular the same clusters after 2000 interactions. The fact that the distributions converge also indicates that they have learned to master the articulatory function.

Just as in the previous version, each simulation produces a unique system of phonemic categories (here vowels). Figures 7.4, 7.5 and 7.6 give other detailed examples of the systems that can be obtained in the simulations. Figure 7.8 gives further examples of configurations of the basins of attraction of the acoustic maps obtained in different populations of individuals. Now, as also explained, statistical regularities characterizing the phoneme inventories can appear at the same time as this diversity exists. As a realistic articulatory vowel synthesizer is used, and as precise databases of statistical regularities are available characterizing the vowel inventories of human languages, it becomes possible to a make comparisons between the vowel systems generated by the model and those of humans.

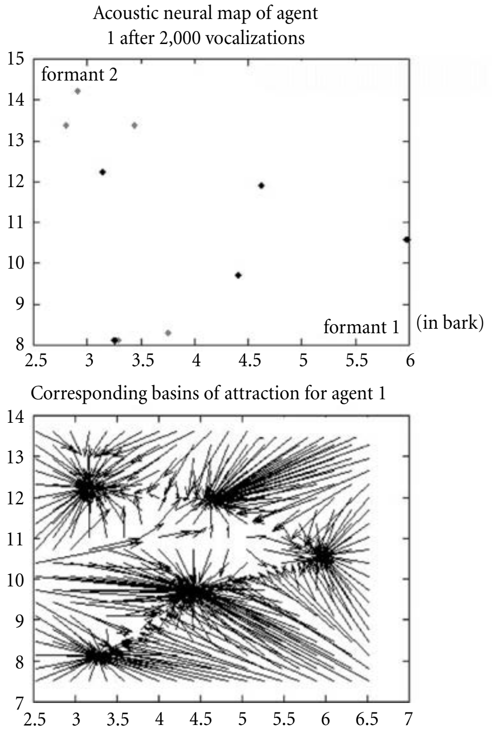


Figure 7.3: The neural maps of the two individuals of the previous figure 2000 vocalizations later. It can be seen that shared clusters have formed in both individuals, defining a particular categorization system shared by both individuals. The system shown in this figure is the system most often obtained in the simulations. It is also the most frequent in the languages of the world. It is the five-vowels /i,u,e,o,a/ system.

The database of human languages which is used is $UPSID_{317}$ (UCLA Phonological Segment Inventory Database), developed by Maddieson (Maddieson, 1984). It contains 317 vowel systems, belonging to 20 different language families, chosen for representativeness in terms of geography and population genetics (see Figure 7.9). Each vowel system consists of a list of vocalic segments said to be "representative". In fact, in a single language, and even in a single person, a vowel can be pronounced differently depending on the preceding or following phonemes or on the rhythm of the utterance (these are co-articulation phenomena). The set of pronunciations of a phoneme is called the set of allophones of this phoneme. To represent a phoneme by a single point in UPSID, the most frequent allophone is chosen. Moreover, the segments of this database are not directly used: instead, regularities are identified according to the groupings made by Schwartz, Boë, Vallée and Abry (Schwartz et al.,1997b). In fact, no two languages exist with exactly the same vowel prototypes (for example, two languages can have slightly different ways of pronouncing [e]). The method used, which turned out to be efficient for setting out regularities, is similar to that used by Crothers (Crothers, 1978). It consists in grouping vowel systems according to the relative positions of vowels

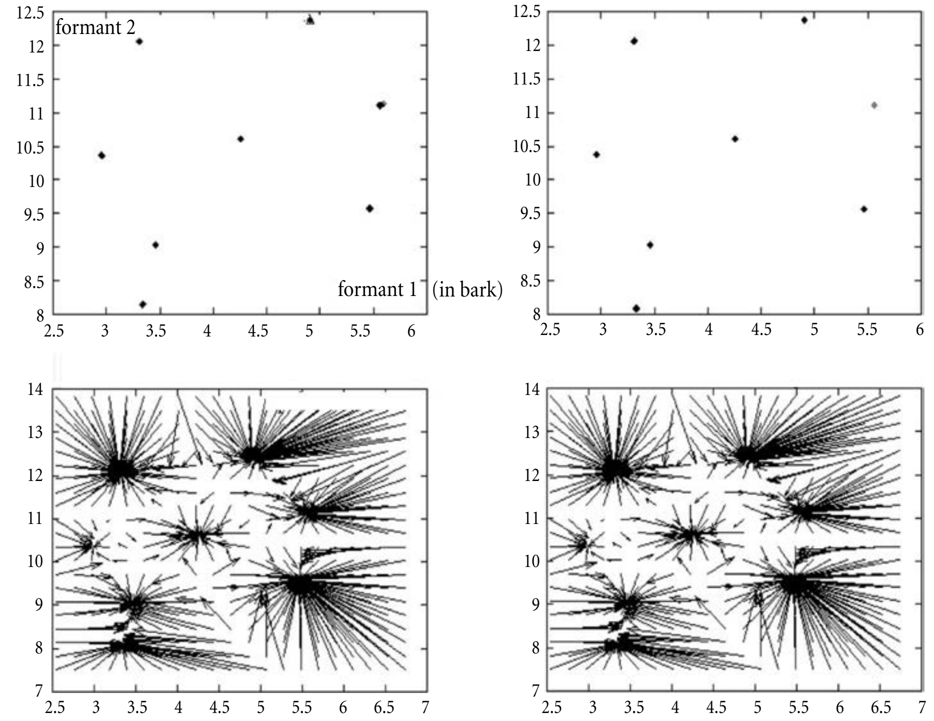


Figure 7.4: Another example of a system that can be obtained. This is a six-vowel system.

with respect to each other rather than according to their absolute positions. Also, vowels are represented in the acoustic space $(F_1, F_2^{'})$, in the same way as they are represented in the perceptual maps in the simulations. Figure 7.10 shows the set of possible patterns used in this classification: black circles on the vowel triangle represent possible phoneme locations. There are 12 possible locations, which makes more than 4 million possible vowel systems of less than 8 vowels. Since we are interested in the relative positions of phonemes, small shifts, rotations and scalings are permissible in the process of matching particular vowel systems to generic patterns of vowel systems. Another abstraction from the UPSID database used as a basis of comparison is the number of systems with 2 vowels, 3 vowels, 4 vowels, etc. This measure is inherently much more direct than going by the structures of the vowel systems.

Five hundred simulations were run, all with the same parameters. Each time, the number of vowels in the inventory and the relative positions of the vowels were recorded. This second measure was done by hand, which appears to be the most precise method, also used by de Boer (de Boer, 2001) to compare the vowel systems generated by his simulations with those of human languages. Here, to determine the number and location of vowels, the representation of the attractor landscape, which models individuals' categorization behaviour, is used. This is more efficient than looking at the distribution of preferred vectors in the perceptual maps, because clusters are sometimes difficult to visualize due to rare neurons with preferred vectors with values far from these clusters. In addition, the categorization representation

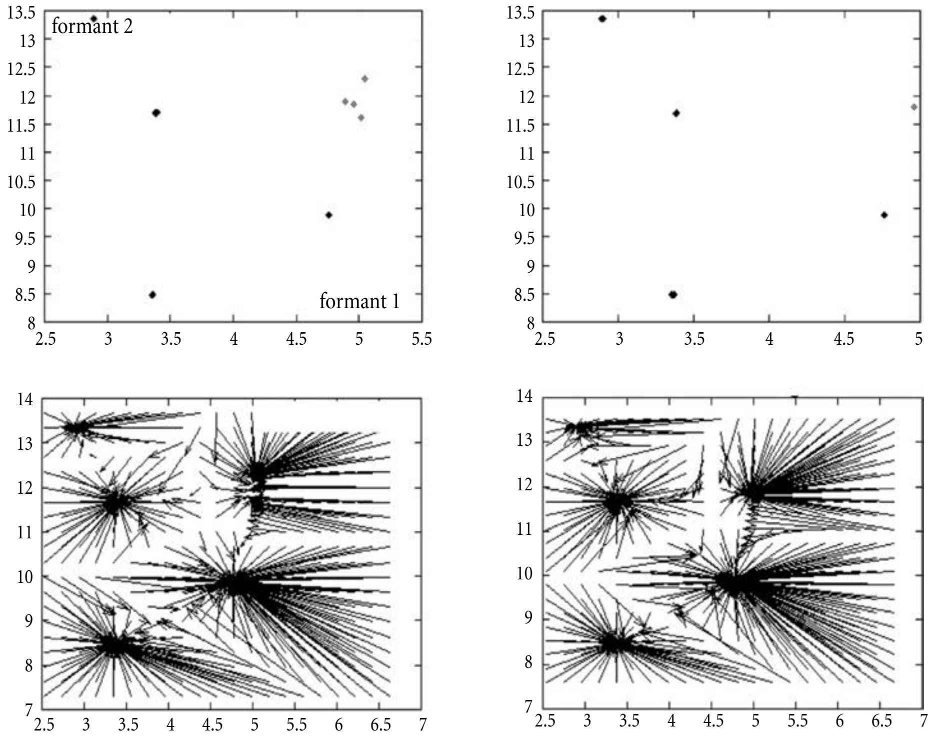


Figure 7.5: Another example vowel system.

directly gives a prototype for each category, the attractor point of the coding/decoding cycle of the population vector.

The results are given in Figures 7.11 and 7.12. Figure 7.11 shows that the sizes of vowel systems in human languages and those of the populations of individuals are very similar. In particular, they are characterized by a maximum of 5 vowels. Figure 7.12 shows the distribution of vowel system structures in human languages and in the systems generated by the individuals. Only the two most frequent systems are represented in each case, up to 8 vowels. Despite the enormous space of possible systems, the two most frequent systems are the same for n=3,6, and the two most frequent systems in humans are in the three most frequent artificial systems for n=4,5,7. If we consider percentages, we also see a degree of correspondence. The prediction for the most frequent vowel system /i,u,e,o,a/ is 25 percent, as against 28 percent in UPSID. The human systems represented in the figure correspond to 59.5 percent of the systems in UPSID, and the artificial systems in the figure are 75.6 percent of the total of artificial systems generated. This shows that the relationship between frequent systems and more eccentric ones is roughly followed. It also shows at the same time the diversity of the systems generated. By contrast, the predictions of the simulations deteriorate for systems of more than 7 vowels, because these are not generated at all, while they do exist in humans (but are certainly less frequent).

Given the enormous number of possible vowel systems, the similarities between the

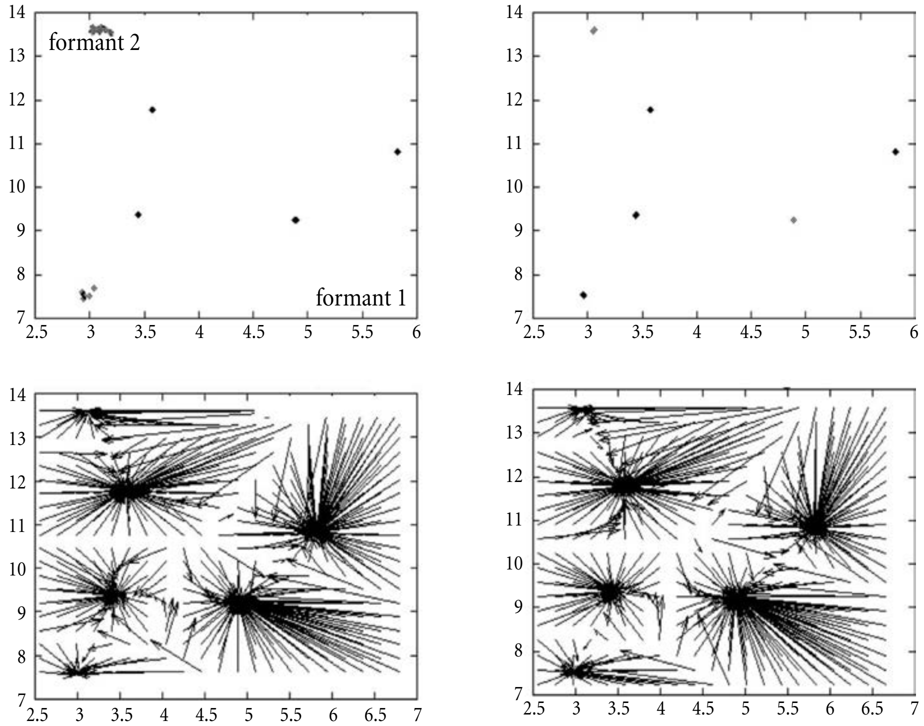


Figure 7.6: Another example vowel system.

human and artificial systems are striking. Moreover, the diversity of the systems is obtained without any changes in parameters between the simulations.

The differences between the predictions and the human systems should be interpreted in the light of the way the UPSID systems and the artificial systems were constructed. In fact, as explained in previous chapters, the models we discuss deal with the formation of a speech code before it was recruited for purposes of communication. That is to say that the generated systems are pre-linguistic systems which have not yet undergone cultural evolution under the functional pressure of linguistic communication. By contrast, the UPSID systems are vowel systems of contemporary languages. They are vowel systems which have evolved under functional pressure for effective communication regarding a large and open set of referents, and this evolution has already been going on for a very long time. To summarize, the simulations we have described are based on mechanisms modelling vocal exploration processes that are prior to the mechanisms that gave rise to the modern languages in UPSID (the first kind of mechanism operated at a time before complex communication appeared, whereas the second operated long after the dawn of language). The similarities and the differences between the structures of the two systems are therefore interesting, but their significance should not be exaggerated.

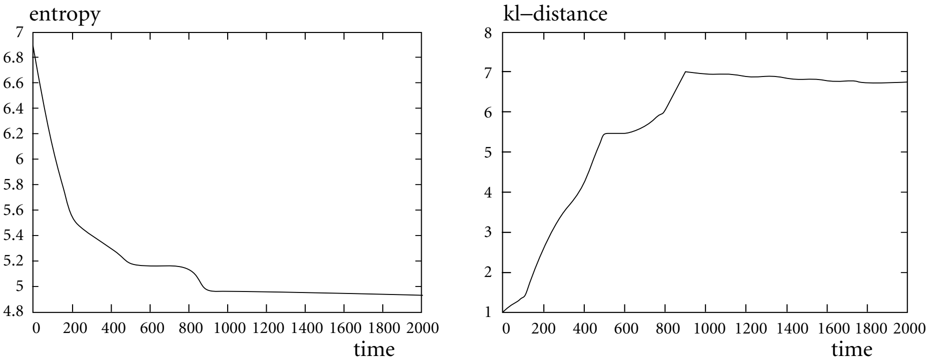


Figure 7.7: Evolution of average entropy of individuals' distributions of preferred vectors in one simulation, and evolution of the average kl-distance between individuals' distributions compared pairwise. The entropy curve shows the formation of clusters. The kl-distance curve shows that all individuals' clusters are the same (it rises a little, but this variation is negligible, given that the kl-distance between two random distributions of five clusters is of the order of 10^5).

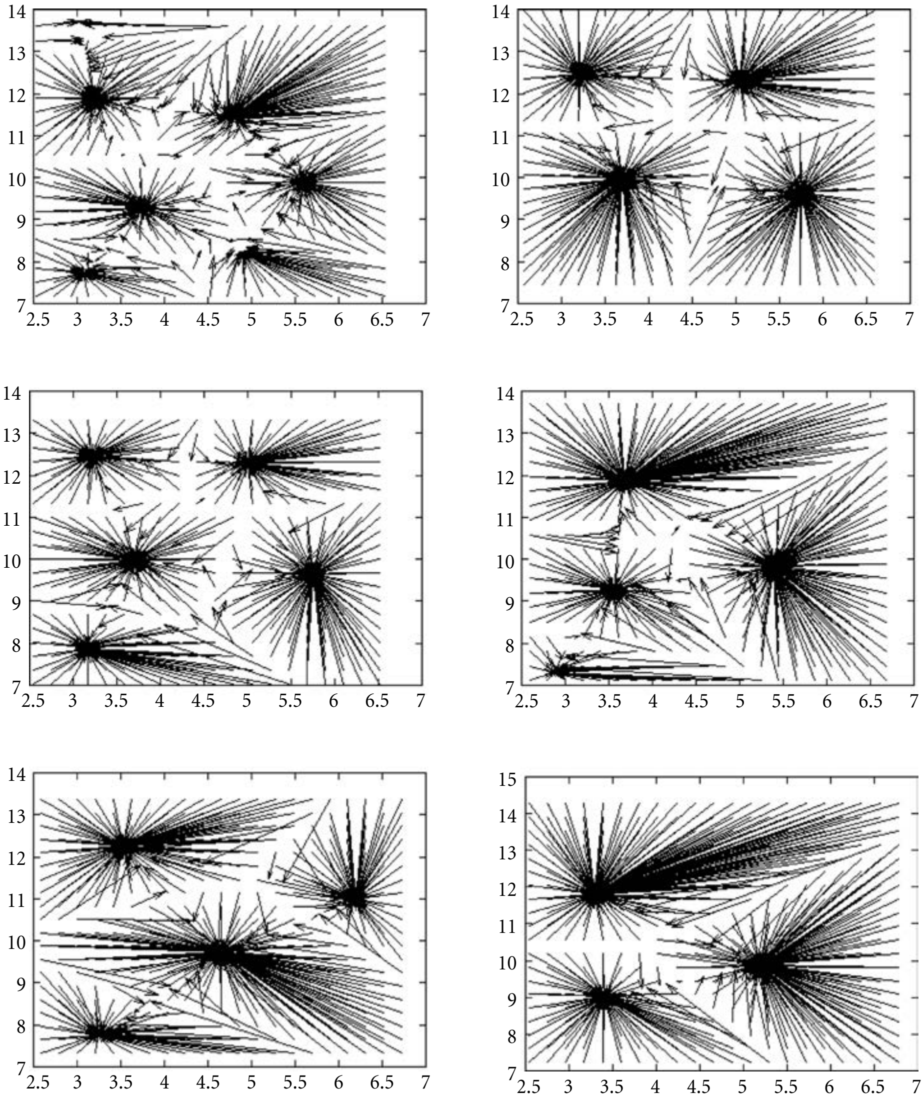


Figure 7.8: Further examples of vowel systems that can be obtained. It can be seen that both the number and the shapes of the basins of attraction are varied, although the simulation parameters remain unchanged.

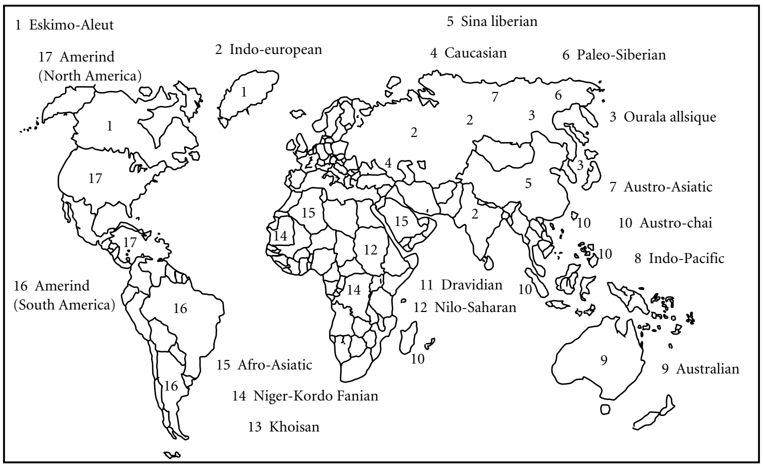


Figure 7.9: The UPSID database contains 317 vowel systems belonging to 20 different language families. Adapted from (Escudier and Schwartz, 2000).

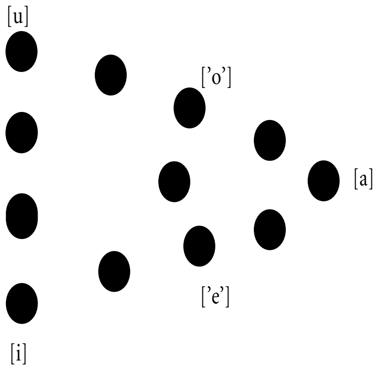


Figure 7.10: The patterns with which each generated system has been identified are combinations of the locations shown in this figure. There are 12 locations, making the number of systems with less than 8 vowels exceed 4 million. Note that here the same notation as Schwartz, Boë, Vallée and Abry (Schwartz et al., 1997) are used, with the horizontal axis corresponding to the first formant and the vertical axis to the effective second formant, but with the high values at the bottom and the lower values at the top (the axis switches direction in relation to the figures given earlier).

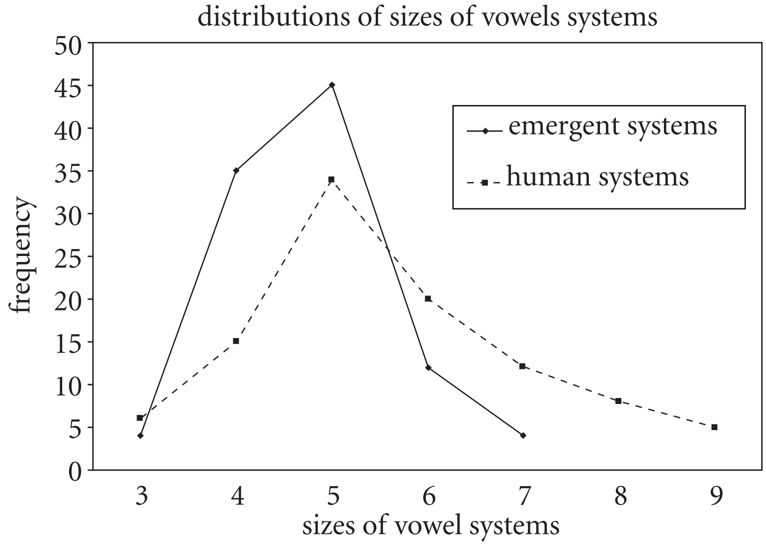


Figure 7.11: Distributions of sizes of vowel systems obtained in the simulations and in the languages of the UPSID database.

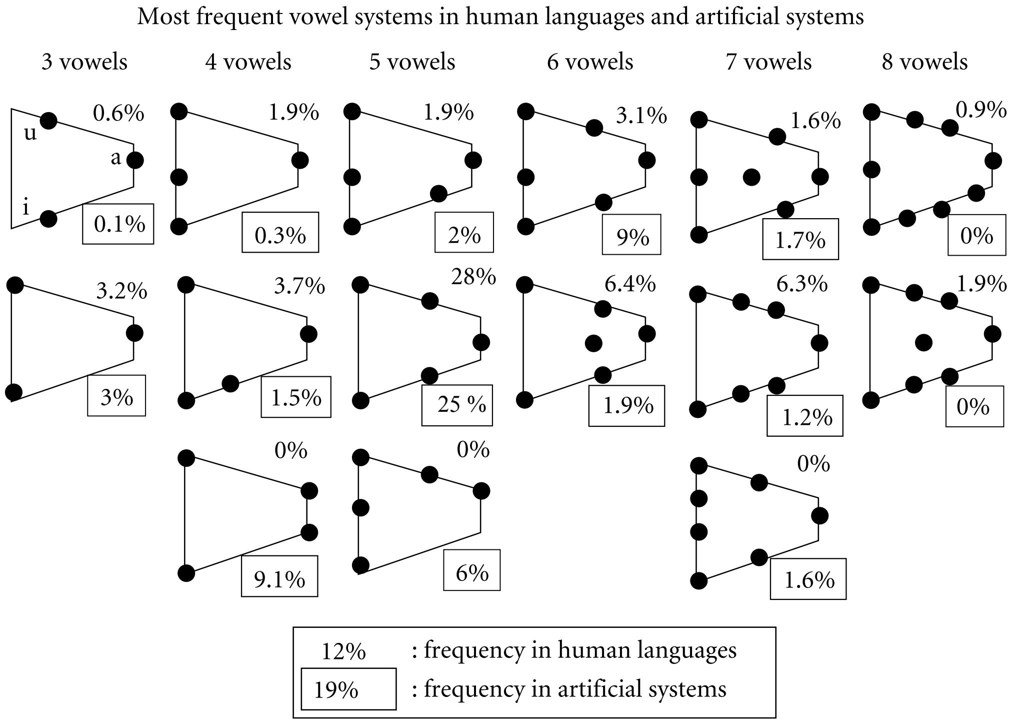


Figure 7.12: Distribution of vowel inventories obtained in the simulations and in the human languages of the UPSID database. The notation is the same as in Figure 7.10.

Chapter 8

Origins of the Syntax of Sounds

In the models described in the two previous chapters, the individuals produced vocalizations whose articulatory targets were chosen randomly from the repertoire defined by the neural networks. We saw how this repertoire of articulatory targets could self-organize, passing from a quasi-continuum to a set of discrete phonemes. We also saw that because the number of these discrete articulatory units (phonemes) was small compared to the number of vocalizations an individual could produce in its lifetime, there was necessarily systematic re-use of these phonemes in composing the vocalizations. By contrast, there was no organization to the specific ways in which these phonemes were sequenced – it was random. This means that individuals could produce all possible sequences of phonemes in their repertoire.

However, the sound systems used by humans, as explained in Chapter 2, organize very strictly the ways in which phonemes can be combined. In particular, every language only allows certain sequences of phonemes, and not others. For example, in English "spink" is a possible word, while "npink" or "ptink" are impossible. In Tashliyt Berber the words pronounced [tgzmt] et [tkSmt] are allowed, but they are impossible in French. Each language is not only a code defining a shared inventory of phonemes, but also a code defining a shared repertoire of possible combinations of phonemes. An essential unit of combination in Linguistics is the syllable, defined as a vocalization produced during one oscillation of the jaw (McNeilage, 1998).

The inventories of syllables permitted in the world's languages therefore have organization, in the sense that there is structure in the set of permitted phoneme combinations. A syllable comprises a certain number of locations, such as the "onset", the "nucleus" and the "coda", where in many cases only certain phonemes can occur. There are languages such as Japanese where syllables can contain only two phonemes: consonants can only occupy the first position and vowels in the second position (these syllables are termed "CV"). Sometimes it is groups, or clusters, that can occur in certain locations only. So the repertoires of syllables in human languages are organized according to patterns (e.g. CV, CCV). The term "phonotactics" is used to refer to these rules of syntax. Certain patterns are statistically preferred over others across the world's languages. For example, all languages allow syllables of the type CV, while many do not allow consonant clusters at the beginnings of syllables.

In this chapter, we will show how an extension of the model presented in Chapter 6 not only enables shared discrete phoneme inventories to appear, but also shared syllable repertoires organized in patterns according to rules governing the syntax of sounds. We will also show how certain patterns can be statistically preferred when articulatory and energetic constraints are introduced.

8.1 Self-organized death of temporal neurons

Instead of selecting articulatory targets by an algorithm that randomly activates neurons in the motor map, which can be seen as a "spatial" map, a map of "temporal" neurons will carry out this task. That is, to the neural map modelling relations between organs (the spatial map), one adds a neural map modelling sequences of articulatory targets (that is, sequences of activations of neurons in the spatial map).

These temporal neurons are each connected to several motor/spatial neurons. They can both send and receive signals through these connections. These neurons are temporal because their activation function has a temporal dimension: their activation does not depend solely on the amplitude of activation of the spatial neurons to which they are connected, but also on the order in which they are activated. Thus, when they receive signals from the spatial map, their activation is calculated from the temporal evolution of the activity of the spatial neurons[1]. This activation mechanism is such that each connection has a weight that varies over time according to a Gaussian function, and that at any time the temporal neuron is the sum of the activations of the spatial neurons to which it is connected, weighted accordingly. The activation of the temporal neuron after a vocalization is the sum of all the weighted sums obtained during perception.

The spatial neurons to which a temporal neuron is connected determine the articulatory targets of the sequence coded by the temporal neuron. The activation of a temporal neuron results in the activation of its connected spatial neurons according to a temporal pattern. Here the temporal pattern will be regular, first activating a single neuron corresponding to the first articulatory target, then a second neuron corresponding to the second articulatory target, then a third, a fourth, and so on. The simulations will here in fact use only two articulatory targets per vocalization: each temporal neuron will therefore be connected to only two spatial neurons. This will make the temporal map easier to represent visually.

Initially, a large number of temporal neurons are created (500, compared with 150 neurons in the spatial map), and these are connected randomly to the spatial map with random values of their internal parameters. Using so many neurons means that a large part of the space of possible combinations can be covered. Unlike the neurons in the spatial map, that behave like in Chapter 6, the properties of temporal neurons do not change over time. So, for the moment, the presence of these neurons has no impact on the simulation in Chapter 6: vocalizations are still composed of arbitrarily selectable articulatory targets. An additional mechanism is therefore required in order to achieve the goal that we have set ourselves in this part.

This mechanism is inspired from apoptosis, or programmed cell death in the human brain (Ameisen, 2000), and fits with the theory of neural epigenesis developed by Changeux

[1] The mathematical formula for computing the activation of a temporal neuron i is:

$$tuneTemporel_i = \sum_{t=0}^{T} \sum_{j=1}^{Ntarget} \frac{1}{\sqrt{2\pi}\sigma}.e^{|t-T_j|^2/\sigma^2}.\frac{1}{\sqrt{2\pi}\sigma}.e^{|tune_{j,t}|^2/\sigma^2}$$

with T denoting the duration of the received signal (i.e. the duration of the perceived vocalization), $Ntarget$ the number of spatial neurons to which it is connected, T_j a parameter which determines when the temporal neuron i is sensitive to the activation of the spatial neuron j, and $tune_{j,t}$ the activation of the spatial neuron j at t.

and Danchin (Changeux and Danchin, 1976). The theory basically proposes that neural epigenesis consists of an initial massive generation of random neurons and connections, which are afterwards pruned and selected according to the level of neurotrophins they receive. Neurotrophins, which block an automatic suicide system of neurons, are above all generated and provided to neurons that are sufficiently activated. Stimulated neurons survive, while the the others die. This phenomenon has sometimes been called "activity-dependent growth" (Ooyen, 2003).

This principle of generation and pruning is applied to the temporal neurons, depending on their mean activity level[2]. The *vitalThreshold* constant defines the level of activity below which the neuron is pruned. This threshold remains the same for all neurons in the map. The pruning process does not begin immediately after the creation of the cell, but after a few dozen vocalizations, when the mean activity level has had time to stabilize a little.

8.2 Dynamic formation of patterns of combinations

A population of 5 individuals interacts in the same way as in Chapter 6: they are located randomly in a virtual environment, and at random moments produce a sound by random activation of a neuron in their temporal network. The nearest individual, and the vocalizing individual itself, hear the vocalization and update the neurons of their two networks.

We will make the same assumptions as in Chapter 6 concerning the nature of the spaces involved and their dimensionality. Again, we assume that the individuals can pass from the acoustic space to organ relation space, and also between organ relation space and the space of muscular activations. So our focus here is only on organ relation space. Moreover, as at the beginning of Chapter 6, a 1-dimensional organ relation space will be used. This will make it possible to visualize the neurons in the temporal map: since they are only connected to two neurons in the spatial map, they are defined by two values (two 1-dimensional articulatory targets), and thus by a point in a 2-dimensional space. Figure 8.1 represents the networks of two individual at the beginning of a simulation. Both the horizontal and vertical axes represent organ relation space. The points on the two axes are the preferred vectors of the neurons of the spatial map. Thus the same points occur on both axes. Then the points in the body of the plotted space represent neurons of the temporal map. Each one has an X and a Y value at one of the preferred vectors of the spatial neurons: the value on the horizontal axis corresponds to the first articulatory target coded by the temporal neuron, and that on the vertical axis corresponds to the second articulatory target coded by the temporal neuron. The figure shows that the set of temporal neurons covers just about all the continuous space of possible combinations defining of vocalizations. The little areas around each point represent the average activation level (here initial) of the temporal neurons. The larger the area, the higher the level of activation (it is initially the same for all neurons).

[2] The mean activity of a temporal neuron j is computed with the formula:

$$< tuneTemporel_j >= \frac{<tuneTemporel_j>*(window-1)+tuneTemporel_j}{window}$$

where $< tuneTemporel_j >$ represents the mean level of activity and $tuneTemporel_j$ the last activity level observed following the perception of a vocalization. This mean activity level is computed in such a way that it is not necessary to memorize the previous activity levels. The value of *window* is 50, and the initial value of $tuneTemporel_j$ is equal to $2 * vitalThreshold$.

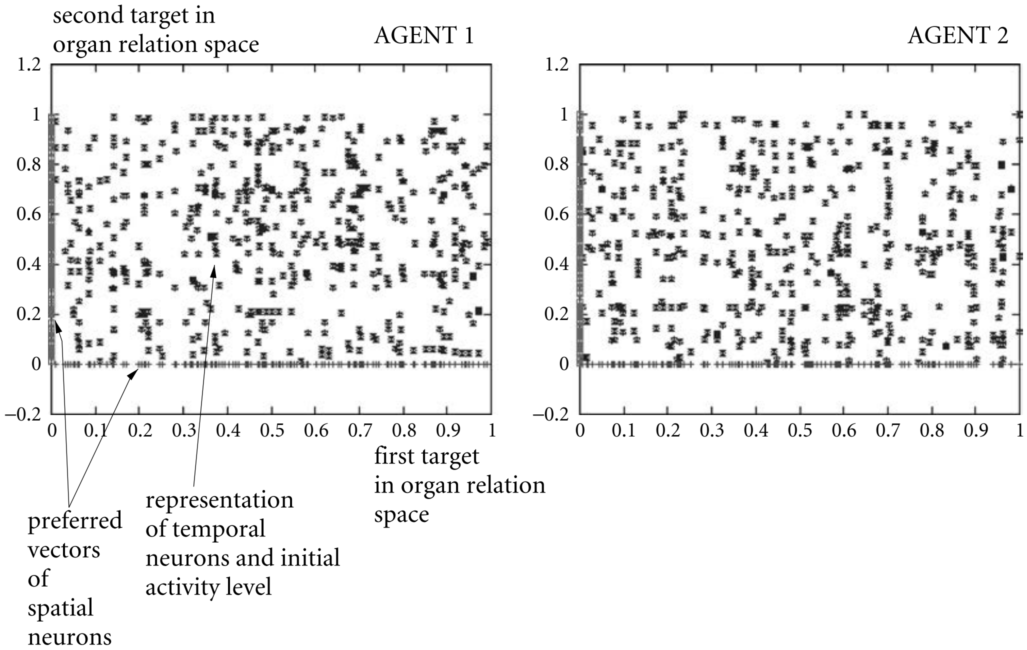


Figure 8.1: The two neural maps of two individuals at the beginning of the simulation: the spatial map (whose preferred vectors are represented by points on the axes), and the temporal map (whose neurons are represented by little areas).

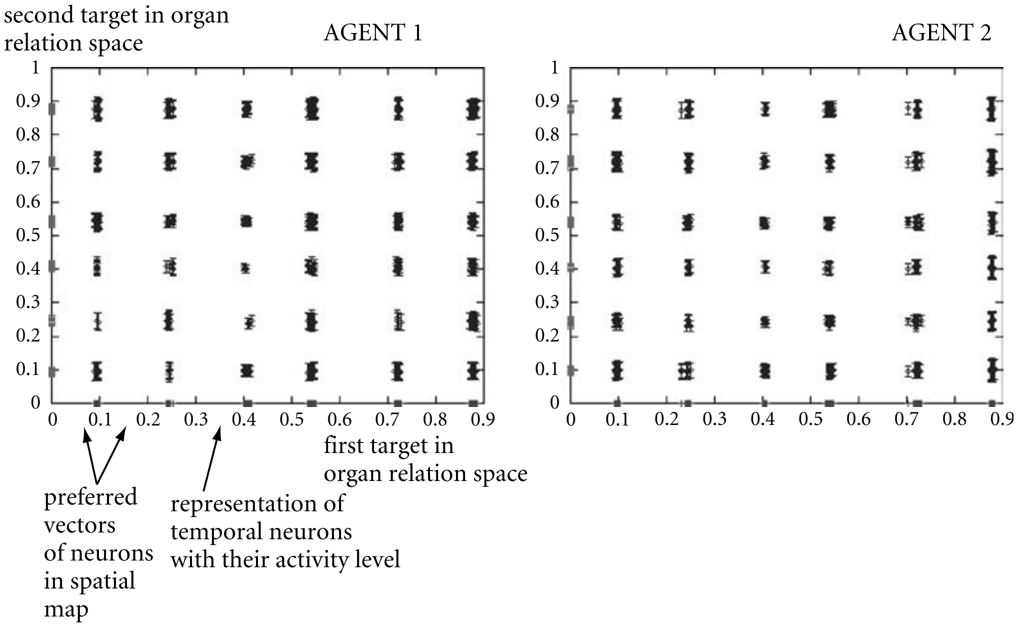


Figure 8.2: Illustration of the sort of temporal map obtained if the elimination mechanism for temporal neurons is not used. The rectangle on the right shows the maps of another individual chosen at random from the population.

Before presenting the results yielded by the generation/elimination dynamic of the temporal neurons, Figure 8.2 shows what happens if no neurons at all are eliminated. It shows an individual's map after 1000 interactions in a population of 10 individuals. Here, as in the simulations presented in the previous chapters, a clustering appears in the spatial map. As far as the temporal neurons are concerned, we observe that they cover all the mathematically possible combinations of two phonemes (whose categories are defined by belonging to clusters). There is thus no constraining principle affecting the combination of phonemes, thus no phonotactics, and even fewer patterns.

Figure 8.3 now shows what is obtained if elimination of insufficiently activated neurons is implemented. Here we have the networks of two individuals, taken randomly in a population of 10 interacting individuals, after 1000 interactions. Note first that the neurons in the temporal map no longer cover all the possible combinations. Furthermore, they cover the same combinations in both individuals. This implies that the vocalizations they produce only exhibit certain phoneme combinations and not others, and that the rules of combination are shared: thus we observe the appearance of culturally shared phonotactic rules. These rules are different for each simulation, and therefore diverse. Figure 8.4 gives another example of a result.

A second observation might be made concerning the organization of temporal neurons surviving in the space of combinations that they encode (that is, in the space represented in the figure). The combinations appearing after several hundred interactions are not dis-

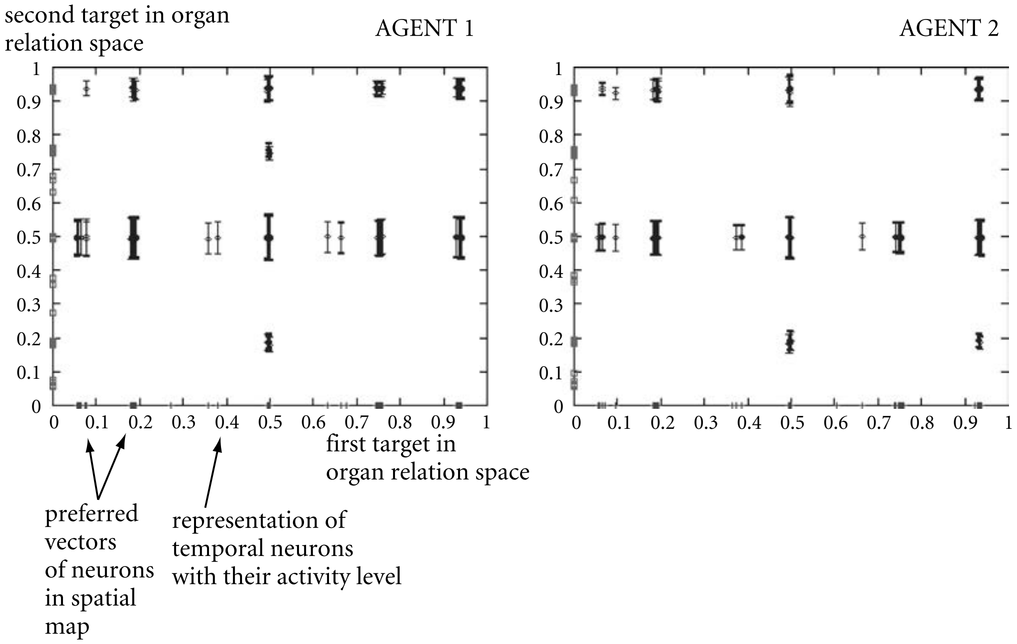


Figure 8.3: If the pruning mechanism on temporal neurons is used, this is the kind of pattern obtained at the end of a simulation. On the one hand, note that not all phoneme combinations are represented in the temporal map, and thus the individuals only produce those combinations for which they have temporal neurons. On the other hand, the possible phoneme combinations are organized in patterns: rules of sound syntax have emerged through self-organization.

tributed randomly among those that are possible. They organize themselves into rows and columns. For example, in Figure 8.4, there are two columns and one row. If the phonemes associated with the eight clusters of the spatial map are labelled $p_1, p_2, ..., p_8$, as indicated in the figure, then the syllable repertoires of these individuals can be summarized as $(p_6, *)$, $(p_8, *)$ et $(*, p_7)$ where $*$ is a 'wild card' meaning 'any of the phonemes $p_1, ..., p_8$". The repertoire is thus organized in patterns, just as the repertoires of human languages can be summarized using patterns (Japanese, for example, is encapsulated in the patterns "CV" , "CVC" and "VC").

The formation of shared repertoires, and also the formation of patterns, is the result of a dynamic of competition and cooperation among temporal neurons. In fact, the vital threshold for these neurons was chosen so that in the case of Figure 8.2, where there are temporal neurons for all combinations, the frequency with which each combination is produced would be too weak to raise them above the critical activation level. Imagine that the initial situation in the simulation is that in Figure 8.2 where the clusters have already formed in the spatial network. In other words, imagine that we disengage the dynamic of the spatial neurons from that of the temporal neurons. In that case, every neuron has an initial activation level ($MA_{j,0} = 2.vitalThreshold$) twice as high as its vital threshold level. This vital level was fixed to be higher than the average activation level generated by this

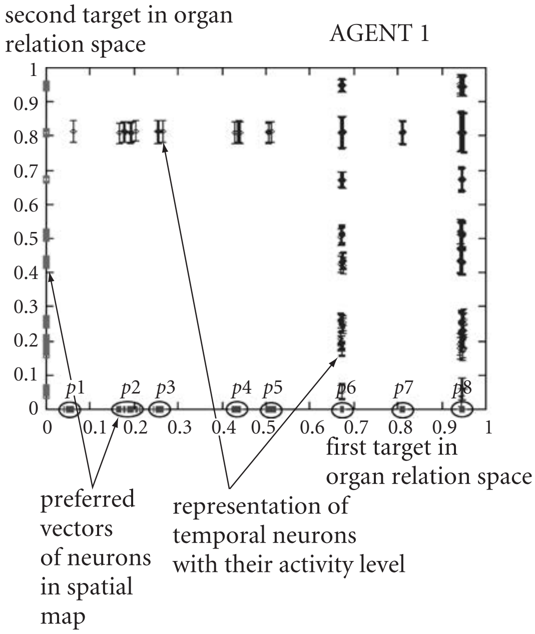


Figure 8.4: Another example of a phonotactic system generated by a population of individuals using the pruning mechanism.

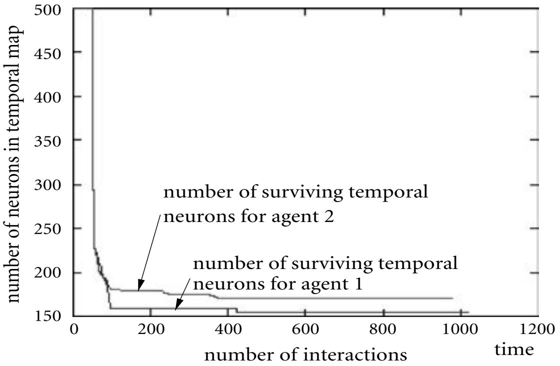


Figure 8.5: Evolution of the number of surviving temporal neurons in the simulation which generated the system in Figure 8.3. At first a phase is seen in which a certain number of neurons die, and then there is a stabilization phase in the system during which the remaining neurons succeed in surviving.

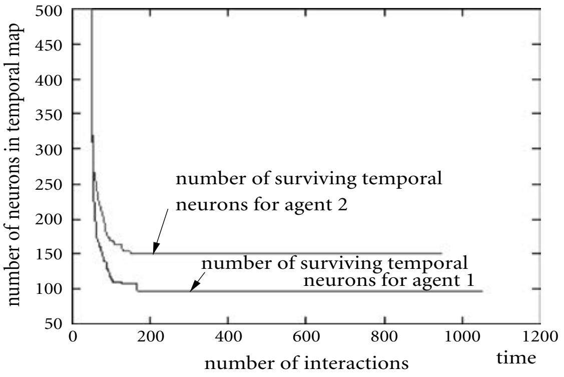


Figure 8.6: Evolution of the number of surviving temporal neurons in the simulation which generated the system in Figure 8.4. We can observe that here the two individuals do not possess exactly the same number of surviving neurons: this is due to the intrinsic stochastic nature of the system. Nevertheless, as Figure 8.4 indicates, they share the same phonotactics and the same patterns.

initial configuration. As a consequence, the mean level of activity of all neurons is going to decrease at the beginning of a simulation. Because there is natural stochasticity in the system, due to the random choice of temporal neurons when vocalizations are produced and the not quite equal sharing of neurons in the clusters of the spatial maps, these activation levels will not all decrease at exactly the same rate. In particular, some neurons will get below their vital level before others, and therefore die. The survival of a neuron in a cluster of an individual's temporal map depends in part on the number of neurons corresponding to the same combination in other individuals, whose survival itself depends on the density of the cluster in question in the first individual. This creates positive feedback loops with the result that when by chance a certain number of neurons die in an individual's cluster, this facilitates the death of corresponding neurons in other individuals. In the same way, healthy clusters (meaning that they have many neurons with high activation levels, which prolongs their effectiveness) will facilitate the survival of corresponding neurons in other individuals.

In this way, the interaction of forces of competition and cooperation takes advantage of fluctuations to lead the system into a stable state where the no more cells are dying. In this state each neuron ends up with an activation level higher than the vital level. Figures 8.5 and 8.6 show the evolution of the number of surviving neurons within the temporal maps of two individuals from simulations whose outcomes are shown in figures 8.3 et 8.4. There is a clear separation between the phase of neuron death and the stabilization phase.

Mutual reinforcement among neurons does not only take place between neurons of clusters corresponding to the same phoneme combination in different individuals. It can also occur between clusters sharing only one phoneme at the same place in a vocalization. We described above how a temporal neuron is activated: computation on the sum of local similarities between signals (filtered by a Gaussian function) is continuous. Let us denote p_1, p_2, p_3 and p_4 four distinct articulatory targets belonging to four distinct clusters. If the similarity of two vocalizations with the same sequence of phonemes is about 1, then the similarity between the vocalization coded by the sequence (p_1, p_2) and the vocalization coded by the sequence (p_1, p_3) is about 0.5, and the similarity between (p_1, p_2) and (p_3, p_4) is about 0. This means that the level of activity "provided" to the temporal neurons of a cluster *cl* thanks to two clusters of temporal neurons in other individuals which share exactly one phoneme in the same location, is about the same as the level of activity provided to the neurons in *cl* thanks to the cluster in other individuals which corresponds to temporal neurons sharing all the phonemes in the right location with those in *cl*. As a consequence, groups of clusters reinforcing each other will form during the self-organization of the temporal neuron map through positive feedback loops. These are the lines and the columns that we saw earlier. This explains why we observe the formation of phonological patterns in the individual's repertoire of syllables. To summarize, the interactions between competition and cooperation among individual clusters explains the formation of shared and stable repertoires of allowed phoneme sequences, while the interaction between competition and cooperation between the "lines" and the "columns" of clusters explains the formation of phonological patterns.

8.3 Role of articulatory and energetic constraints

With the mechanism presented in the previous sections, if a large number of simulations are run, there will be no statistical preferences for localization of either temporal neurons or spatial neurons. This is similar to the simulation without articulatory bias in Chapter 6 in which the clusters had as much chance of appearing at one spot in organ relation space as at another.

As in Chapter 6, we will see here how an articulatory bias can introduce preferences. This articulatory bias, modelling nonlinearities in the function mapping sounds to articulatory configurations, will be modelled through the initial distribution of preferred vectors. The initial distribution of preferred vectors of the spatial map will be similar here to the initial distribution of preferred vectors in the simulation with articulatory bias of Chapter 6: the density of preferred vectors will increase between 0 and 1. This means that initially there will be a higher density of preferred vectors closer to 1, and the lowest density will be near 0. Figure 8.7 gives two examples of the initial distribution of preferred vectors.

One can consider another bias introducing energetic constraints. In a human, for example, each vocalization calls for the displacement of a certain number of organs, and this displacement costs energy: some vocalizations are easier to pronounce than others from the point of view of the muscular energy expended. Moreover, several researchers (Lindblom, 1992; Redford et al., 2001) have already proposed that energetic cost is an important constraint in the formation of the sound systems of human languages. This constraint will be modelled here by assigning each vocalization an energy cost, to be defined, which will influence the level of neurotrophins received by the neurons coding for them (thus, at equal frequency, a neuron coding for an easily pronounced syllable will receive more neurotrophins than a neuron coding for a syllable that is hard to pronounce, and thus will have a greater chance of survival). The combination of these two constraints will enable us to show how the interaction of two different biases can lead to the formation of repertoires of syllables with structures whose statistical properties are not deducible from any single bias on its own.

The energy cost associated with a vocalization is defined by the displacement of the vocal organ from a position of rest fixed as coded by point 0 in organ relation space. The

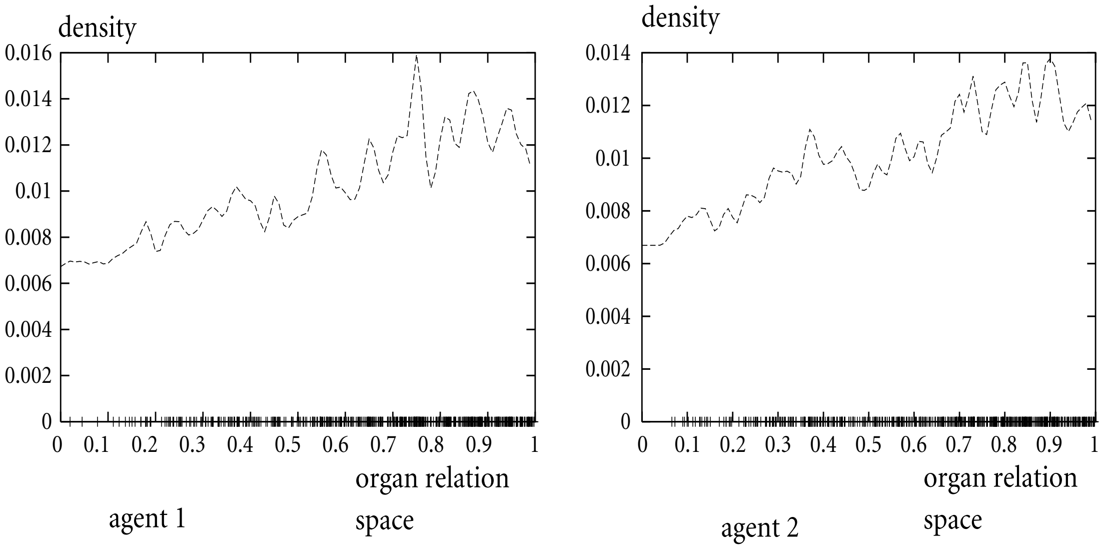


Figure 8.7: Example of the initial distribution of preferred vectors in the spatial maps of two individuals, biased by articulatory constraints. At the start there are more neurons coding for articulatory configurations close to 1 than neurons coding for articulatory configurations close to 0.

most distant position from this is coded by 1[3]. This energy will influence the survival of the temporal neurons. Whereas previously the survival of temporal neurons depended only only on their level of activity, here a term corresponding to the energy is added. This term does not directly represent the energy, but $c_1 - energy$, thus giving a low value when a lot of energy is expended (and the neuron therefore receives fewer neurotrophins), and a high value when not much energy is expended. The constant c_1 is chosen such that on average the activity and energy terms are of the same order of magnitude[4]. The vital threshold is once again set quite low so that not all temporal neurons can survive.

Figure 8.8 shows the initial temporal map of two individuals, with the initial levels of neurotrophins of the different neurons. Only 150 neurons randomly chosen from among the 500 are depicted so that the disparity in densities can easily be seen. The segments here are shown with their initial values. Only 150 randomly chosen temporal neurons are

[3] To calculate the energy expended, the articulatory trajectory is discretized and the sum of all the distances between the points on this trajectory and the resting position is calculated. If p_1 and p_2 are two phonemes used as articulatory targets in a vocalization, and $p_1, p_{int_1}, p_{int_2}, ..., p_{int_{N-1}}, p_2$ are the successive points in the trajectory generated after interpolation, then the energy expended is:

$$e(p_1, p_2) = p_1^2 + p_{int_1}^2 + p_{int_2}^2 + ... + p_{int_{N-1}}^2 + p_2^2$$

[4] The level of neurotrophins received by each neuron j is thus

$$level(j) =< tuneTemporal_j > +(c_1 - energy(j))$$

With 150 neurons in the spatial map, 500 neurons initially in the temporal map, the same Gaussian parameters that were used at the beginning of this chapter and a discretization of the articulatory trajectory into 10 points to compute the energy, we use $c_1 = 15$ and $vitalThreshold = 0.03$.

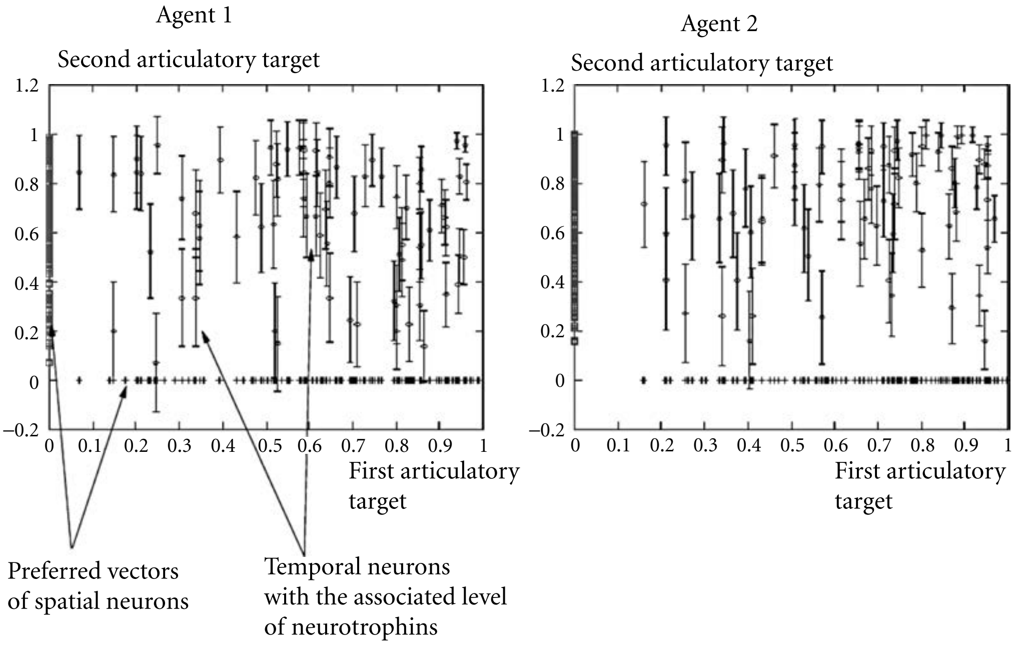


Figure 8.8: Example of biased initial temporal neural map: the segments associated with the temporal neurons show the initial level of neurotrophins of these temporal neurons. We observe that temporal neurons close to $(0, 0)$ have the largest initial level of neurotrophins, but that the temporal neurons close to $(1, 1)$ are more numerous and so will be activated more often initially, which means that they will receive more activity-dependent neurotrophins than those close to $(0, 0)$.

represented here rather than all 500, so that the disparity in densities can be seen clearly. The segments on each neuron represent their initial level of neurotrophin. This initial value is not the same for all neurons because it is computed with an initial value for mean activity that remains fixed[5] and an energy value which varies with each temporal neuron. A look at Figure 8.8 shows that there are more temporal neurons in the top right, but that those in the lower left have a lower energy cost, and therefore receive more neurotrophins at the same level of activation. There are thus two forces, each favouring the survival of one corner of the space to the detriment of another. If there were no energy cost, as in Chapter 6, a statistical preference for clusters of spatial neurons close to 1 would appear, leading automatically to a statistical preference for clusters of temporal neurons in the upper right quadrant. Likewise, if only neurons corresponding to vocalizations which do not use much energy were to survive, there would be a statistical preference for temporal clusters in the lower left. The combination of these two forces will give a different statistical preference. To evaluate this 500 simulations were run. Figure 8.9 thus represents the set of surviving temporal neurons after 500 simulations. There is a clear statistical preference for neurons located in the centre of the space, and not at the edges.

From a qualitative point of view, these statistical preferences have structural similarities

[5] The initial value for mean activity is 0.02

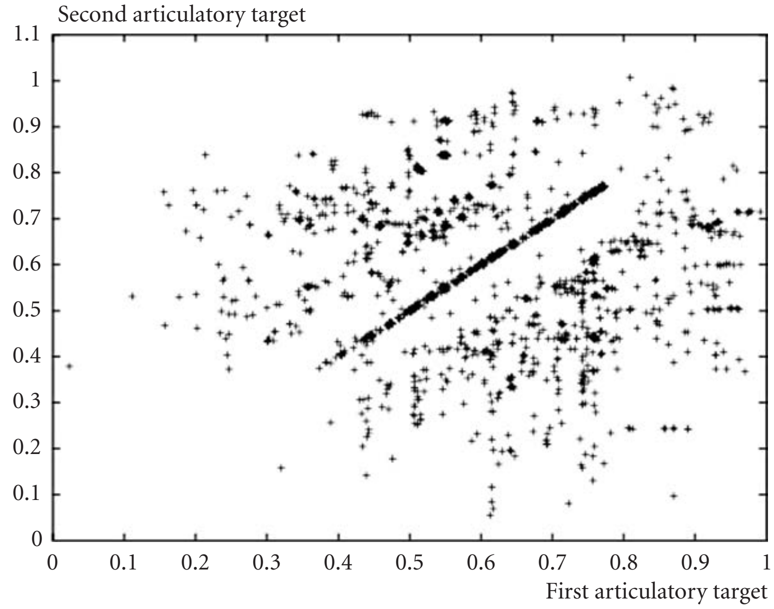


Figure 8.9: Distribution of surviving temporal neurons in 500 simulations. The self-organized repertoire of syllables re-uses mainly elementary articulatory configurations in the centre of the space, reflecting a compromise between constraints due to nonlinearities and the articulatory cost of vocalizations.

with those found in human languages between various syllable repertoires: human languages prefer CV syllables over CVC syllables, which are in turn preferred over CC syllables, then CCV, then CVVC/CCVC/CVCC. Unlike in the simulations in chapters 6 and 7, where it was possible to model realistically the articulatory constraints related to the production and perception of vowels and thus create a direct parallel between the model and human languages, no model exists that is both realistic and capable of being simulated efficiently enough in the context of full syllable production. This is why the model presented in this chapter has not attempted to model carefully any particular physiological constraints. It was nevertheless shown, as in Chapter 6, how populations of individuals might generate speech systems having qualitative structural similarities to human languages.

Chapter 9

Evolutionary Scenarios: Speech Codes Emerging from Babbling and Adaptive Vocal Imitation

In the previous chapters, we have described mechanisms enabling a population of artificial individuals to develop spontaneously a system of vocalizations that had human-like properties (cultural sharing, discreteness, combinatoriality, syntax of sounds). These models were built on several assumptions, in particular concerning the cognitive and neural architecture of individuals. We will now discuss these assumptions in relation with evolutionary scenarios that could explain aspects of the origins of speech.

First, we will argue that the neural models we used share general functional correspondances with human neural systems, both in terms of architecture and plasticity. This is made possible by the fact that they are abstract and generic. Indeed, they are not models of the physiology of biological neurons, but rather models of certain functions of structures encountered in biological brains. Thus, the neural models presented so far constitute a mathematical formal language to model these functions. They are also compatible with several families of models of the physiology of neurons elaborated in computational neuroscience (Arbib, 2005). This kind of model can be reformulated using another formal language, that of Bayesian probabilities, often used to model brain functions (Knill and Pouget, 2004), and like Moulin-Frier, Schwartz, Diard and Bessière have achieved in other work on the morphogenesis of speech (Moulin-Frier et al., 2008, 2015).

Second, we will study in what ways these models can illuminate the problem of the origins of speech, using the framework of explanation we developped in chapter 3. In particular, we will first explain that they are compatible with adaptationnist scenarios of the origins of speech, strengthening them through providing an explanation of mechanisms that can generate the structures of speech (complementing mechanisms of natural selection). It appears that the low complexity of the models' assumptions, especially in comparison with the speech code they generate, makes it plausible that natural selection could evolve them relatively easily under pressure for linguistic communication.

Third, the genericity of the mechanisms involved in these models will allow us to propose a scenario in which they could have evolved without direct relation with a selective pressure for linguistic communication. We propose that shared speech structures could be collateral effects of the formation of other structures that are older from an evolutionary point of view, such as vocal imitation, or as we will study in the next chapter, intrinsic motivation systems that push organisms to explore their body and their environment driven by pure curiosity. This leads us to suggest that contemporary speech codes may be exaptations, which first versions may have been the result of self-organization of structures not initially associated to language.

9.1 Functional correspondences with human neural systems

In chapter 7, we showed the similarities between vowel systems generated by artificial individuals in the simulations and human vowel systems. We will argue here that the assumptions of mechanisms which lead to these results have several strong functional correspondences with the human brain. While we employed terms such as "neural unit" and "neural network" to describe the internal architecture of artificial individuals, one has to note that they are not models of the physiology of human brains. These terms were used for functional analogy. These artificial neural architectures, and their mechanisms of plasticity, are abstract mathematical models of certain functions of biological neural networks. Let us look at these correspondences.

First, the basic elements of the brains of the individuals in the models are artificial neural units that respond maximally for a certain stimulus, and this response decreases as the stimuli presented become more different from the "preferred" stimulus. The spatial neural units introduced in chapter 6 rely on mechanisms commonly used in the neuroscience litterature (Arbib, 2002), with a Gaussian tuning function that is often used in such models. Other decreasing tuning functions could have been used, as long as they converge to zero far from the preferred stimulus, and would lead to very similar results. The temporal neural units introduced in chapter 8 are less frequently used but are part of the conceptual landscape in neuroscience. For example, the neurons used in visual cortex have such spatio-temporal activation functions (Dayan and Abbot, 2001). Similar neurons appear to be used also in auditory processing which implies a temporal component of perception (Kandel, 2001).

The plasticity mechanism that allows neurons to adapt their preferred vectors upon the perception of new stimuli is also relatively common. The reinforcement of the sensitivity of neural units to stimuli that activate them strongly constitutes the basis of Kohonen self-organizing maps (Kohonen, 1982), themselves considered as models of cerebral cortex (Afflalo and Graziano, 2006). This law of adaptation is often formulated in terms of hebbian reinforcement of inter-neuron connections (Arbib, 1995). Other models similar to those we have studied, and focused on the perception of vowels, have lead to significant advances about the understanding of perceptual development in speech (Guenther and Gjaja, 1996).

In the models, neural units of individuals are organized as maps peculiar to each modality, which is an important property of the organization of biological brains. Furthermore, these maps are connected in such a way that individuals can learn to master the correspondences from one space to another: they can find the articulatory configuration which correspond to a sound and vice versa. This competence to translate from one space to the other is obviously present in humans as this is needed for imitation. Below we will discuss the possible origins of such a capacity. For now, let us just note that this capcity is found in many other animals: for example, an equivalent capacity to master the coordination between hand movements and vision exists in animals who have hands. Recent discoveries concerning mirror neurons

(Rizzolatti et al., 1996) have stimulated a lot of animation in the scientific community. These are neurons, initially observed in monkeys but also present in humans, which are normally activated when the animal produces a motor action (e.g. take an apple in the hand), but which are also activated when they see another animal doing the same action. However, this discovery does not go much further than identiying certain parts of the neuronal path that enable to translate from one space to another, and it is still unclear whether this circuitry is innate or the result of learning. The structure of connections between neural maps in the simulation was designed to be as generic as possible and does not assume such a precise innate circuit mapping certain actions to certain perceptions. Rather, it relies on hebbian adaptation mechanisms that learn to map the activities of motor neurons to the activities of perceptual neurons.

The coupling between perceptual and neural maps and the dynamic adaptation of preferred stimuli of neural units produces the phenomenon of "perceptual attunement" well-established in the phonological development of infants (Vihman, 1996): the perception of a sound increases the probability to produce a similar sound in the future.

The preferred stimuli of neural units, as well as the connections between neural units, are initially random and adapt, or even survive and disappear for temporal neurons, as a function of their activation across the life of the individual. Activity-dependant adaptation, growth and pruning is a brain building mechanism supported by selectionnist theories of neuronal epigenesis proposed by Changeux, Courrège and Danchin (Changeux et al., 1973) and Edelman (Edelman, 1993). These theories were recently reinforced by the discovery of apoptosis, or programmed cellular death, which shows that all neurons have an internal suicide program which has to be inhibited to survive (Ameisen, 2000). This inhibition happens through reception of neurotrophin, in particular when neurons are sufficiently stimulated. Several approaches in neuroscience, such as the theory of Changeux, Courrège and Danchin (Changeux et al., 1973), as well as more recent theories like the one proposed by Fernando and Szathmary (Fernando et al., 2013), conceptualize learning in the brain as initial generation of random neural material which is progressively sculpted through interaction with the environment.

In the simulations we presented, the categorization of stimuli by each neural map has been modelled through the dynamic relaxation of recurrent neural networks, until they fell into an attractor (which was a fixed point). The category of the stimulus was the identity of the attractor. Conceptualizing perceptual categories as attractors of brain activity considered as a dynamical system is an idea proposed several times in the scientific community, for example by Edelman (Edelman, 1993) or Kaneko and Tsuda (Kaneko and Tsuda, 2000). Freeman (Freeman, 1978) conducted experiments on the dynamics of electric activity of the olfactive bulb in rabbits that confirm the plausibility of this hypothesis.

Vocal production by the artificial individuals we considered was happening through the selection of several articulatory objectives, which were reached in sequence through a continuous movement of the vocal tract organs. This organization of motor control along two levels has been the focus of several works on human motor control (Flash and Hochner, 2005). At the higher level, a motor program specifies the objectives to be reached (the gestural score in the case of speech). At the lower-level, other neural mechanisms execute the program taking into account the current physical and functional constraints. In the preceding chapters, we studied the higher level of motor control, modelling their low-level execution with a relatively simple mechanism (interpolation). Thus, we did not consider in detail the interesting phenomena of co-articulation and their potential influence on the construction of phonemic repertoires.

9.2 From models to evolutionary scenarios of the origins of speech forms

We have so far focused on how the mechanisms we presented, including their assumptions, could generate systems of human-like vocalizations in groups of individuals. We will now discuss the possible evolutionary origins of these mechanisms, and of the function of the biological building blocks they model. In particular, we will investigate whether the most important of these functions could have evolved independantly of a function/pressure for linguistic communication.

If the perceptual and motor systems in these models are considered independantly from each other, it is possible to explain their origins outside the function of communication. All mammals with ears have also neuronal ensembles dedicated to sound processing, and other neurons dedicated to the control of their mouth and vocal organ. As we have seen, the way motor control operates in the models is very generic and based on neural map structures that can be used in multiple modalities. Furthermore, plasticity in these neural maps (which increases the sensitivity to perceived stimuli) is typical of unsupervised learning mechanisms, also called latent learning, in mammals.

However, two mechanisms in these models are associated to behaviors and capacities which distinguis humans from many other species (but not all). The first mechanism corresponds to the plastic neural structure connecting auditory and motor/phonatory systems. In the models, these structures allow individuals to learn perceptuo-motor correspondences of their own vocalizations, as well as to map the vocalizations produced by other individuals to their own motor commands. What could be the evolutionary origins of such a connection structure? Has it evolved under linguistic selective pressures? In section 9.3, we discuss such an adaptationist scenario. Or could it evolve without a pressure for linguistic communication? Its genericity allows to formulate the hypothesis that adaptive vocal imitation could be the original function for which it evolved. Interestingly, as we will discuss in section 9.4, this capacity is shared by several species which do not have language, such as certain species of birds or whales, but have combinatorial systems of vocalization with complex rules of combination.

A second mechanism that is special and key in the models we discussed is babbling: a mechanism that pushes individuals to spontaneously explore their vocal space. In the models, this mechanism is crucial to allow individuals to experiment their vocal tract, collect sensorimotor observations, and learn the auditory-motor mapping thanks to the plastic intermodal neural connections. It is also key for the generation of a diversity of speech sounds influencing the vocalizations of others. One hypothesis is that an ad hoc vocal babbling mechanism could have evolved as a building block in the morphogenesis of speech, in the specific functional context of adaptive vocal imitation.

A complementary hypothesis to understand the origins of babbling, which we will study in the next chapter, is that it may be a natural consequence of a general intrinsic drive to learn and gain information about the world, which we call "curiosity" in everyday language, and that is expressed particularly strongly in humans. Indeed, a distinctive trait of humans is their propensity to explore spontaneously their own body and their environment, driven by the pure pleasure to learn, to acquire new information, to practice activities for themselves and not to accumulate extrinsic rewards such as food or social recognition. Such a propensity to explore driven by curiosity is observed as well in infants who explore avidly objects they can grasp, as in adults when they play games like sudoku or read novels. These mechanisms of "intrinsic motivation", a term used by psychologists, were studied in many research programs in educational and developmental psychology (Deci and Ryan, 1985; Berlyne, 1960; Csikszenthmihalyi, 1991; Oudeyer et al., 2016), and more recently in neuroscience (Oudeyer and Kaplan, 2007; Gottlieb et al., 2013; Gottlieb and Oudeyer, 2018). They appear to be expressed with a unique amplitude in humans. We will see that these mechanisms, probably evolutionarily older than language, can spontaneously generate and self-organize vocal development. In this context, we will discuss in the next chapter computational models of curiosity-driven learning (Oudeyer et al., 2007; Kaplan and Oudeyer, 2007). These models will allow us to understand how such mechanisms of intrinsically motivated exploration can automatically form organized developmental learning trajectories in individuals, where vocal exploration and imitation appear spontaneously out of the same mechanism that drives learning of object manipulation (Oudeyer and Kaplan, 2006b; Moulin-Frier and Oudeyer, 2012).

9.3 An adaptationist scenario

An evolutionary scenario to consider is one in which the plastic structure connecting auditory and motor systems appeared after specific genetic innovations selected under a pressure for linguistic communication. The models we presented are compatible with such a scenario, and can enlighten the chicken and egg evolutionary problem we discussed in chapter 2. This problem was that it is hard to understand how a convention like a speech (or gestural) code, which is necessary to establish any linguistic communication, can form when there is not already one. Indeed, other models of the evolution of speech (de Boer, 2001; Goldstein, 2003; Moulin-Frier et al., 2008), lexicon or grammar (Steels, 2012), show how a linguistic convention can form when one assumes that individuals can already interact following interaction protocols called "language games". Such language games constitute complex conventionalized interactional structures, which involve signals used to identify who is the speaker and the hearer and what should be done by whom and when, like the rules of a board game. These rules have a syntactic structure and are partly arbitrary (many variations can be imagined for each game). They are in themselves a (pre-)linguistic communication system. Thus, it is hard to see how such interactions can happen without a system of forms (visual or acoustic) permitting to transfer information from one individual to another to regulate their interactions.

Interestingly, the models we studied in previous chapters do not involve language games - they do not rely on any form of explicit social coordination. Yet in such a context, we showed that it was possible to bootstrap a conventional speech code. This opens a door to understanding how such mechanisms could be used by evolution for generating and selecting fundamental building blocks of language that could later on be used in establishing more complex forms of linguistic communication, and be further shaped by sophisticated cultural pressures for communication such as in the models of language dynamics studied by de Steels, Kirby, de Boer, Goldstein or Moulin-Frier.

The nature of the system's components also provide a very different light than the ideas proposed by the innatist perspective proposed by Pinker and Bloom (Pinker et Bloom, 1990) or Chomsky (Chomsky, 1975). These perspectives advocate the genetic pre-wiring of speech sound organization, expressed through brain maturation. The models we studied suggest another hypothesis: no specific "program" encodes specifically digital and shared speech codes in neural networks. Indeed, there is no innate precise wiring between the neural perceptual and motor structures: rather the innate wiring is random and changes according to very generic hebbian rules of plasticity. The initial preferred vectors of perceptual and motor maps are also random. Yet, in spite of this initial randomness, we showed that spontaneous speech structures systematically formed and were shared across individuals. The idea that emerges is that it may not have been a difficult search problem for evolution to find a genetic program that allows individuals to develop speech codes, especially when this capacity has a direct adaptive advantage. Simulations show that there is no need to explore the huge and complicated space of genetic programs that could encode for the maturation of highly specific neural networks. On the contrary, small variations of simple neural structures, such as creating random plastic connections between auditory and motor maps, could be enough. The fundamental role of development in individuals, also called ontogenesis, allows to channel phylogenetic evolution, and vice versa: this is the basis of a new research field called evolutionar-developmental biology (West-Eberhard, 2003; Carroll, 2005).

The limited complexity of the neural structures required to generate such self-organization of speech code also point to novel and stimulating evolutionary scenarios as we will now discuss.

9.4 Exaptation: speech as a collateral effect of adaptive vocal imitation?

Vocal interaction is a very common phenomenon in the animal world. Humans use them for linguistic communication, but they serve many other functions in animals. Several animal species have a sophisticated vocal plasticity (see figure 9.1), both motor and perceptual, but do not have a natural language system: they do not use these vocalizations as symbolic forms arbitrarily and culturally associated to semantic categories [1]. Such vocal plasticity is typically associated with capacities for adaptive vocal imitation (Mercado et al., 2005; Caldwell and Whiten, 2002), as studied in several species of birds (Marler et al., 2004; Kelley et al., 2008) and cetaceans (Frankel, 1998; Handel et al. 2009) capable to learn and repeat the vocalizations of their conspecifics, or even sometimes the sounds of other species or of their surrounding environment (Kelley et al., 2005). The vocal and song repertoires of these species are also often combinatorial, with rules that regulate how sounds can be combined (Balaban, 1988; Frankel, 1998), as illustrated on figure 9.2, a structure that shares strong similarities with human speech structures. While there is not yet a general consensus in the scientific community to explain the functions of these capacities for adaptive vocal imitation and learning, as well as to explain the potential function of combinatorial structures, many hypothesis were proposed and do not involve linguistic communication. For example, such capacities could be used by individuals to find sexual mates by showing their vocal "know-how", or to mark their affiliation to a group that shares the same vocalizations and increase social cohesion, to mark their territory, or to produce sounds that repel their predators (Kelley et al., 2005).

Interestingly, the neural mechanisms in the models we discussed in previous chapters correspond basically to the elementary "neural kit" needed for any form of adaptive vocal imitation. It is hard to imagine simpler mechanisms that would allow an individual to learn perceptuo-motor correspondences and repeat the vocalizations of its conspecifics. It is thus natural to formulate the hypothesis that similar plastic neural structures could have evolved

[1] Several species have a fixed repertoire of alarm calls, used to alert when a danger is coming, but these systems are neither adaptive nor linguistic.

Figure 9.1: Humpback whales and swamp sparrows are two animal species capable of adaptive vocal imitation. They can learn the songs of conspecifics, and songs can vary from one group to the other. Their songs are structured around the reuse of elementary notes taken in a sound repertoire that is specific to each group. (Photos Dr. Louis M. Herman/NOAA et H. Stephen Kerr).

in humans under a pressure for adaptive vocal imitation, for non linguistic functions similar to those we just mentionned for other species. In this perspective, the combinatorial and syntactic properties of sound systems, as well as the capacity to categorize sounds, could be a collateral effect self-organized out of the basic biological material for adaptive vocal imitation. Once such shared combinatorial vocal structures have evolved, their recruitment for efficient linguistic communication is possible. In this context the use of combinatorial vocal sounds in linguistic systems would be an exaptation.

As a consequence, these models of the formation of speech sounds also open a stimulating perspective for the understanding of how organized and shared vocal systems form in animal species that are capable of adaptive vocal imitation. Indeed, the presence of combinatorial sound systems with syntactic rules of sound/note combination, the equivalent of phonemic coding in humans, has so far remained a mystery. The functions attributed to bird or whale songs could often be achieved with much simpler non-combinatorial vocal systems. The models we discussed, which can be transposed to other vocal tract and auditory physiologies, suggest that these syntactic and combinatorial structures could be a collateral and spontaneous effect of the elementary mechanisms of adaptive imitation. Thus, these models open new hypotheses to explain how digital and combinatorial song systems could have self-organized through cultural interaction in groups of birds or cetaceans, as well as how they can vary and evolve across several groups of individuals.

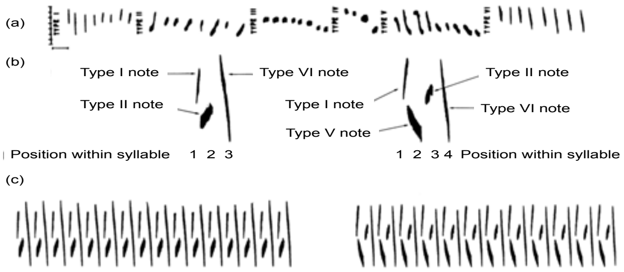


Figure 9.2: The song structure of swamp sparrows (adapté de Balaban (Balaban, 1988)). On line (a), examples of acoustic units (notes) systematically reused by these birds to construct their songs. On line (b), examples of note groupings, also called syllables, that are themselves systematically reused to construct songs. The building of these syllables is constrained by phonotactics rules which vary from one group to the other. Birds from a geographical zones have preferences for using certain notes at certain specific locations in syllables. On line (c), two examples of songs from swamp sparrows. There is a strong similarity between the structure of vocalizations used in these bird species and human speech sounds, as both are based on digitality, combinatoriality, syntactical rules for sound combination, and cultural sharing. Computational models presented in preceding chapters could be applied to account for the morphogenesis of vocal systems in such bird species with adaptive vocal imitation. This opens a new vision on the way digital and combinatorial sound systems can self-organize in groups of birds (or cetaceans with similar capacities), and at the same time how this vocal system vary and evolve across groups of individuals.

Chapter 10

Curiosity, Speech Development and Emergence of Communication

[1]

Vocal babbling appears to be a crucial exploration mechanism in the development and evolution of speech, as discussed in the previous chapter. Did this mechanism evolve specifically for speech? Vocal exploration as it is practiced by human infants in their first months, shares strong similarities with body exploration as a whole - and scientists talk of "body babbling". Beyond exploring the perceptual consequences of their vocal tract movements, infants also explore spontaneously and systematically how the movements of their arms or legs provoke visual changes, proprioceptive or gustatory stimuli when they put them in the mouth, or produce effects on physical objects around them. Such spontaneous body exploration is expressed in infants to an extent that is unparalleled in any other animal.

Rather than imagining ad hoc mechanisms that are specific to exploration and learning in each modality, would it be possible to explain vocal babbling out of general intermodal body exploration mechanisms? This is the hypothesis we study in this chapter. We will see that certain mechanisms called intrinsically motivated learning, implementing a form of what we call "curiosity" in everyday language, can self-organize vocal exploration and interactional structures in individuals through the same process than the discovery of how arms and legs can be used to interact with the physical world. In such a perspective, vocal babbling forms through the dynamics of general sensorimotor development in humans, and this could be a stepping stone to guide and constrain the evolution of speech.

[1] Several parts of this chapter reuse text, with permissions, from the following articles co-authored with Linda Smith, Clément Moulin-Frier, Mai Nguyen and Sébastien Forestier:
Oudeyer, P-Y. and Smith. L. (2016) How Evolution may work through Curiosity-driven Developmental Process, Topics in Cognitive Science.
Moulin-Frier, C., Nguyen, S.M., Oudeyer, P-Y. (2014) Self-organization of early vocal development in infants and machines: the role of intrinsic motivation, Frontiers in Psychology (Cognitive Science), 4(1006)
In particular, sections 10.1 and 11.2 was co-written with L. Smith, and section 10.3 was co-written with C. Moulin-Frier and M. Nguyen.
Forestier S, Oudeyer P-Y. 2017. A Unified Model of Speech and Tool Use Early Development. Proceedings of the 39th Annual Meeting of the Cognitive Science Society.
Citations of the work described here should refer to these articles.

10.1 Intrinsic motivation and spontaneous exploration in infants

Learning experiences do not passively "happen" to infants. Rather, infants' own activities create and select these experiences. Piaget (1952) described a pattern of infant activity that is highly illustrative of this point. He placed a rattle in a four-month-old infant's hands. As the infant moved the rattle, it would both come into sight and also make a noise, arousing and agitating the infant and causing more body motions, and thus causing the rattle to move into and out of sight and to make more noise. The infant had no prior knowledge of the rattle but discovered through activity the task and goal of rattle shaking. As the infant accidentally moved the rattle, and saw and heard the consequences, becoming captured by the activity and outcomes, the infant may be said to have gained intentional control over the shaking of the rattle and the goal of making noise (Thelen, 1994). This example -a body movement that leads to an interesting outcome and thus more activity and the re-experience and building of expectations about the outcome- may be foundational, not just to developmental process, but also to how evolution works through developmental process.

Infants' exploration of rattles are instances of action and learning motivated by curiosity and may reflect intrinsic motivations that select "interesting" sensorimotor activities for the pure pleasure of learning, the pleasire of gaining information and control about the world (Lowenstein, 1994; Gottlieb et al., 2013). However, what is interesting is history dependent. Once all the variations in rattle shaking are easily predicted, and all the outcomes expected, then playing with a rattle is not very interesting, as evident in the disinterest of older infants in rattles. What is interesting depends on what one knows and what one does not know.

Across many different fields, theorists have suggested that "interest" is engaged by what is just beyond current knowledge, neither too well known nor too far beyond what is understandable. This theoretical idea has been offered many times in psychology, through concepts like cognitive dissonance (Kagan, 1972), optimal incongruity (Hunt, 1965), intermediate novelty (Berlyne, 1960; Kidd, Piantadosi, Aslin, 2012) and optimal challenge (Csikszenthmihalyi, 1991). There have been several recent theoretical advances in these ideas in developmental robotics (Baldassarre and Mirolli, 2013; Gottlieb et al., 2013; Oudeyer et al., 2007), in models about the evolutionary origins of intrinsic motivation systems (Singh et al., 2010; Barto, 2013), and in recent findings in neuroscience linking intrinsic motivation with attention (Gottlieb et al., 2013), as well as in new formal models of infant visual attention (Kidd, Piantadosi, Aslin, 2012).

In general, learning in these curiosity driven activities progresses to yield an improvement of prediction or control over a repeated activity and thus a reduction of uncertainty (Friston, 2012, 2015; Oudeyer et al., 2007; Schmidhuber, 1991). Such intrinsically motivating activities have been called "progress niches" (Oudeyer et al., 2007): progress in learning in and for itself generates intrinsic rewards and an action selection system directly aims to maximize this reward. In reinforcement learning frameworks, this search for progress niches leads to spontaneous exploration (Gottlieb et al., 2013; Kidd et al, 2012; Oudeyer et al., 2007; Schmidhuber, 1991). In this view, progress in prediction or control is a primary driver (and accordingly, intrinsic rewards for learning progress/uncertainty reduction may be primary rewards). These theoretical advances lead to a definition of curiosity as an epistemic motivational mechanism that pushes an organism to explore activities for the primary sake of gaining information (as opposed to searching for information in service of achieving an external goal like finding food or shelter). Thus, activities and stimuli that arouse curiosity and stimulate exploration are those for which there is a particular relation between internal predictive models of these activities or stimuli and what is actually observed through experimentation. From a machine learning perspective, mechanisms of information seeking are called active learning, where the learner probabilistically and through its own activity selects experiences according to their potential for reducing uncertainty and for improving world models (Cohn et al., 1996; Oudeyer et al., 2007; Lopes and Oudeyer, 2010; Lopes and Montesano, 2014).

Such a motivational mechanism of curiosity will often be only one of several motivational mechanisms operating in any living being, and at any given time curiosity may interact, complement or conflict with other motivations such as motivation to satisfy physiological needs like hunger or physical and social contact with social peers. In theoretical models, information seeking mechanisms have been used either as an "exploration bonus" mechanism in service of efficient maximization of a task-specific reward, or as primary rewards driving models of curiosity-driven learning (Gottlieb et al., 2013).

Curiosity-driven learning through the search of information gain, is what fuels children's motivation to make sense of their environment, as if they were little scientists making informed choices of which are the informative experiments to perform in order to build better predictive theories about the world (Gopnik et al., 1999; Shulz, 2012; Oudeyer, 2018). Further than guiding exploration, intrinsic motivation also modulates brain mechanisms for learning, for example producing positive effects on memory (Stahl and Feigenson, 2015). Thus, these mechanisms are at the center of research in educational sciences, with studies of how the balance between intrinsic and extrinsic motivation can influence cognitive development (Bruner, 1962; Cameron and Pierce, 2002).

Brain circuits associated to spontaneous exploration and curiosity are still little known, but recent work has been converging towards the identification of neural mechanisms associated to intrinsic motivation (Kaplan et Oudeyer, 2007; Gottlieb et al., 2013; Gottlieb et al., 2014; Gottlieb and Oudeyer, 2018). For example, Panksepp identified in the brain a lateral hypothalamic corridor, starting from ventral tegmental area to nucleus accumbens, playing a central role in spontaneous exploration (Panksepp, 1998). This system is also located at a strategic location in dopaminergic circuits, which use multiple forms of phasic and tonic signals to influence the processing and response to novel stimuli or errors in reward predictions (Schultz, 1998; Hooks and Kalivas, 1994; Fiorillo, 2004). The degeneration of dopaminergic neurons in Parkinson's disease does not only provoke psychomotor impairments, but also decreases exploration behaviors and interest in cognitive tasks (Bernheimer et al., 1973). Reversely, artificial electric or chemical stimulation of the dopaminergic system triggers spontaneous exploration and curiosity in humans and animals (Panksepp, 1998).

Intrinsic motivation, interacting with several brain and behavioral structures, form a complex system and understanding it requires here again formal and operational modelling to complement these works in psychology and neuroscience. In particular, it appears crucial to formulate precisely concepts such as "optimal challenge" or "intermediate complexity", as they are key to define interest in such motivational systems. How could the brain measure in practice intermediate complexity? How could it exploit it to decide which experiments to make, which situations to explore? What is the long-term impact of intrinsic motivation on the structuration of sensorimotor, cognitive and social skills? In recent years, mathematical, computational and robotic models have approached these questions and have provided surprising ideas.

10.2 A robotic model of artificial curiosity

Developmental robotics is a research domain that uses robotic modelling to study the mechanisms by which embodied individuals can continuously acquire know-how and knowledge across their life span, either through self-exploration or through social learning (Weng et al., 2001; Oudeyer, 2011). In particular, it focuses on modelling several families of developmental constraints that guide and structure learning in the real world. Examples include the study of the role of morphological properties of bodies, maturation of sensorimotor organs which grow progressively, motor synergies that simplify high-dimensional motor control, attention and emotions in social interaction, imitation, cognitive biases that facilitate statistical inference, and motivational systems which is the topic we discuss here.

Several strands of computational and robotics models of spontaneous active exploration, serving the acquisition of new skills, were elaborated so far (Oudeyer, 2018). Some of these works take the purely engineering objective to build machines that can learn efficiently repertoires of complex tasks, anchoring themselves in the domain of statistical inference and machine learning (e.g. Fedorov, 1976; Schmidhuber, 1991, 2011; Kaplan and Oudeyer, 2003; Barto et al., 2004; Lehman and Stanley, 2008; Lopes et al., 2012; Baranes and Oudeyer, 2013; Martius et al., 2013; Bellemare et al., 2016). Other works were have targeted to model exploration mechanisms in the living, trying to help us understand better mechanisms of curiosity and development in humans (Oudeyer et al., 2007; Kaplan and Oudeyer, 2007; Schembri et al., 2007; Baldassarre, 2011).

The models we will now discuss, developped in particular with Frédéric Kaplan, Clément Moulin-Frier, and Sébastien Forestier, belong to the latter perspective. They were also among the first models of artificial curiosity to be implemented in robots, allowing progressive learning and development of new and diverse motor skills. As we will describe, several robot experiments illustrate how mechanisms of curiosity-driven exploration, dynamically interacting with learning, physical and social constraints, can self-organize developmental trajectories and in particular lead a learner to successively discover two important functionalities, object affordances and vocal interaction with its peers.

10.2.1 The Playground Experiment: discovering affordances

In the Playground Experiment, a quadruped learning robot (the learner) is placed on an infant play mat with a set of nearby objects and is joined by an “adult” robot (the teacher), see figure 10.1 (Oudeyer and Kaplan, 2006; Kaplan and Oudeyer, 2007b; Oudeyer et al., 2007). On the mat and near the learner are objects, and associated affordances, for discovery: an elephant (which can be bitten or grasped by the mouth), a hanging toy (which can be “bashed” or pushed with the leg). The robot teacher is pre-programmed to imitate the sound made by the learner when the learning robot looks to the teacher while vocalizing at the same time.

The learner is equipped with a repertoire of motor primitives parameterized by several continuous numbers that control movements of its legs, head and a simple simulated vocal tract. Each motor primitive is a dynamical system controlling various forms of actions: (a) turning the head in different directions; (b) opening and closing the mouth while crouching with varying strengths and timing; (c) rocking the leg with varying angles and speed; (d) vocalizing with varying pitches and lengths. These primitives can be combined to form a large continuous space of possible actions. Such motor primitives allow to constrain the movements that the robot can explore, providing a structure which will allow it to produce

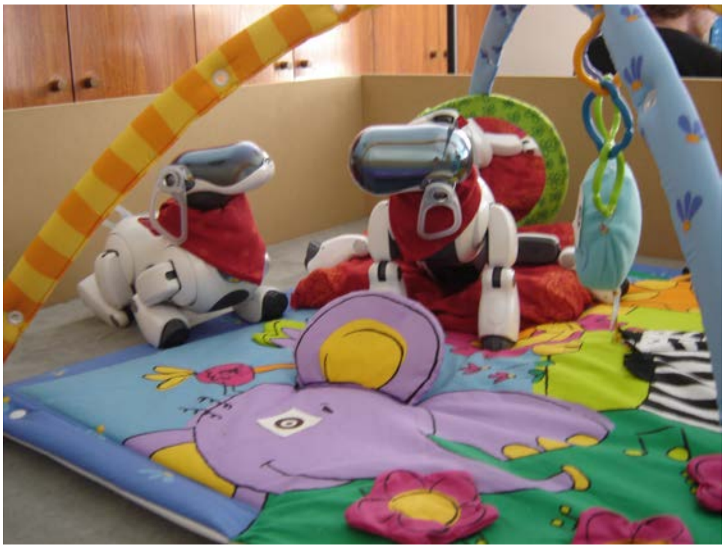


Figure 10.1: The Playground Experiment (Oudeyer and Kaplan, 2006; Oudeyer et al., 2007). The robot in the centre explores through adaptive babbling driven by a model of curiosity, and learns to predict the effects that its movements cause on its environment. He generates continuous movements based on setting the parameters of motor primitives in its repertoire, which serve as lego bricks that can be assembled and continuously varied. It can move its legs, its head, its jaw and produce vocalizations). It observes perceptual effects with a repertoire of sensori primitives, which allow it to measure movement, proprioceptive effects (e.g. whether he has grasped something in its mouth), or the characteristics of sounds it hears. Its cognitive architecture is presented on figure 10.2.

a diversity of controllable effects on its environment. They can be related to motor primitives such as central pattern generators (CPGs) observed in animals (Flash and Hochner, 2005). Similarly, sensory primitives allow the robot to detect visual movement, salient visual properties (visual patches with certain characteristics), proprioceptive touch in the mouth, and pitch and length of perceived sounds. For the robot, these motor and sensory primitives are initially black boxes and he has no knowledge about their semantics, effects or relations, and how the values of their parameters change their effects. However, the environment is such that it contains structures that the robot has to discover. For example, an object in front of him can be grasped with the mouth, which it can measure with its proprioceptive sensory primitive. An object on its left is too far to be grasped, but can be pushed in diverse ways with the leg, which the robot can observe if at the same time it looks in the direction of this object, turning its visual sensor in the appropriate direction.

The robot learns how to use and tune these primitives to produce various effects on its surrounding environment, and exploration is driven by the maximization of learning progress, by choosing physical experiences (experiments) that improve the quality of predictions of the consequences of its actions. What the robot will find interesting, and thus the notion of optimal complexity, corresponds here to the situations for which its errors decrease fastest: thus, it will avoid familiar and trivial situations, as well as situations which are too complex for it at this particular moment of cognitive development. As we will see, this will push the robot to progressively explore activities of increasing complexity.

This requires capabilities for learning, meta-cognition and action selection, as used in the R-IAC computational architecture on figure 10.2 (Oudeyer, Kaplan et al. 2007; Moulin-Frier et al., 2014). A prediction machine (M) learns to predict the consequences of actions taken by the robot in given sensory contexts. For example, this module might learn to predict (with a neural network) which visual movements or proprioceptive perceptions result from using a leg motor primitive with certain parameters. To do this, each time the robot makes an action and observes its consequences in a given context, it collects this data and use it to update an internal predictive model. The update of this predictive model consists in detecting statistical regularities in the data observed by the robot, allowing it to predict better in the future the consequences of similar actions through forms of interpolation or extrapolation.

On top of this lower-level learning module, a meta-cognitive module estimates the evolution of errors in prediction of M in various regions of the sensorimotor space. This module estimates how much errors decrease in predicting an action, for example, in predicting the consequence of a leg movement when this action is applied towards a particular area of the environment. These estimates of error reduction are used to compute the intrinsic reward from progress in learning. This reward is an internal quantity that is proportional to the decrease of prediction errors, and the maximization of this quantity is the goal of action selection within a computational reinforcement-learning architecture (Oudeyer and Kaplan, 2007; Oudeyer et al., 2007). Importantly, the action selection system chooses most often to explore activities where the estimated learning progress is high. However, this choice is probabilistic, which leaves the system open to learning in new areas and open to discovering other activities that may also yield progress in learning [2]. Since the sensorimotor flow does not come pre-segmented into activities and tasks, a system that seeks to maximize differences in learnability is also used to progressively categorize the sensorimotor space into regions. This categorization thereby models the incremental creation and refining of cognitive categories differentiating activities/tasks.

10.2.2 Self-organization of an ordered learning curriculum

In the Playground experiment, multiple experimental runs lead to two general results. First, one observes a self-organization of developmental trajectories, where an ordered learning curriculum spontaneously forms through active self-exploration. Second, if one computes statistics over many experiments, one discovers the formation of a mixture of regularities and diversities in the developmental patterns: some behavioural or cognitive structures appear often in a certain order, but some robots follow a quite different trajectory even when starting from the same mechanisms and environment (Oudeyer and Kaplan, 2006; Oudeyer et al., 2007). Figure 10.3 shows an example of the formation of several behavioral structures in one run of the experiment.

To illustrate how the intrinsically motivated exploration mechanism can automatically generate such ordered learning stages, let us first imagine a learner confronted with four categories of activities, as shown on figure 10.4. The practice of each of these four activities, which can be of varying difficulty, leads to different learning rates at different points in time (see the top curves, which show the evolution of prediction errors in each activity if the learner were to focus full-time and exclusively on each). If, however, the learner uses

[2] Technically the decision on how much time to spend on high learning progress activities and other activities is achieved using Multi-Armed Bandit algorithms for the so-called exploration/exploitation dilemma (Audibert et al., 2009).

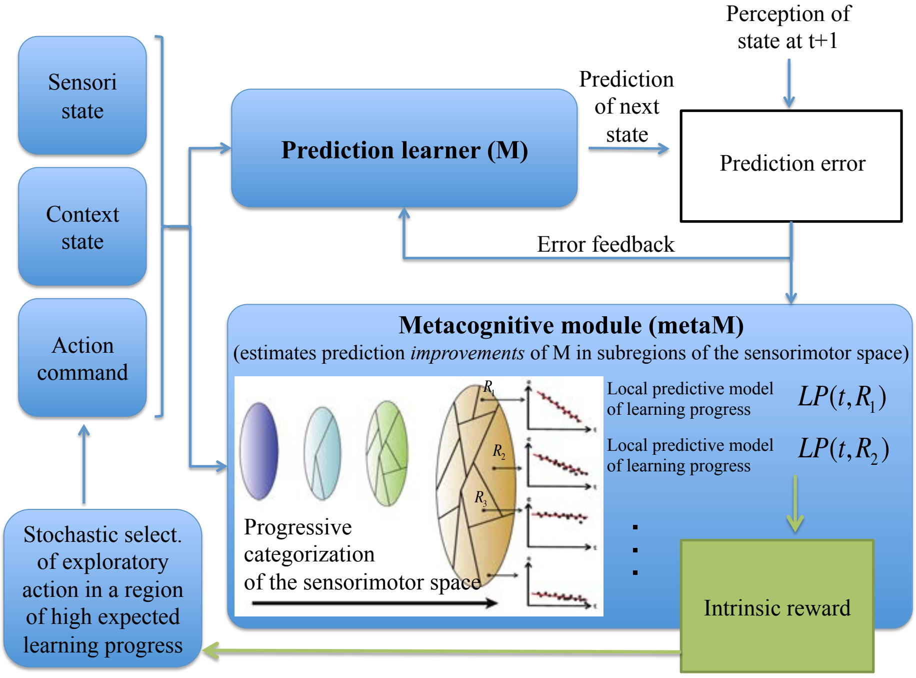

Sensori state
Context state
Action command
Prediction learner (M)
Prediction of next state
Perception of state at t+1
Prediction error
Error feedback
Metacognitive module (metaM)
(estimates prediction *improvements* of M in subregions of the sensorimotor space)
R_1
R_2
R_3
Progressive categorization of the sensorimotor space
Local predictive model of learning progress
$LP(t,R_1)$
Local predictive model of learning progress
$LP(t,R_2)$
Intrinsic reward
Stochastic select. of exploratory action in a region of high expected learning progress

Figure 10.2: IAC cognitive architecture for curiosity-driven learning (Oudeyer et al., 2007).

curiosity-driven exploration to decide what and when to practice by focusing on progress niches, it will avoid activities already predictable (curve 4) or too difficult to learn to predict (curve 1), in order to focus first on the activity with the fastest learning rate (curve 3) and eventually, when the latter starts to reach a plateau to switch to the second most promising learning situation (curve 2). Thus, such robots will show a regular developmental course - one that will be "universal" for learners with similar internal processes learning in similar environments. Embodied exploration driven by learning progress creates an organized exploratory strategy: the system systematically achieves these learning experiences in an order and does so because they yield (given the propensities of the learner and the physical world) different patterns of uncertainty reduction.

Now, let us come back to what happens in the Playground experiment and focus on the self-organization of such an ordered learning curriculum. In all of the runs, one observes the formation of structured developmental trajectories, where the robot explores objects and actions in a progressively more complex stage-like manner while acquiring autonomously diverse affordances and skills that can be reused later on and that change the learning progress in more complicated tasks. The following developmental sequence is typically observed:

1 In a first phase, the learner achieves unorganized body babbling;

2 In a second phase, after learning a first rough model and meta-model, the robot stops combining motor primitives, exploring them one by one, but each primitive is explored itself in a random manner;

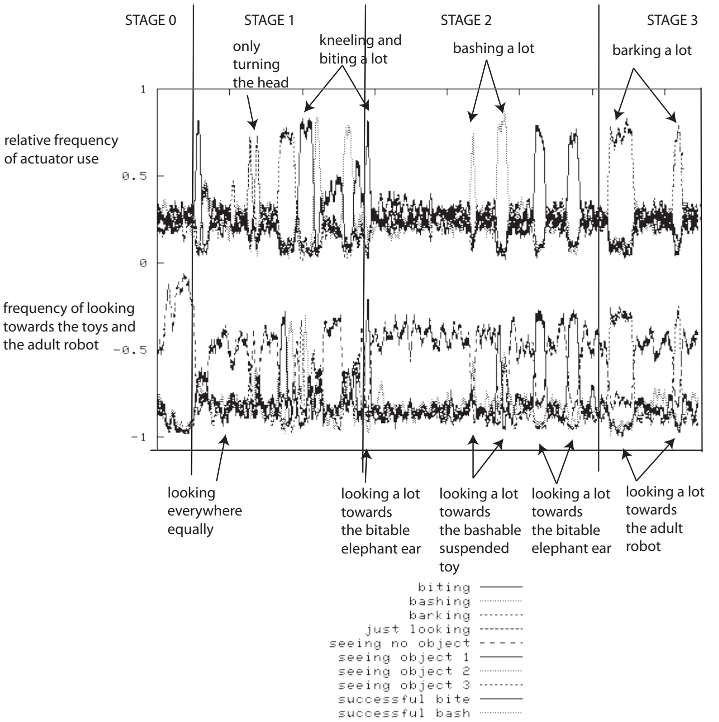


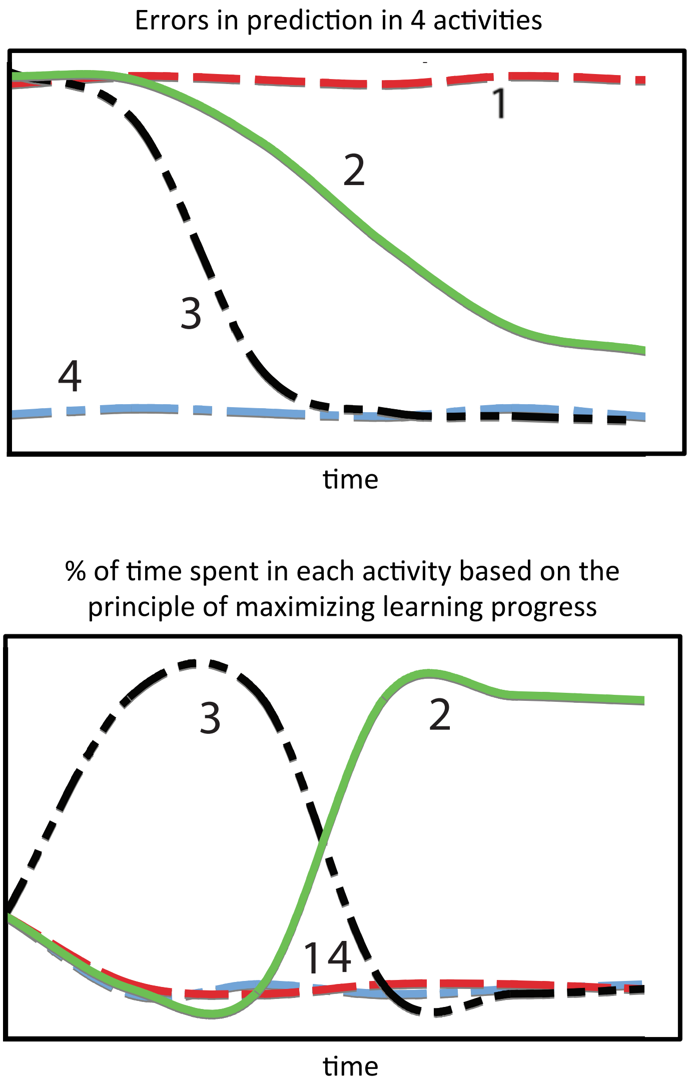
Figure 10.3: Measures of the evolution of the behavior of the developmental robot in the Playground Experiment. Several phases appear spontaneously as the result of the dynamical interaction between learning, curiosity-driven exploration, the body and the environment. These phases appear following an order characterized by an increase of the complexity of sensorimotor activities explored by the robot.

3 In a third phase, the learner begins to experiment with actions towards zones of its environment where the external observer knows there are objects (the robot is not provided with a representation of the concept of "object"), but in a non-affordant manner (e.g. it vocalizes at the non-responding elephant or tries to bash the teacher robot which is too far to be touched);

4 In a fourth phase, the learner now explores the affordances of different objects in the environment: typically focussing first on grasping movements with the elephant, then shifting to bashing movements with the hanging toy, and finally shifting to explorations of vocalizing towards the imitating teacher.

5 In the end, the learner has learnt sensorimotor affordances with several objects, as well as social affordances, and has mastered multiple skills. None of these specific objectives were pre-programmed. Instead, they self-organized through the dynamic interaction between curiosity-driven exploration, statistical inference, the properties of the body, and the properties of the environment.

These playground experiments do not simply simulate the acquisition of particular skills (such as batting at toys to make them swing or vocalizations) but simulate an ordered and


Errors in prediction in 4 activities

% of time spent in each activity based on the principle of maximizing learning progress

Figure 10.4: Illustration of a self-organized developmental sequence where the robot automatically identifies, categorizes and shifts from simple to mode complex learning experiences. Figure adapted with permission from (Kaplan and Oudeyer, 2007).

systematic developmental trajectory, with a universality and stage-like structure that may be mistakenly taken to indicate an internally-driven process of maturation. However, the trajectory is created through activity and through the general principle that sensorimotor experiences that reduce uncertainty in prediction are rewarding. In this way, developmental achievements can build on themselves without specific pre-programmed dependencies but nonetheless - like evolution itself - create structure (see Smith and Breazeal, 2007; and Smith 2013, for related findings and arguments).

10.2.3 Regularities and diversity in developmental trajectories

Because these are self-organizing developmental processes, they generate not only strong regularities but also diversity across individual developmental trajectories. For example, in most runs one observes successively unorganized body babbling, then focused exploration of head movements, then exploration of touching an object, then grasping an object, and finally vocalizing towards a peer robot (pre-programmed to imitate). This can be explained as gradual exploration of new progress niches, and those stages and their ordering can be viewed as a form of attractor in the space of developmental trajectories. Yet, with the same mechanism and same initial parameters, individual trajectories may invert stages, or even generate qualitatively different behaviours. This is due to stochasticity, to even small variability in the physical realities and to the fact that this developmental dynamic system has several attractors with more or less extended and strong domains of attraction (characterized by amplitude of learning progress). We see this diversity as a positive outcome since individual development is not identical across different individuals but is always, for each individual, unique in its own ways. This kind of approach, then, offers a way to understand individual differences as emergent in developmental process itself and makes clear how developmental process might vary across contexts, even with an identical learning mechanism.

10.2.4 Bootstrapping communication

A further result to be highlighted in the Playground experiment is the early development of vocal interaction. With a single generic mechanism, the robot both explores and learns how to manipulate objects and how to vocalize to trigger specific responses from a more mature partner (Oudeyer and Kaplan, 2006; Kaplan and Oudeyer, 2007a). Vocal babbling and language play have been shown to be key in infant language development; however, the motivation to engage in vocal play has often been associated with hypothesized language specific motivation. The Playground Experiment makes it possible to see how the exploration and learning of communicative behavior might be at least partially explained by general curiosity-driven exploration of the body affordances, as also suggested by Oller (2000).

Furthermore, the mechanism for incremental categorization used in this architecture allows the robot to distinguish sensorimotor activities based on their degree of learnability/predictability. This leads to the formation of internal abstract representations that grow as the robot explores its world. As argued in (Oudeyer et al., 2007) and in (Kaplan and Oudeyer, 2007), this makes it possible for the robot to discriminate stimuli corresponding to its own body - the "self"- (e.g. its hand moving in its visual field, very predictable and controllable), from stimuli corresponding to external physical objects (e.g. an object that can be pushed only in certain conditions, predictable but less than its own body), and from stimuli corresponding to special external entities: the "others", which are structured but less predictable and controllable than physical objects.

10.3 Self-organization of early vocal development: the onset of imitation

The Playground experiment shows how active exploration mechanisms driven by intrinsic motivation can self-organize global developmental trajectories that span multiple modalities,

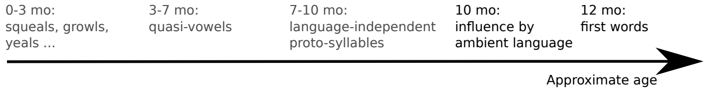


Figure 10.5: The first year of infant vocal development.

and sharing qualitative properties with the progressive structuration of development in infants. If one focuses on the details of what is happening in each modality in infants, one also observes complex pattern formation. For example, the development of arm reaching capabilities follows the so-called proximo-distal law, where infants tend to first explore using articulations close to the shoulder while freezing articulations close to the hand that are progressively freed (Berthier et al., 1999). Computational models of adaptive learning, based on principles similar to the models discussed so far, showed how such proximo-distal law can be formed as a side effect of learning mechanisms (Schlesinger et al., 2000; Stulp and Oudeyer, 2018). In what follows, we will focus on properties of early vocal development, showing that mechanisms of curiosity-driven learning, combined with the influence of social peers, can generate ordered vocal structures that closely match what is observed in infants.

Early on, babies seem to explore vocalizations as if it was a game in itself, as reported by Oller who cites two studies from the 19th century (Oller, 2000):

> "[At] three months were heard, for the first time, the loud and high crowing sounds, uttered by the child sponteaneously, [...] the child seemed to take pleasure in making sounds." (Sigismund, 1856)
>
> "[He] first made the sound *mm* spontaneously by blowing noisily with closed lips. This amused [him] and was a discovery for [him]."[3] (Taine, 1856)

Such play with her vocal tract, where the baby discovers the sounds she can make, echoes other forms of body play, such as exploration of arm movements or how he can touch, grasp, mouth or throw objects. It appears to be closely related to the concept of intrinsic motivation that we discussed in previous sections. Although spontaneous vocal exploration is an identified phenomenon, occurring in the early stages of infant development, the specific mechanisms of such exploration and the role of intrinsic motivation for the *structuration* of early vocal development had not received much attention until recently. As we will argue here, mechanisms of intrinsically motivated spontaneous exploration play an important role in speech acquisition, by driving the infant to follow a self-organized developmental sequence which will allow him to progressively learn to control his vocal tract. Let us first look at the properties of vocal development in infants.

10.3.1 Infant vocal development

Despite inter-individual variations in infant vocal development (Vihman et al., 1986), strong regularities in the global structuration of vocal development are identified (Oller, 2000, Kuhl, 2004). Figure10.5 schematizes this vocal development during the first year of infant. First,

[3] We have changed the gender of the subject to a male in this quotation, in order to follow the convention of the present article. Throughout this paper, we will use "he" for an infant, "she" for a caregiver (e.g. the mother) and "it" for a learning agent (the model).

until the age of approximately 3 months, an infant produces non-speech sounds like squeals, growls and yeals. During this period, he seems to learn to control infrastructural speech properties, e.g. phonation and primitive articulation. Then, from 3 to 7 months, he begins to produce vowel-like sounds (or quasi-vowels) while he probably learns to control his vocal tract resonances. At 7 months, canonical babbling emerges where well-timed sequences of proto-syllables are mastered. But it is only around the age of 10 months that infant vocal productions become more influenced by the ambient language, leading to first word productions around 1 year of age.

Two features of this developmental sketch are particularly salient.

- Infants seem to first play with their vocal tracts in a relatively language-independent way, and then are progressively influenced by the ambient speech sounds.

- In the initial phase, when sounds produced by their peers influence little their vocalizations, infants seem to learn skills of increasing complexity: normal phonation, then quasi-vowels and finally proto-syllables. According to Oller (2000), such a sequence displays a so-called natural, or logical hierarchy. For example, it is impossible to master quasi-vowel production without previously mastering normal phonation.

10.3.2 Models of speech development

Several computational models of speech development, where speech acquisition is organized along a developmental pathway, have been elaborated so far. They have shown how such stage-like organization can ease the acquisition of complex realistic speech skills.

The DIVA model (Guenther et al., 1998; Guenther, 2006), as well as Kröger's model (Kröger et al., 2009), propose architectures partly inspired by neurolinguistics. They involve two learning phases. The first one is analogous to infant babbling and corresponds to semi-random articulator movements producing auditory and somatosensory feedbacks. This is used to tune the correspondences between representation maps within a neural network. In the second phase, the vocal learner is presented with external speech sounds analogous to an ambient language and learns how to produce them adequately. The Elija model (Howard and Messum, 2011) also distinguishes several learning phases. In the first phase of exploration, the agent is driven by a reward function, including intrinsic rewards such as sound salience and diversity (which is related to models of curiosity-driven learning), as well as articulatory effort. Various parameterizations of this reward function allows the model to produce vocalizations in line with Oller's vocal developmental stages of infants. In a subsequent phase, the sounds produced by the model attract the attention of a caregiver, providing an external reinforcement signal. Other models also use a reinforcement signal, either from social reinforcement of human listeners (Warlaumont, 2013b), or based on sound saliency as a form of intrinsic reinforcement (Warlaumont, 2013a), and show how this can influence a spiking neural network to produce canonical syllables.

These computational models of speech acquisition pre-determine the global ordering and timing of learning experiences, which amounts to preprograming the developmental sequence (for example the switching between self-exploration and imitation/socially influence babbling). Understanding how a vocal developmental sequence can be formed is still a mystery to solve. As we will argue, mechanisms of active curiosity-driven learning, shown to self-organize learning curricula in the Playground experiment, may also be good candidates for explaining the structuration of vocal development.

Within such a perspective, Clement Moulin-Frier, Mai Nguyen and I have been studying how active exploration mechanisms, driven by a drive to improve predictability and controllability of one's own body, can lead to a specific organization of vocal development (Moulin-Frier and Oudeyer, 2012; Moulin-Frier et al., 2014). This model uses a similar curiosity-driven learning architecture than in the Playground Experiment, but with three differences. First, it is applied to the exploration of a complex model of the vocal tract and auditory system, reusing the articulatory synthesizer of the DIVA model (Guenther, 2006). Second, it is based on a diffrent form of curiosity-driven learning, called intrinsically motivated goal exploration, and detailed in section 10.3.3. Third, it combines intrinsically motivated goal exploration and social guidance, as detailed in section 10.3.4.

10.3.3 Intrinsically motivated learning by autonomous goal setting

Curiosity-driven learning can take many different forms (Oudeyer and Kaplan, 2007b). In the Playground Experiment, the learning robot decides which sensorimotor experiment to make by selecting motor commands that are expected to provide high improvement in predicting sensory consequences. From the reinforcement learning perspective, this amounts to modeling curiosity as a learning mechanism driven by the maximization of intrinsic rewards that measure forms of information gain, uncertainty or learning progress associated to taking action a in a given state s.

Another very important form of curiosity-driven learning is "intrinsically motivated goal exploration" (Baranes and Oudeyer, 2013; Forestier et al., 2017). With this mechanism, used in the model below, the robot directly imagines and selects goals (defined as target properties of behavioural and state trajectories), as well as their associated cost function to measure the adequacy between what they target and what they do. After selecting a goal, the robot then uses its current internal world model to infer the corresponding motor program, and tries to improve its competences towards the self-selected goal by experimenting around its current best known motor program. In turn, new observations collected through such phases of self-exploration are used to improve the internal world model. Autonomous exploration by directly and actively self-generating and self-selecting goals was shown to be particulary efficient for learning diverse repertoires of skills and world models in real-world high-dimensional sensorimotor spaces (Baranes and Oudeyer, 2013; Rolf and Steil, 2013).

In the speech development model below, goals are defined as target auditory trajectories that should pass through a number of acoustic key points in a given order. The internal world model is encoded as a mixture of gaussians encoding the joint distribution between motor programs and auditory trajectory outcomes, and is updated incrementally as exploration unfolds. The selection of these goals is based on evaluating expected progress in producing a given auditory goal, which is a form of learning progress called "competence progress" (Oudeyer and Kaplan, 2007b). In the model below, a meta-learning system incrementally tracks the learning progress across the space of goals, and enables to predict the expected progress associated to selecting a goal.

Importantly, intrinsically motivated goal exploration processes (IMGEPs) define a novel family of learning mechanisms that is an alternative to the reinforcement learning framework. There are three properties of IMGEPs that differentiate these mechanisms from reinforcement learning when integrated together:

- First, in IMGEPs learning happens without external supervision from engineers, and especially without an externally defined reward measuring a scalar number to be maximized and corresponding to a specific task (e.g. maximizing a score in a game).

- Second, IMGEPs approaches learning as driven by intrinsically generated goals and associated cost functions (as oppposed to intrinsically motivated reinforcement learning mechanisms that do not include the concept of goal).

- Third, goals in IMGEPs are in the general case defined as a set of properties being targeted over a behavioral/state trajectory, and associated to a self-defined cost function (a key property being that the learner knows the cost function algorithm as it generates it). Thus, intrinsically generated goals are more than black-box reward functions (parts of the RL litterature uses the term "goal" to denote the objective encoded by a reward function, which is a much more restrictive concept of goal than in IMGEPs). Also, as IMGEPs goals are defined as global properties of behavioural/state trajectories, the cost functions measuring the adequacy between a trajectory and a goal cannot be defined as only depending on the current state and action. Thus, in the general case, IMGEPs goals cannot be framed within the markovian framework used in reinforcement learning.

Beyond providing a conceptual approach to autonomous learning, IMGEPs afford several mechanisms that enable very efficient learning of diverse skills and world models in large spaces. As goals and cost functions are internally generated, the learner has several opportunities. It can choose learning goals which are most adapted to its current skills through automatic curriculum learning, actively controlling the growth of complexity of its self-generated learning situations (Baranes and Oudeyer, 2013). Also, as he has access to the "code" defining its goals, it can leverage its knowledge of these goals semantics (e.g. by differentiating goals and cost functions). Furthermore, and crucially, this framework affords hindsight learning (Forestier et al., 2017): when the learner selects and targets a goal, generating a trajectory, it can learn concurrently to improve its skills to reach other goals. Indeed, it can a posteriori imagine other goals, compute the cost of the trajectory for these goals, and update its world model relative to these other imagined goals. For example, in the model below, a machine targeting an auditory goal defined in terms of passing through sounds ['b', 'a', 'b', 'i'] in this order, might produce a motor commands that fails importantly toward this goal (e.g. may produce ['d', 'a', 'p', 'o']) , but learn at the same time how to produce other goals such as ['d', 'a'], ['p','o'] or ['d','a','p','o'].

10.3.4 Interaction between autonomous goal setting and social guidance

The speech development model outlined below also considers how vocal exploration by autonomous goal setting can be influenced by the vocalization of social peers. Here, vocalizations of social peers are stored in a short-term memory and can be selected as auditory goals if the learner estimates that such imitation trials can provide competence progress at this stage of its development. Thus, the learner not only decides which part of the sensorimotor space to explore when it is in self-exploration mode, but it also decides when to imitate adult vocalizations based on a measure of learning progress. Such a learning strategy has been called *strategic learning*, where the learner makes active choices to decide *What*, *How*, *When* and *with/from Whom* to learn (Lopes and Oudeyer, 2012; Nguyen and Oudeyer, 2013). This is a generalization of the concept of active learning as used in machine learning, and extends intrinsically motivated learning to the selection of learning strategies, including the possibility to select the interaction with peers as a strategy to gain information and improve one's own world model.

In the following sections, we will present an outline of this model, as well as the results of several experiments. These experiments will show how such a mechanism can explain the adaptive transition from vocal self-exploration with little influence from the speech environment, to a later stage where vocal exploration becomes influenced by vocalizations of peers. Within the initial self-exploration phase, we will see that a sequence of vocal production stages self-organizes, and shares properties with infant data: the vocal learner first discovers how to control phonation, then focuses on vocal variations of unarticulated sounds, and finally automatically discovers and focuses on babbling with articulated proto-syllables. As the vocal learner becomes more proficient at producing complex sounds, imitating vocalizations of peers starts to provide high learning progress explaining an automatic shift from self-exploration to vocal imitation.

10.3.5 Model

In this section, we describe the models that we use for the vocal tract and auditory signals. We describe the learning of the internal model of the sensorimotor mapping, and the intrinsic motivation mechanism which allows the learner to decide adaptively which vocalization to experiment at given moments during its development, and whether to do so through self-exploration or through imitation of external sounds.

Sensorimotor system

Vocal Tract and Auditory System This computational model uses the articulatory synthesizer of the DIVA model described in (Guenther, 2006)[4], which is itself based on Maeda's model (Maeda, 1989). The model corresponds to a computational approximation of the general speech production principles illustrated in figure 10.6.

The model receives 13 articulatory parameters as input. The first 10 are from a principal component analysis (PCA) performed on sagittal contours of images of the vocal tract of a human speaker, allowing to reconstruct the sagittal contour of the vocal tract from a 10-dimensional vector. The effect of the 10 articulatory parameters from the PCA on the vocal tract shape is displayed figure 10.7. In this study, only use the 7 first parameters are used (the effect of the others on the vocal tract shape is negligible), fixing the 3 last in the neutral position (value 0 in the software). Through an area function, associating sections of the vocal tract with their respective area, the model can compute the 3 first formants of the produced signal if phonation occurs. Phonation is controlled through the 3 last parameters: glottal pressure controlling the intensity of the signal (from quiet to loud), voicing controlling the voice (from voiceless to voiced) and pitch controlling the tone (from low-pitched to high-pitched). It is then able to compute the formants of the signal (among other auditory and somato-sensory features) through the area function. Here, only the glottal pressure and voicing parameters are used. In addition to the 7 articulatory parameters from the PCA, a vocal command is therefore defined by a 9-dimensional vector. From the vocal command, the synthesizer computes the auditory and somatosensory consequences of the motor command, thus approximating the speech production principles of figure 10.6.

[4] Available online at `http://www.bu.edu/speechlab/software/diva-source-code`

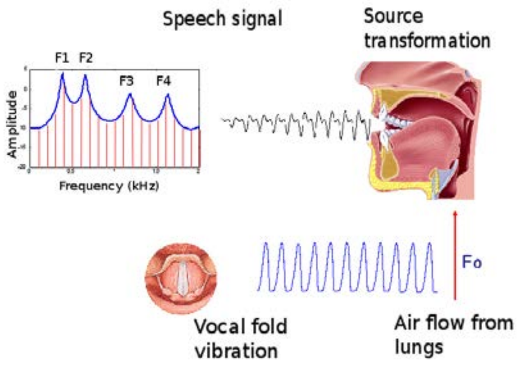


Figure 10.6: Speech production general principles. The vocal fold vibration by the lung air flow provides a source signal: a complex sound wave with fundamental frequency F_0. According to the vocal tract shape, acting as a resonator, the harmonics of the source fundamental frequency are selectively amplified or faded. The local maxima of the resulting spectrum are the formants, ordered from the lower to the higher frequencies. They belong to the major features of speech perception and are used in the model of curiosity-driven speech development (design adapted from L-J. Boë and S. Jacopin).

On the perception side, the model uses the first two formants of the signal, $F1$ and $F2$, approximately scaled between -1 and 1. A third parameter I is also defined, which measures the intensity (or phonation level) of the auditory outcome. I is supposed to be 0 when the agent perceives no sound, and 1 when it perceives a sound. Technically, $I = 1$ if and only if two conditions are checked: (1) both pressure and voicing parameters are above a fixed threshold (null value) and (2) the vocal tract is not closed (i.e. the area function is positive everywhere). In human speech indeed, the formants are not measurable when phonation is under a certain threshold. This is modelled by setting that when $I = 0$, the formants do not exist anymore and are set to 0. This drastic simplification is yet arguable in term of realism, but what the model encodes here is the fact that no control of the formant values can be learnt when no phonation occurs.

Dynamical properties Speech production and perception are dynamical processes and the principles of figure 10.6 have to be extended with this respect. Humans control their vocal tract by variations in muscle activations during a vocalization, modulating the produced sound in a complex way. Closure or opening movements during a particular vocalization, coupled with variations in phonation level, are able to generate a wide variety of modulated sounds. A vocalization is here defined as a trajectory of the 9 motor parameters over time, lasting 800 milliseconds, from which the articulatory synthesizer is able to compute the corresponding trajectories in the auditory space (i.e. trajectories in the 3-dimensional space of $F1$, $F2$ and I). The agent is able to control this trajectory by setting 2 commands for each articulator: one from 0 to $250ms$, the other one from 250 to $800ms$. Then, the motor system is modeled as an overdamped spring-mass system driven by a second-order dynamical equation[5]. Thus, the agent's policy for a vocalization is defined by two vectors

[5] $\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2(x - m) = 0$, where x is a motor parameter, and m is the command for that motor parameter. ζ is set to 1.01, ensuring that the system is overdamped (no oscillation), and ω_0 to $\frac{2\pi}{0.8}$ (0.8

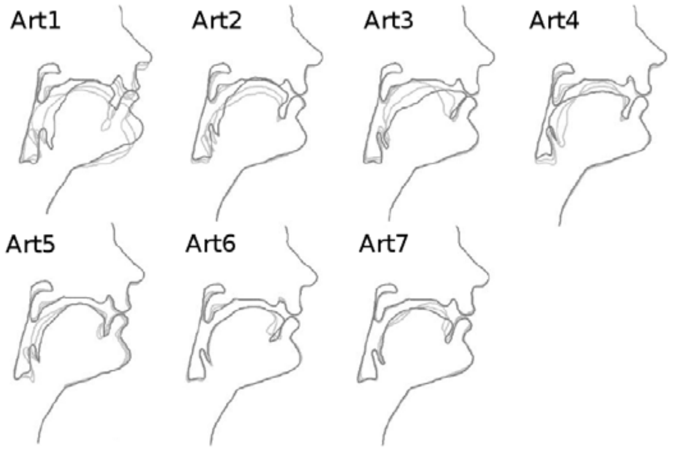


Figure 10.7: Articulatory dimensions controlling vocal tract shape (10 dimensions, from left to right and top to bottom), adapted from the documentation of the DIVA source code. Each subplot shows a sagittal contour of the vocal tract, where we can identify the nose and the lips on the right side. Bold contours correspond to a positive value of the articulatory parameter, the two thin contours are for a null (neutral position) and negative values. These dimensions globally correspond to the dimensions of movements of the human vocal tract articulators. For example, Art_1 mainly controls the jaw height, whereas Art_3 rather controls the tongue front-back position.

m_1 and m_2 (one for each command) of 9 real values each (one for each motor parameter). The policy space is 18-dimensional. The first command is applied for the beginning of the vocalization to $250ms$, the second one from $250ms$ to $800ms$.

Figure 10.8A illustrates the process by showing a typical syllabic vocalization. In this illustrative example, the controlled articulators are the first and third articulators of figure 10.7 (roughly controlling the jaw height and the tongue front/back dimensions), as well as pressure and voicing. The two last ones are set to 0.5 and 0.7 respectively, for both commands, to allow phonation to occur. The "jaw parameter" ($art1$ on the figure) is set to 2.0 (jaw closed) for the first command and to -3.0 for the second one (jaw open). We observe that these commands, quite far from the neutral position, are not completely reached by the motor system. This is due to the particular dynamics of the system, defined with ζ and ω_0 in the dynamical system. For the third articulator ($art3$), the commands are both at 2.0. We observe that, whereas the value 2.0 cannot be achieved completely at $250ms$, it can however be reached before the end of the vocalization.

This motor system implies interaction between the two commands, i.e. a form of co-articulation. Indeed, a given motor configuration may sometimes be harder to reach if it is set as the first command, because time allocated to reach the first command is less than for the second command. Reversely, some movements may be harder to control in the second command because the final articulator positions will depend both on the first and the second commands (e.g., it is harder to reach the value -3.0 for the second command if the first command is set to 2.0, than if the first command is set to -3.0, as seen in the example of

being the duration of the vocalization in seconds)

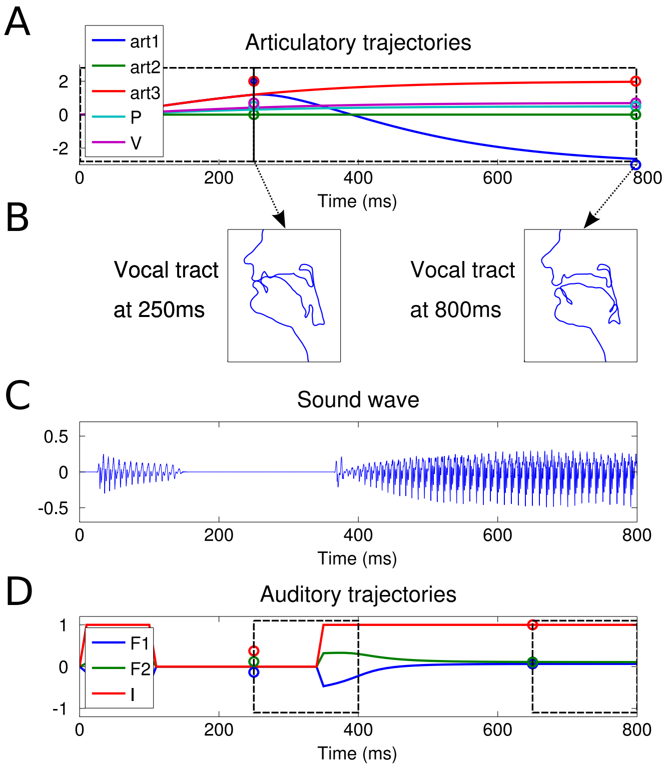


Figure 10.8: An illustrative vocalization example. A) Articularory trajectories of 5 articulators during the 800ms of the vocalization (4 articulators, from *art*4 to *art*7 are not plotted for the sake of readability but display the same trajectory as *art*2). Circles at 250 and 800ms represents the values of the first and second commands, respectively, for each trajectory. The first commands are active from 0 to 250ms and second ones from 250 to 800ms, as represented by dotted black boxes. The trajectories are computed by a second order dynamical equation encoding a spring-mass system, starting in a neutral position (all articulators set to 0). B) Resulting vocal tract shapes at the end of each command, i.e. at 250 and 800ms. Each subplot displays a sagittal view with the nose and the lips on the left side. The tongue is therefore to the right of the lower lip. C) Sound wave resulting from the vocalization. D) Trajectories of the 3 auditory parameters, the intensity I and the two first formants $F1$ and $F2$. Dotted black boxes represent the two perception time windows. The agent perceives the mean value of the auditory parameters in each time window, represented by the circles at 250 and 650ms.

figure 10.8).

These characteristics are the results of modeling speech production as a damped spring-mass system, which is a common practice in the litterature (Markey, 1994, Boersma,1998; Howard and Messum, 2011).

Figure 10.8B shows the resulting vocal tract shape at the end of the 2 commands (i.e. at $250ms$ and at $800ms$). We observe that the vocal tract is closed at the end of the first command, open at the end of the second one.

Figure 10.8C shows the resulting sound. We observe that there is no sound during vocal tract closure.

Figure 10.8D shows the resulting trajectories of auditory parameters. In our experiments, we model the auditory perception of the agent of its own vocalization as the mean value of each parameter I, F1 and F2 in two different time windows lasting $150ms$: the first one from 250 to $400ms$, the second one from 650 to $800ms$. The auditory representation of a vocalization is therefore a 6-dimensional vector $(I(1), I(2), F1(1), F1(2), F2(1), F2(2))$. Perceived auditory values are represented by circles on figure 10.8D. Note that the agent does not have any perception of what happens before $250ms$, and that $I(1)$ and $I(2)$ can take continuous values in $[0, 1]$ due to the averaging in a given perception time window. We will refer to the perceived "phone" of a given command for the perception occurring around the end of that command, although such an association will not be assumed in the internal sensorimotor model of the agent. Indeed, this sensorimotor system has the interesting property that the perceptions in both time windows depend on both motor commands. In the example of figure 10.8, the perception for the first command, i.e. the mean auditory values between 250 and $400ms$, would not be the same if the second motor command did not cause the vocal tract opening.

Vocalization classification For the purpose of analyzing the experiments presented below, three types of phones are defined according to the value of I for a given command. In this description, we use common concepts like vowels or consonants to make an analogy with the human types of phones, although this analogy is limited.

- Those where $I > 0.9$: , i.e. phonation occurs during almost all the $150ms$ of perception around the end of the command. We call them *Vowels* (V).
- Those where $I < 0.1$, i.e. there is almost no phonation during the $150ms$ of perception around the end of the command. We call them *None* (N).
- Those where $0.1 < I < 0.9$, i.e. phonation occurs partially during the 0.15s of perception around the end of the command. This means that the phonation level I has switched during that period. This can be due either to a closure or opening of the vocal tract, or to variations in the pressure and voicing parameters. We call them *Consonants* (C), although they are sometimes more comparable to a sort of prosody (when due to a variation in the phonation level).

This classification will be used as a tool for the analysis of the results in section 10.3.6, but is never known by the agent (which only has access to the values of I, $F1$ and $F2$).

Thus, each vocalization produced by the agent, belongs to the combination of 2 of these 3 types (because a vocalization corresponds to 2 commands), i.e. there are $3^2 = 9$ types of vocalizations: VV, VN, VC, NV, NN, NC, CV, CN, CC.

Then, we suggest to group these 9 types into 3 classes.

- The class *No Phonation* contains only NN: the agent has not produced an audible sound. This is due either to the fact the pressure and voicing motor variables have never been sufficiently high (not both positive, as explained in the description of the motor system) during the two $150ms$ perception periods, or that the vocal tract was totally closed.
- The class *Unarticulated* contains VN, NV, CN, NC: the vocalization is not well-formed. Either the first or the second command produces a phone of type *None* ($I < 0.1$, see above).
- The class *Articulated* contains CV, VC, VV and CC: the vocalization is well-formed, in the sense that there is no *None* phone. Phonation is modulated in most cases (i.e. except in the rare case where the two commands of a VV are very similar). Note that according to the definition of *consonants*, phonation necessarily occurs in both the perception time windows.

It is important to note that the auditory values of these vocalization classes span subspaces of increasing complexity. Indeed, whereas various articulatory configurations belong to the *No Phonation* class, their associated auditory values are always null, inducing a 0-dimensional auditory subspace (i.e. a point). Regarding the *Unarticulated* class, the associated auditory values span a 3-dimensional subspace because at least one command produces a phone of type *None* (i.e. the corresponding auditory values are null). Finally, in the *Articulated* classes, the auditory values span the entire 6-dimensional auditory space. These properties will have important consequences for the learning of a sensorimotor model by the agent, as we will see.

Internal sensorimotor model

[6]

During its life time, the agent iteratively updates an internal sensorimotor model by observing the auditory results of its vocal experiments. We denote motor commands M and sensory perceptions S. We call $f : M \to S$ the unknown function defining the physical properties of the environment (including the agent's body). When the agent produces a motor command $m \in M$, it then perceives $s = f(m) \in S$, modulo an environmental noise and sensorimotor constraints. In the sensorimotor system defined in the previous section, M is 18-dimensional and S is 6-dimensional. f corresponds to the transformation defined in section 10.3.5 and illustrated figure 10.8, and has a Gaussian noise with a standard deviation of 0.01. By collecting (m, s) pairs through vocal experiments, the agent learns the joint probability distribution defined over the entire sensorimotor space SM (therefore 24-dimensional). This distribution is encoded in a Gaussian Mixture Model (GMM) of 28 components, i.e. a weighted sum of 28 multivariate normal distributions[7]. Let us note G_{SM} this GMM. It is learnt using an online version of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) proposed by Calinon (Calinon, 2009) where incoming data are considered incrementally. Each update is executed once each sm_step (=400) vocalizations are collected. G_{SM} is thus refined incrementally during the agent life, updating each time a

[6] The sensorimotor internal model and aspects of the intrinsic motivation system used in this model were initially described in in a more general context where the goal was to compare various exploration strategies in (Moulin-Frier and Oudeyer, 2013a,b)

[7] We empirically chose a number of components which is a suitable trade-off between learning capacity and computational complexity.

number *sm_step* of new (m, s) pairs are collected. Moreover, we adapted this online version of EM to introduce a *learning rate* parameter α which decreases logarithmically from 0.1 to 0.01 over time. α allows to set the relative weight of the new learning data with respect to the old ones.

This GMM internal model is used to solve the inverse problem of inferring motor commands $m \in M$ that allow the learner to reach a given auditory goal $s_g \in S$. From this sensorimotor model G_{SM}, the agent can compute the distribution of the motor variables knowing a given auditory goal to reach s_g, noted $G_{SM}(M \mid s_g)$. This is done by Bayesian inference on the joint distribution, and results in a new GMM over the motor variables M (see e.g. (Calinon, 2009)), from which the agent can sample configurations in M.

The sensorimotor system we specified in the previous section involves a 24-dimensional sensorimotor space (18 articularory dimensions and 6 auditory ones). As we have already noted, the three vocalization classes (*No Phonation, Unarticulated* and *Articulated*) span subspaces of the 6-dimensional auditory space with increasing dimensionality. Learning an inverse model using GMMs with a fixed number of Gaussians is harder, i.e. requires more sensorimotor experiments, as the spanned auditory subspace is of higher dimensionality. Thus, learning an inverse model to produce *No Phonation* requires fewer learning data than learning an inverse model to produce various *Articulated* vocalizations, as the range of sensory effect is much larger in the second case.

Intrinsically motivated exploration of dynamic auditory goals

In order to provide training data to the sensorimotor model, the agent autonomously and adaptively decides which vocal experiments to make. The key idea is to self-generate and choose goals for which the learner predicts that experiments to reach these goals will lead to maximal competence progress. Below, we will present two series of experiments considering two variants of this exploration mechanism. A first variant involves self-exploration without considering vocalizations of social peers, whereas the second variant allows the learner to actively decide when to try to imitate the vocalizations of its peers.

The specific model we use in the first series of experiments (see section 10.3.6) is a probabilistic version of the SAGG-RIAC architecture (Baranes and Oudeyer, 2013). As explained above, this architecture was itself an evolution of the model of curiosity-driven exploration used in the Playground Experiment and presented above. It combines two principles: 1) autonomous goal exploration; 2) self-selection of goals based on the maximization of empirically measured learning progress. In practice, the learner self-generates its own auditory goals in the sensory space S. One goal is here a sequence of two auditory targets encoded in a 6-dimensional vector $s_g = (I(1), I(2), F1(1), F1(2), F2(1), F2(2))$ (see section 10.3.5). For each goal, it uses the current sensorimotor estimation to infer a motor program $m \in M$ in order to reach that goal. Through the sensorimotor system, this produces a vocalization and the agent perceives the auditory outcome $s \in S$, hence a new (m, s) training data. Goals are selected stochastically so as to maximize the expected competence progress (i.e. the learner is interested in goals where it predicts it can improve maximally its competence to reach them at a particular moment of its development). This allows the learner to avoid spending too much time on unreachable or trivial goals, and progressively explore self-generated goals/tasks of increasing complexity. As a consequence, the learner self-explores and learns only sub-parts of the sensorimotor space that are sufficient for reachable goals: this allows to leverage the redundancy of these spaces by building dense tubes of learning data only where it is necessary for control.

We define the competence c associated to a particular experiment (m, s) to reach the goal s_g as $c = comp(s_g, s) = e^{|s_g - s|}$ This measure is in $[0, 1]$ and exponentially increases towards 1 when the Euclidean distance between the goal and the actual realization $s = f(m) + \epsilon$ tends to 0.

The measure of competence progress uses another GMM, G_{IM}, learnt using the classical version of EM on the recent goals and their associated competences. This GMM provides an interest distribution $G_{IM}(S)$ used to sample goals in the auditory space S maximizing the competence progress in the recent sensorimotor experiments of the agent. This was firstly formalized in Moulin-Frier and Oudeyer, 2013a,b. Figure 10.9 provides a graphical explanation of the process.

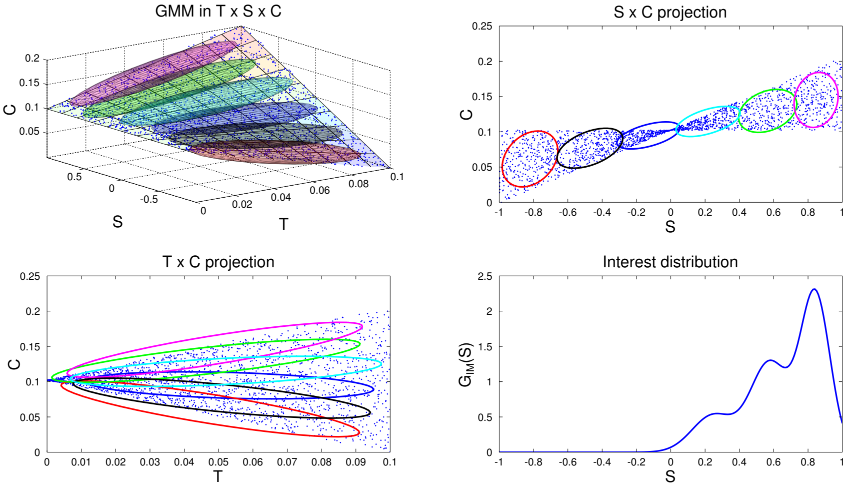


Figure 10.9: Illustration of how the "interestingness" of auditory goals is computed. Top-left: the recent history of competences of the agent, corresponding to blue points in the space $T \times S \times C$, where T is the space of recent time indexes (in $\mathbb{R}^+$), S the space of recently chosen goals s_g (mono-dimensional in this toy example) and C the space of recent competences of reaching those goals (in $\mathbb{R}^+$). For the sake of the illustration, the competence variations over time are here hand-defined (surf surface) and proportional to the values in S (increases for positive values, decreases for negative values). A GMM of 6 components, G_{IM}, is trained to to learn the joint distribution over $T \times S \times C$, represented by the six 3D ellipses. Projections of these ellipses are shown in 2D spaces $S \times C$ and $T \times C$ in the top-right and bottom-left plots. To reflect the competence progress in this dataset, we then bias the weight of each Gaussian to favor those which display a higher competence progress, that is measured as the covariance between time and competence for each Gaussian (in the example the purple ellipse shows the higher covariance in the bottom-left plot). Gaussians with a negative covariance between time T and competence C (blue, black and red ellipses) are weighted with a negligible factor, such that they do not contribute to the mixture. Using Bayesian inference in this biased GMM, the distribution of interestingness $G_{IM}(S)$ over the goal space S is computed, favoring regions of S displaying the highest competence progress (bottom-right).

To summarize, the agent possesses the following abilities:

- Producing a complex vocalization, sequencing two motor commands interpolated in a dynamical system. It is encoded by a 18-dimensional motor configuration $m \in M$.
- Perceiving the 6-dimensional auditory consequence $s = f(m) + \epsilon \in S$, computed by an articularory synthesizer. f is unknown to the agent.
- Iteratively learning a sensorimotor model from lots of (m, s) pairs it collects by vocalizing through time. It is encoded in a GMM G_{SM} over the 24-dimensional sensorimotor space $M \times S$.
- Controling its vocal tract to achieve a particular goal s_g. This is done by computing $G_{SM}(M \mid s_g)$, the distribution over the motor space M knowing a goal to achieve s_g.
- Actively choosing goals to reach in the sensory space S by learning an interest model G_{IM} in the recent history of experiences. By sampling in the interest distribution $G_{IM}(S)$, the agent favors goals in regions of S which maximizes the competence progress.

This agent is thus able to act at two different levels. At a high level, it chooses auditory goals to reach according to its interest model G_{IM} maximizing the competence progress. At a lower level, it attempts to reach those goals using Bayesian inference over its sensorimotor model G_{SM}, and incrementally refines this latter with its new experiences. The combination of both levels results in a self-exploration algorithm (**Algorithm 1**).

Algorithm 1 Self-exploration with active goal babbling (stochastic SAGG-RIAC architecture).

```
1: initialise $G_{SM}$ and $G_{IM}$
2: while true do
3:     $s_g \sim G_{IM}(S)$
4:     $m \sim G_{SM}(M \mid s_g)$
5:     $s = f(m) + \epsilon$
6:     $c = comp(s_g, s)$
7:     $update(G_{SM}, (m, s))$
8:     $update(G_{IM}, (s_g, c))$
9: end while
```

The agent starts in line 1 with no experience in vocalizing. Both GMMs have to be initialized in order to be used. To do this, the agent acquires a first set of (m, s) pairs, by sampling in M around the neutral values of the articulators (see figure 10.7). Regarding the pressure and voicing motor parameters, the neutral value is at -0.25, which leads to *no phonation* (recall that both these parameters have to be positive for phonation to occur, section 10.3.5). This models the fact that the agent does not phonate in its neutral configuration, and has at least to raise the pressure and voicing parameters to be able do do it. The agent then executes this first set of motor configurations (mostly not phonatory), observes the sensory consequences, and initialises G_{SM} with the corresponding (m, s) pairs using incremental EM. G_{IM} is initialised by setting the interest distribution $G_{IM}(S)$ to the distributions of the sounds it just produced with this first set of experiences. Thus, at the first iteration of the algorithm, the agent tries to achieve auditory goals corresponding to the sounds it produced during the initialisation phase. Then, in the subsequent iterations, the interest distribution $G_{IM}(S)$ reflects the competence progress measure, and is computed as explained above.

Line 3, the agent thus selects stochastically $s_g \in S$ with high interest values. Then it uses $G_{SM}(M \mid s_g)$ to sample a vocalization $m \in M$ to reach s_g (line 4). The execution of m will actually produce an auditory outcome s (line 5), and a competence measure to reach the goal, $c = comp(s_g, s)$, is computed (line 6). This allows it to update the sensorimotor model G_{SM} with the new (m, s) pairs (line 7). Finally, it updates the interest model G_{IM} (line 8) with the competence c to reach s_g

Active imitation system

In language acquisition and vocalization, the social environment plays naturally an important role. Thus the model considers an active speech learner that not only can self-explore its sensorimotor space, but can also learn by imitation. In a second series of experiments (described below), the model is extended by integrating the mechanism described in the previous section in the SGIM-ACTS architecture, which has been proposed in Nguyen and Oudeyer, 2012.

We consider here that the learning agent can use one of two learning strategies, that are selected adaptively:

- explore autonomously with intrinsically motivated goal babbling, as described previously,
- or explore with imitation learning. There are several forms of imitation, such as mimicry, in which the learner copies the policies of others without an appreciation of their purpose, or emulation, where the observer witnesses someone producing an outcome, but then employs its own policy repertoire to reproduce the outcome (Call and Carpenter, 2002; Lopes et al., 2010). As the learner a priori can not observe the vocal tract of the demonstrator, it can only emulate the demonstrator by trying to reproduce the auditory outcome observed, by using its own means, finding its own policy to reproduce the outcome. The model considers here that the the social peer has a finite set of vocalizations, and every time the learner chooses to learn by social guidance, it chooses at random an auditory goal within this set to emulate.

The learner can monitor the competence progress resulting from using each of the strategies. This measure is used to decide which strategy is the best progress niche at a given moment: a strategy is chosen with a probability directly depending on its associated expected competence progress. Thus, competence progress is used at two hierarchical levels of active learning, forming what is called strategic learning (Lopes and Oudeyer, 2012): at the higher-level, it is used to decide when to explore autonomously, and when to imitate; at the lower-level, if self-exploration is selected, it is used to decide which goal to self-explore (as in the previous model). Since competence progress is a non-stationary measure and is continuously re-evaluated, the individual *learns* to choose both the strategy $str \in \{autonomous\_exploration, social\_guidance\}$ and the auditory goals $s_g \in S$ to target, by choosing which combination enables highest competence progress.

For the particular implementation of SGIM-ACTS of this paper, we use the same formalism and implementation as in **Algorithm 1** and consider that the strategy is another choice made by the agent. This leads to **Algorithm 2**, where the interest model G_{IM} now learns an interest distribution as in section 10.3.5. The difference is that the space of interest is now the union of the strategy space $\{autonomous\_exploration, social\_guidance\}$ and the auditory space S. We call $StrS$ this new space $StrS = \{autonomous\_exploration, social\_guidance\} \times S$. Hence G_{IM} is a distribution over $StrS$ (**Algorithm 2**, line 3). If the self-exploration strategy is chosen ($str = autonomous\_exploration$), the agent acts as in Algorithm 1. If the social guidance strategy is chosen ($str = social\_guidance$, line 4), the learner then emulates an auditory demonstration $s_g \in S$ chosen randomly among the demonstration set of adult sounds (line 5), overwriting s_g of line 3. It then uses its sensorimotor model G_{SM} to choose a vocalization $m \in M$ to reach s_g, by drawing according to the distribution $G_{SM}(M \mid s_g)$ (line 7), as in the self-exploration strategy. The execution of m will produce an auditory outcome s (line 8), from which it updates its models G_{IM} and G_{SM} (lines 10 and 11).

Algorithm 2 Strategic active exploration (active goal babbling and imitation with stochastic SGIM-ACTS architecture).

```
 1: Initialise $G_{SM}$ and $G_{IM}$
 2: while true do
 3:     $(str, s_g) \sim G_{IM}(StrS)$
 4:     if $(str = social\_guidance)$ then
 5:         $s_g \leftarrow$ random auditory demonstration from the ambient language
 6:     end if
 7:     $m \sim G_{SM}(M \,| s_g)$
 8:     $s = f(m) + \epsilon$
 9:     $c = comp(s_g, s)$
10:     $update(G_{SM}, (m, s))$
11:     $update(G_{IM}, (str, s_g, c))$
12: end while
```

Thus, this new exploration algorithm is augmented with yet another level of learning, allowing to choose between different exploration strategies. This strategy choice moreover uses the same mechanism as the choice of auditory goals, by means of the interest model G_{IM}.

10.3.6 Experiments: transition from self-exploration to imitation

We first present experiments where individuals in the model learn in a pure self-exploration mode (**Algorithm 1**), without any social environment or sounds to imitate. In a second series of experiments, an auditory environment is introduced to study the influence of ambient language (**Algorithm 2**)[8].

[8] In these experiments, the model does not consider that infants and adults have different vocal tracts, and thus that the pitch and formant of vocalizations are different. In addition, this model does not study the social interaction aspect of the teacher and in particular how the behavior of the adult in response to the learner behavior can influence speech learning. Adressing the difficult matching problem has been addressed in other models, in particular by considering the dynamics of social interaction, such as in the work of Howard and Messum (2014) and Miura, Yoshikawa and Asada (2012).

Emergence of developmental sequences in autonomous vocal exploration

The first series of experiments involves nine independent simulations of **Algorithm 1** with the same parameters but different random seeds, of 240.000 vocalizations each[9]. Most of these nine simulations display the formation of a developmental sequence, as we will see. Before describing the regularities and variations observed in this set of simulations, let us first analyse a particular one where the developmental sequence is clearly observable. Figure 10.10 exhibits such a simulation. We observe three clear developmental stages, i.e three relatively homogeneous phases with rather sharp transitions. These stages are not preprogrammed, but emerge from the interaction of the vocal productions of the sensorimotor system, learning within the sensorimotor model, and the active choice of goals by intrinsically motivated active exploration. First (until $\simeq 30.000$ vocalizations), the agent produces mainly motor commands which results in *no phonation* or in *unarticulated* vocalizations (in the sense of the classes defined section 10.3.5). Second (until $\simeq 150.000$ vocalizations), phonation almost always occurs, but the vocalizations are mostly *unarticulated*. Third, it produces mainly *articulated* vocalizations.

The visualisation of the developmental sequence of the 9 independent simulations is provided in figure 10.11. This figure shows important interindividual variations whereas initial conditions are statistically similar due to initialisation in line 1 of **Algorithm 1**. These variations can be understood through the interaction of the sensorimotor system f, the internal sensorimotor model G_{SM} and the interest model G_{IM}, resulting in a complex dynamical system where observed developmental sequences are particular attractors (see e.g. van Geer, 1991; Smith, 1993). Moreover the sensorimotor and the interest models are probabilistic, thus inducing a non-negligible source of variability all along a particular simulation. Another factor is that using an online learning process on a GMM can result in a sort of forgetting, leading sometimes to the re-exploration of previously learnt parts of the sensorimotor space[10]. However, the sequence *No phonation* → *Unarticulated* → *Articulated* appears as a global tendency, as shown in **Table 10.1**. We observe that despite variations, most simulations begin with a mix of *no phonation* and *unarticulated* vocalizations, then mainly produce *unarticulated* vocalizations, and often end up with *articulated* vocalizations. An analogy can be made with human phonological systems, which are all different in the details but display strong statistical tendencies (Maddieson, 1984; Schwartz et al., 1997b).

This suggests that the agent explores its sensorimotor space by producing vocalizations of increasing complexity. The class *no phonation* is indeed the easiest to learn to produce for two reasons: the rest positions of the pressure and voicing motor parameters do not allow phonation (both around -0.25 at the initialisation of the agent, line 1 of **Algorithm 1**) ; and there is no variations on the formant values, which makes the control task trivial as soon as the agent has a bit of experience. There is more to learn with *unarticulated* vocalizations, where formant values are varying in at least one part of the vocalization, and still more with *articulated* ones where they are varying in both parts (for the first and second command).

Figure 10.12 shows what happens in the particular simulation of figure 10.10 in more details.

This developmental sequence is divided into 3 stages, I, II and III, stages being separated by vertical dark lines on figure 10.12, identical on each subplot (stage boundaries are the

[9] Each simulation involves several hours of computing on a desktop computer, due to the complexity of Bayesian inference and update procedures.

[10] This is why the simulations are limited to 240.000 vocalizations each, in order to avoid this unwanted effect of forgetting. However, the fact that the system is able to adaptively re-explore sensorimotor regions that have been forgotten is an interesting feature of curiosity-driven learning.

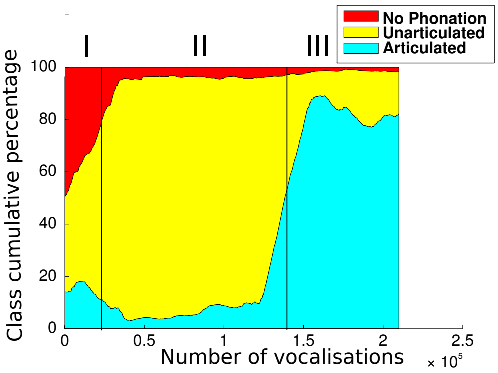


Figure 10.10: Self-organization of vocal developmental stages. At each time step t (x-axis), the percentage of each vocalization class between t and $t + 30.000$ is plotted (y-axis), in a cumulative manner (sum to 100%). Vocalization classes are defined in section 10.3.5. Roman numerals shows three distinct developmental stages. I: mainly no phonation or unarticulated vocalizations. II: mainly unarticulated. III: mainly articulated. The boundaries between these stages are not preprogrammed and are here manually set by the authors, looking at sharp transitions between relatively homogeneous phases.

Types of sounds produced	Stage I	Stage II	Stage III	Stage IV
No phonation-Unarticulated	**7**	0	2	0
Unarticulated	0	**7**	0	3
Articulated	0	2	**4**	0
Other	2	0	1	0

Table 10.1: Count of vocalization stages in the 9 simulations of the supplementary data. The “types of sounds produced” (first column of the table) correspond to the most prominent class in a given stage, where stages are manually set, looking at sharp transitions between relatively homogeneous phases. These developmental stages are therefore subjective to a certain extent, in the sense that another observer could have set different ones (but hopefully also would observe major structural changes). “No phonation-Unarticulated” means a mix between *No phonation* and *Unarticulated* classes (as defined in section 10.3.5 in that stage. A number x in a cell means *this type of vocalizations (row) appears x times at the n^{th} stage of development (column) in the set of 9 simulations.* Two to four developmental stages were identified in each simulation, explaining why the “Stage I” and “Stage II” columns sum up to 9 (the total number of simulations), but not the “Stage III” and “Stage IV” columns.

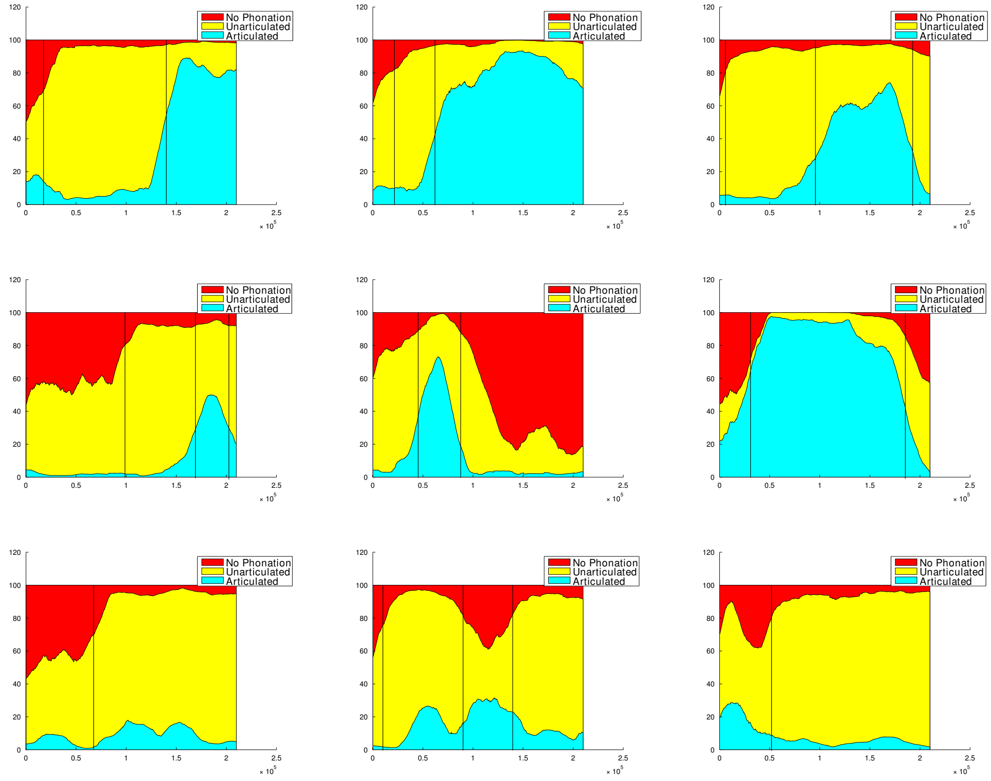


Figure 10.11: Developmental sequences emerging from the 9 simulations for the experiment described in section 10.3.6. Each subplot follows the same convention as in **Figure 10.10**. The simulations have been ordered, also in a subjective manner, from those which display a clear developmental sequence of the type *No phonation* → *Unarticulated* → *Articulated* to those less organized (from left to right, then top to bottom).

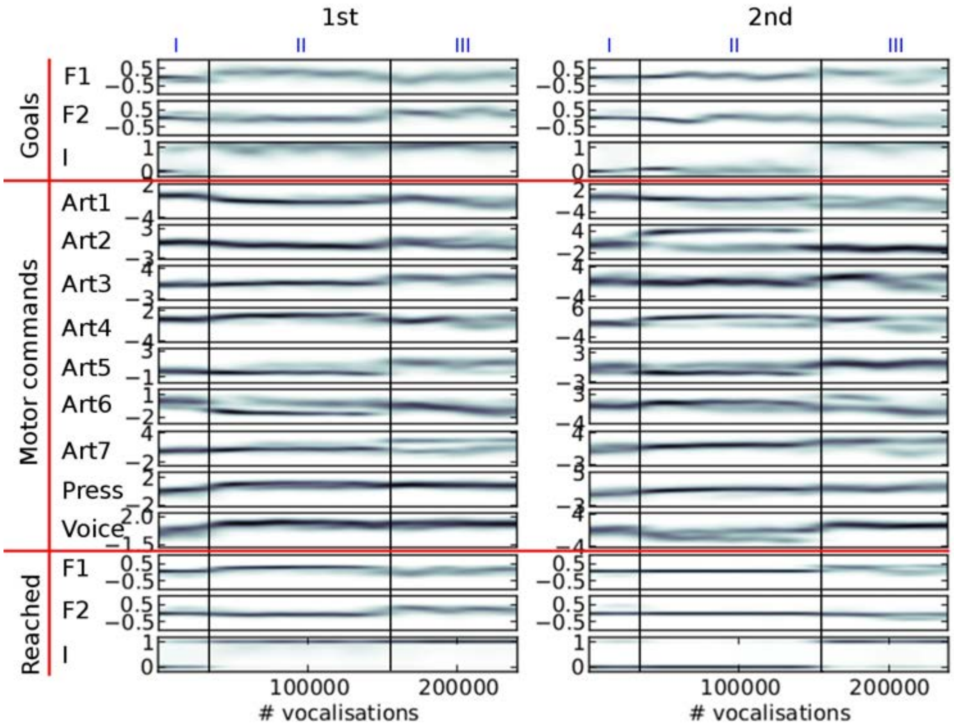


Figure 10.12: Evolution of the distribution of auditory goals, motor commands and sounds actually produced over the life time of a vocal agent (the same agent as in figure 10.10). The variables are in three groups (horizontal red lines): the goals chosen by the agent in line 3 of **Algorithm 1** (top group), the motor commands it inferred to reach the goals using its inverse model in line 4 (middle group), and the actual perceptions resulting from the motor commands through the synthesizer in line 5 (bottom group). There are two columns (1st and 2nd), because of the sequential nature of vocalizations (two motor commands per vocalization). Each subplot shows the density of the values taken by each parameter (y-axis) over the life time of the agent (x-axis, in number of vocalizations since the start). It is computed using an histogram on the data (with 100 bins per axis), on which we apply a 3-bins wide Gaussian filter. The darker the color, the denser the data: e.g. the auditory parameter I actually reached by the second command ($I(2)$, last row in 'Reached", 2nd column), especially takes values around 0 (y-axis) until approximately 150.000th vocalization (x-axis), then it takes rather values around 1. The three developmental stages of figure 10.10 are reported at the top.

same than in figure 10.10).

In stage I, until approximately 30.000 vocalizations, the agent produces mainly *no phonation* and *unarticulated* vocalizations. We observe that the agent set goals for $I(1)$ either around 0, either around 1, whereas the goals for $I(2)$ stay around 0 (last row in "Goals"). By trying to achieve these goals, the agent progressively refines its sensorimotor model and progresses by raising the values of the pressure and voicing motor parameter in the first command (two last rows of the section "Motor commands", 1st column). Other articulators remain around the neutral position (value 0). The agent is learning to phonate. The percentages of vocalization belonging to each vocalization class is provided **Table 10.2**.

NN	CN	NC	VN	NV	VV	CV	VC	CC
45.3 %	13.4 %	0.6 %	18.9 %	4.5 %	9.9 %	6.6 %	0.7 %	0.2 %

Table 10.2: Percentage of vocalization classes produced in stage I of the studied developmental sequence.

Then, in stage II, from 30.000 to approximately 150.000 vocalizations, the agent is mainly interested in producing vocalizations which begin with a *Vowels* ($I(1) > 0.9$, see the definition of phone types in section 10.3.5) and finish with a *None* ($I(2) < 0.1$). During this stage, it learns to produce relatively high $F1(1)$ values, in particular by decreasing the $Art_1(1)$ parameter (approximately controlling the jaw height, see figure 10.7). Regarding the second command, although the agent self-generates various goals for $F1(2)$ and $F2(2)$, and produces various motor commands to try to reach them, the sound produced mostly corresponds to a *None* ($I(2) = 0$, and therefore $F1(2) = F2(2) = 0$). This is due both to the negative value of the voicing parameter (last row in "Motor commands", second column), and to the fact that the vocal tract often ends in a closed configuration due to the poor quality of the sensorimotor model in this region (because phonation occurs very rarely for the second command, leaving the agent without an adequate learning set). During this stage, the agent explores a limited part of the sensorimotor space both in time (sound only for the first command) and space (around the neutral position), until it finally manages to phonate more globally at the end of this stage. This could be correlated to the acquisition of articulated vocalizations. The percentages of vocalization belonging to each vocalization class is provided in **Table 10.3**.

NN	CN	NC	VN	NV	VV	CV	VC	CC
4.0 %	26.9 %	0.1 %	62.2 %	0.1 %	3.4 %	0.5 %	2.5 %	0.2 %

Table 10.3: Percentage of vocalization classes produced in stage II of the studied developmental sequence.

Finally, in stage III (until 150.000 to the end), phonation almost always occurs during both the perception time windows, producing VV vocalizations (see I densities, both for goals and reached values). This is much harder to achieve for two reasons: firstly because there is a need to control a sequence of 2 articulators movement in order to reach two formant values in sequence (i.e. $F1(1)$, $F1(2)$, $F2(1)$, $F2(2)$) instead of one in the previous stage (the second command leading to no sound), and secondly because the position of the articulators reached for the second command also depends on the position of the articulators reached for the first one (a kind of coarticulation due to the dynamical properties of the motor system). We observe that the range of values explored in the sensorimotor space is larger than for the previous stage (both in motor and auditory spaces). The percentages of vocalizations belonging to each vocalization class is provided in **Table 10.4**.

NN	CN	NC	VN	NV	VV	CV	VC	CC
1.6 %	3.7 %	0.1 %	12.1 %	0.8 %	67.5 %	6.5 %	6.8 %	0.8 %

Table 10.4: Percentage of vocalization classes produced in stage III of the studied developmental sequence.

Influence of the auditory environment

In a second set of experiments, a social environment is integrated and provides a set of adult vocalizations. As explained in section 10.3.5, the learner has an additional choice: it can explore autonomously, or emulate the adult vocalizations. An "ambient language" is here modeled as a set of two speech sounds. To make it coherent with human language and the learning process observed in development, we chose speech-like sounds, typically vowel or consonant-vowel sounds. In terms of our sensorimotor descriptions, the adult sounds correspond to $I1$ with low values and $I2$ with high values. Figure 10.13 shows such vocalizations corresponding to those used by Teacher 1 in figure 10.14.

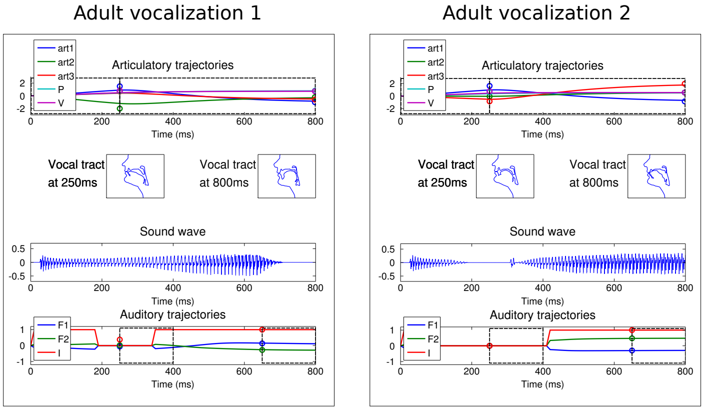


Figure 10.13: The two vocalizations of the adult Teacher 1 used in figure 10.14, with the same convention as in figure 10.8 .

Figure 10.14 shows a significant evolution in the agent's vocalizations. In the early stage of its development, it can only make a few sounds. Most sounds correspond to small values of $I1(2), F1(1), F1(2), F2(1)$ and $F2(2)$, as in the first developmental stage of the previous experiment (see **Table 10.2** and figure 10.12). Therefore the agent is not able to reproduce the ambient sounds of its environment. In contrast, in later periods of its development, its vocalizations cover a wider range of sounds, with notably $I(1)$ and $I(2)$ both positive, which means it now produces more articulated sounds. The development of vocalizations for a self-exploring agent in the last section showed that it progressively was able to produce articulated vocalizations, which we observed in several simulations at the

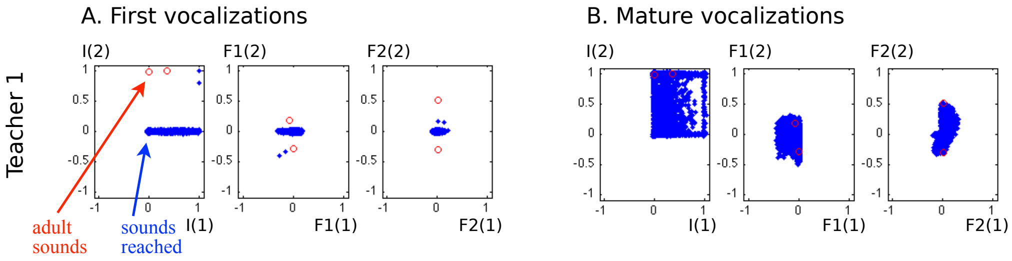


Figure 10.14: Vocalizations of the learning agent in the early and mature stages of vocal development. A) All auditory outcomes s produced by the agent in its early stage of vocalization are represented by blue dots in the 6-dimensional space of the auditory outcomes. The adult sounds are represented in red circles. The actually produced auditory outcomes only cover a small area of physically possible auditory outcomes, and correspond mostly to $I(2) = 0$, which represent vowel-consonant or consonant-consonant types of syllables. B) The auditory outcomes produced by the infant in its mature stage of vocalization cover a much larger area of auditory outcomes and extend in particular over areas in which vocalizations of the social peer are located.

end of its development. This effect has been reinforced by the environment: with articulated vocalizations to emulate, it produces this class more regularly.

Another important result is that mature vocalizations can now reproduce the ambient sounds of the environment: the regions of the sounds produced by the learner (blue dots) overlap the teacher's demonstrations (red circles). It seems that, during the first vocalizations, the agent cannot emulate the ambient sounds because they are too far away from its possible productions, and thus it can hardly make any progress and approach these demonstrations. Figure 10.15 confirms this interpretation. In the beginning, the agent makes no progress with emulation, and it is only around $t = 450$ that it makes progress with the emulation strategy. At that point, as we can see in figure 10.16, it uses equally both strategies. This enables the agent to make considerable progress from $t = 450$ to $t = 800$. Indeed, once its mastery improves and the set of sounds it can produce increases, it then increasingly emulates ambient sounds. Once it manages to emulate the ambient sounds well, and thus its competence progress decreases, it uses less the emulation strategy and more the self-exploration strategy.

To analyse better this emulation phenomenon and assess the influence of the ambient language, we run the same experiment with different acoustic environments. We used two other sets of speech sound demonstrations from simulated peers, and analysed the auditory productions of the agent in figure 10.17. The first property that can be noted is that in the early phase of the vocal exploration (figure 10.17. A and C), the auditory productions of the two agents are alike, and do not depend on the speech environment. On the contrary, the mature vocalizations vary with respect to the speech environment. With Teacher 1, the productions have their values $F2(1)$ and $F2(2)$ along the axis formed by the demonstration (figure 10.14. A, last column). Comparatively, Teacher 2's speech sounds have different $F1(1), F1(2), F2(1)$ and $F2(2)$. As represented in figure 10.17. B, the two speech sounds now differ mainly by their $F1(1)$ (instead of $F1(2)$) and in their subspace $(F2(1), F2(2))$ the speech sounds have approximately rotated from those of Teacher 1. The produced auditory

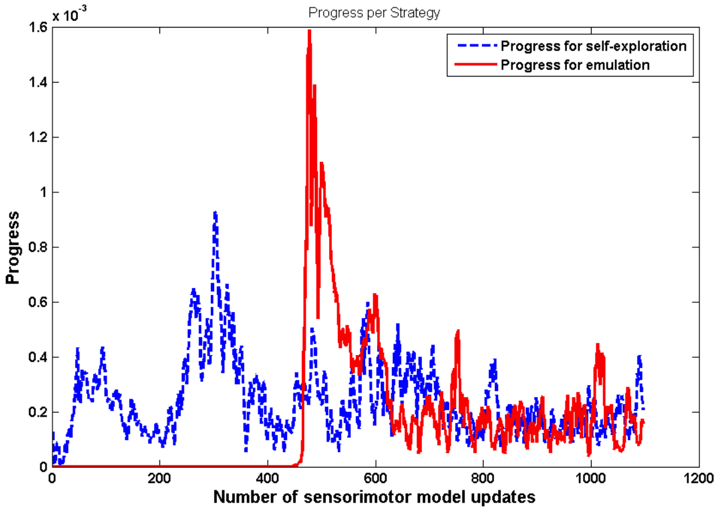


Figure 10.15: Progress made by each strategy with respect to the number of updates of the sensorimotor model G_{SM}. These values have been smoothened over a window of 100 updates. For $t < 450$, the agent makes no progress using emulation strategy. After $t = 450$, both strategies enable the agent to make progress.

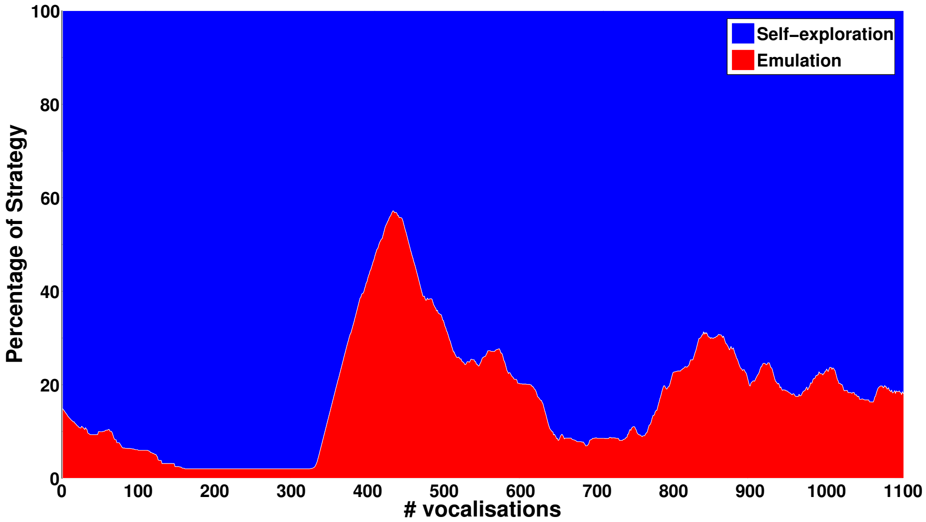


Figure 10.16: Percentage of times each strategy is chosen with respect to the number of updates of the sensorimotor model G_{SM}. These values have been smoothened over a window of 100 updates. For $t < 450$, the agent mainly uses self-exploration strategy. When its knowledge enables it to make progress in emulation, it chooses emulation strategy until it can emulate the ambient sounds well (and its competence progress decreases).

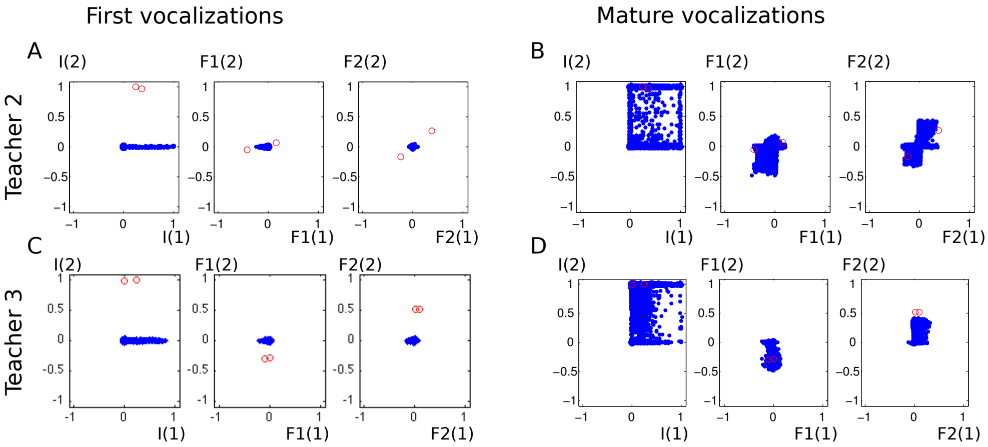


Figure 10.17: Vocalizations of the learning agent in the early and mature stage of vocalization in two different speech environments (Teacher 2 and Teacher 3). A and C) All auditory outcomes produced by the vocal learner in its early stage of vocal development are represented by blue dots in the 6-dimensional space of the auditory outcomes. The sounds of the environment are represented in red circles. The auditory outcomes only cover a small area, and do not depend on the speech environment. B and D) The auditory outcomes produced by the infant in its mature stage of vocal development cover a larger area of auditory outcome, which depend on the speech environment.

outcomes of the learner look like they have changed in the same way. Whereas the reached space (blue area) seemed to be along axis F1(2) and F2(2) and little on F1(1) or F2(1) for Teacher 1, it has extended its exploration along F1(2) and F2(2) for Teacher 2. With Teacher 3, the demonstrations are more localised in the auditory space, with $F1(1) < 0$ and $F2(2) > 0$. The effect we observe in figure 10.17. D is that the exploration is more localised too: the explored space is more oriented toward areas where $F1(1) < 0$ and $F2(2) > 0$. Thus, these three examples strongly suggest a progressive influence of the auditory environment, in the sense that the first vocalizations in figure 10.14 and 10.17 are very similar, whereas we observe a clear influence of the speech environment on the produced vocalizations in later stages.

Altogether, the results of this second series of experiments provide a computational support to the hypothesis that the progressive influence of the ambient language observed in infant vocalizations can be driven by an intrinsic motivation to maximize competence progress. At early developmental stages, attempts to imitate adult vocalizations are certainly largely unsuccessful because basic speech principles, such as phonation, are not yet mastered. In this case, focusing on simpler goals probably yields better progress niches than an imitative behavior. While they are progressively mastered, the interest in these goals decreases whereas the ability to imitate adult vocalizations increases. Imitation thus becomes a new progress niche to explore.

Chapter 11

Discovering Language as a Tool to Influence Others

[1]

Beyond the role of intrinsically motivated exploration for bootstrapping vocal development and imitation, what leads children to discover linguistic communication? How can they discover that certain speech sounds can be used as a currency to manipulate attention and action of social peers, and achieve joint tasks with them?

In this chapter, we describe a model studying further the potential role of intrinsically motivated exploration in these discoveries. In particular, we will show how autonomous exploration can enable individuals to discover progressively their body, how to interact with tools, and how to use speech sounds as tools to influence social peers.

11.1 Discovery of tool use and language: the role of intrinsic motivation

In the model described in the previous section, individuals could only explore how to produce sounds with their vocal tract. They were not grounded in a multimodal environment with objects affording other forms of sensorimotor activities, and the only goals they could generate were auditory goals. Thus, linguistic communication, defined as the ability to produce signals to manipulate the attention and action of others to achieve a goal, was not part of their world's possibilities. In the Playground Experiment, robots were situated in a richer environment where they had the ability to explore a variety of actions ranging from body movement to sound production with their vocal tract, and could similarly perceive the outcomes of their actions across several visuo-auditory-proprioceptive modalities. Towards the end of most runs of this experiment, curiosity-driven learning robots engaged in forms

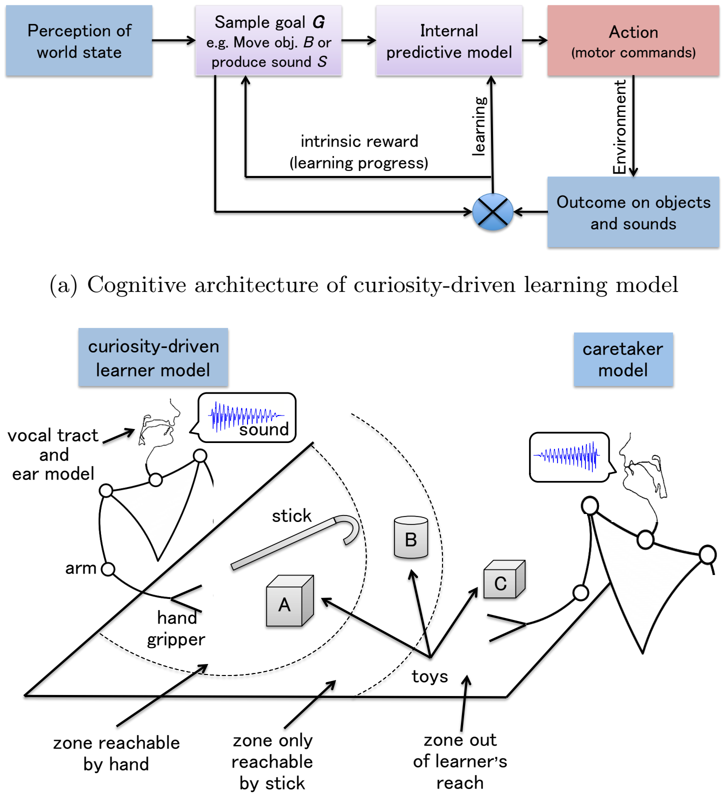
[1] Several parts of this chapter reuse text or graphs, with permissions, from the following articles co-authored with Linda Smith (section 11.2) and Sébastien Forestier (section 11.1):
Oudeyer, P-Y. and Smith. L. (2016) How Evolution may work through Curiosity-driven Developmental Process, Topics in Cognitive Science.
Forestier S, Oudeyer P-Y. 2017. A Unified Model of Speech and Tool Use Early Development. Proceedings of the 39th Annual Meeting of the Cognitive Science Society.
Citations of the work described here should refer to these articles.

of simple vocal interaction with a social peer. However, the cognitive architecture used in the Playground Experiment did not enable individuals to form and pursue goals. They were restricted to exploring questions such as "what would happen if I did action A?". As a consequence, linguistic communication was also beyond the reach of discoveries they could make.

A recent model has explored what may happen if a learner, equipped with a mechanism of intrinsically motivated goal exploration, could explore a rich body and environment affording object manipulation (including tools), speech production and perception, and a social peer producing contingent vocalizations. This mechanism and environment were presented in (Forestier and Oudeyer, 2017), and are illustrated in figure 11.1. In this environment, a simulated robot placed in front of a table with several objects, and a social peer model is situated on the other side of the table. The robot can produce actions either through arm movements or vocal tract movements. The arm has 3 degrees of freedom and movements are parameterized by a 21 dimensional dynamic movement primitive (DMPs are simple parametric dynamical systems used in robotics to produce structured movement, Ijspeert et al., 2013). The vocal tract is based on the DIVA model (Guenther, 2006), and includes 7 degrees of freedom which global movements are parameterized by a 28D dimensional dynamic movement primitive. There are 4 objects on the table, which can potentially be grasped by the hand of the robot, or interact with each other (one of the object, the stick, can be used to grasp other objects). These objects can be positionned either within reach of the hand, or out of reach of the hand but within reach of the stick, or out of reach of both hand and stick (see figure). They are regularly randomly repositionned across the table in experiments. The social peer model is also capable to produce actions with its arm (it can take an object and place it at another location) as well as produce vocalizations. It is programmed to respond contingently over certain actions of the learning robot: if the learning robot touches one of the objects, then the social peer produces the speech sound corresponding to the name of this object. Also, if the learning robot produces a speech sound that is sufficiently similar to the name of one of the object, the social peer grasps this object and places it just in front of the learner, within hand reach. At other random times, when the learning robot does not interact with objects, the social peer also produces speech sounds that are not related to the scene (distractors). The learning robot can perceive objects locations and track their trajectories. It also perceives auditory trajectories of sounds produced by itself or by the caretaker (as a sequence of key points in the formant space).

Initially, the learning robot has no prior knowledge of the potential affordances of its environment: it does not know whether objects can be moved with its arm, or that objects can interact with each other (it does not know the concept of tool). It also does not know that vocal tract movements can produce sounds, and that these sounds can produce effects on the environment (in the social peer in particular). It does not know either that the social peer can produce contingent reactions. To discover its environment, the learning robot experiments through **intrinsically motivated goal exploration** (see top part of figure 11.1), going repeatedly through the following steps:

- **Perception of world state:** at the beginning of each experimental episode, it first observes the state of the environment;
- **Sampling of self-generated goal:** it then samples an imagined goal. It can sample different kinds of goals: moving the hand or any of the objects along a target trajectory (e.g. "Move object B along trajectory starting from initial state and passing though key points $a_1, ..., a_5$"); produce an acoustic trajectory passing through randomly imagined


Perception of world state
Sample goal *G* e.g. Move obj. *B* or produce sound *S*
Internal predictive model
Action (motor commands)
learning
intrinsic reward (learning progress)
Environment
Outcome on objects and sounds
curiosity-driven learner model
caretaker model
vocal tract and ear model
sound
stick
B
C
A
arm
hand gripper
toys
zone reachable by hand
zone only reachable by stick
zone out of learner's reach

(a) Cognitive architecture of curiosity-driven learning model

(b) Experimental setup

Figure 11.1: In this experiment, a curiosity-driven learner (left agent) explores by autonomously sampling goals to discover which effects can be produced by its body actions. Its action repertoire comprises the possibility to move its arm and hand, as well as to move a simulated vocal tract. The effects it can perceive are trajectories of objects (simulated visual system), as well as sound trajectories (simulated ear). Initially the learner knows nothing about the potential relations between actions and effects, and it does not even know which effects are possible (e.g. what object movements and what sounds are possible in this environment). On the table, several objects are positioned, and can be in different zones. One of the objects can be used as a tool to move other objects, e.g. retrieving them when they are beyond hand reach. On the other side of the table, a simulated social peer utters the speech names of objects when the learner touches them, as well as other disctractor speech sounds when it does not touch objects. It also grasps an object and puts it within hand reach of the learner when the learner produces a sound that is sufficiently close to the name of this object. However, from the learner's perspective, the social peer is initially just one object among others, and there is no a priori knowledge about what are its affordances. Initially, most self-generated goals are either impossible or too difficult to achieve for the learner, except goals for moving the hands around. This produces a diversity of hand movement, leading to the discovery of skills for pushing within reach objects. Then, the learner discovers that acting in a peculiar way on one of the objects (the stick) enables to move at the same time other objects that are beyong hand reach. Simultaneously, the learner also makes progress towards controlling its vocal tract, discovering how to produce sounds that look like the ones uttered by the social peer. Finally, in a later stage, the learner discovers that when some objects are beyond reach, even using the stick, another tool can be used to retrieve them: they can utter a particular speech sound (name of this object) that will influence the behaviour of the social peer to bring it closer. Therefore, the learner discovers the linguistic functionality of speech: it can be use a a tool to influence others.

key formant configurations; produce an acoustic trajectory similar to a randomly selected speech sounds produced by the social peer;

- **Infer action through current internal world model:** Once a goal has been selected, it uses its current internal world model to infer which action has a high probability to achieve this goal, starting from the current initial state of the world. This action may be either an arm movement or a vocal tract movement;
- **Run the action and observe outcomes:** The inferred action is then performed, and outcomes are observed (potential movements of all objects and potential sounds perceived);
- **Update internal world model and goal sampling process:** These outcome observations, combined with the initial state, the goal and the action, are then used to update the internal world model. Hindsight learning is used (this is a form of counterfactual learning): even if the initial goal was not achieved, other goals imagined a posteriori and matching the observations are used to update the world model (e.g. if the robot's goal was to move object A on the left, but its action made it move object B on the right, then the world model is updated to learn how to move object B on the right). Also, the feedback observation on the world can be used to update the goal sampling process to bias the selection of goals towards goals that provide maximum expected learning progress (however, in the model below, a simpler mechanism of random goal sampling is used).

As initially the learning robot has no knowledge of its environment, it will fail at most of the goals it samples, and in particular it will not be able to move any object. However, there are two kinds of outcomes that produce rich and contingent feedback from the environment initially: hand movement and self-produced speech sounds. Indeed, even if initial goal trajectories of the hand are not achieved, most motor commands sent to the arm produce actual movement of the hand: through hindsight learning, this enables to learn fast how to move around the hand. Similarly, even if initial random auditory goals are not achieved, most movements of the vocal tract produce auditory trajectories which enables the development of initial vocal skills through hindsight learning.

As sampling random goals about the hand leads to discover a high-diversity of hand movements (Colas et al., 2018), this increases the probability to discover that arm movements can also move objects (at least in some contexts corresponding to the reachable zone), including the stick. Once a few movements of these objects have been found, learning of object manipulation inside the reachable zone is bootstrapped. At this stage, the learning robot does not know yet how to retrieve and move objects that are beyond hand reach. However, generating goals to move the sticks increases the probability to discover how it can interact with other objects, and finally that it can be used as a proxy to retrieve objects beyond hand reach. A first major cognitive milestone is here reached: from the point of view of an external observer, the learning robot has discovered a form of tool use, i.e. that objects beyond hand reach can sometimes be retrieved with an arm movement that first grasps the stick.

In parallel of these early steps of learning of to move objects with the arm, vocal learning also progresses. After an initial self-exploration and progress of how vocal tract movements produce self-generated sounds, bootstrapping the robot's knowledge of the vocal tract-sounds mapping, the learning robot starts to make progress towards imitating sounds produced by the social peer (both object names and distractors). At some point, when these speech sounds become similar enough to the names of objects used by the social peer, the social peer starts to make contingent responses by grasping the corresponding objects and placing it in front of the learner robot. Thanks to the mechanism of hindsight learning, this enables in turn the learner to discover that "if it had selected the goal of retrieving object A in front of it, then producing sound S would have made the social peer realize the appropriate action to achieve the goal". In particular, the learner discovers that such speech sound production strategy to get the social peer to place the toy in front of it is the most successful strategy to retrieve objects when they are beyond the reach of the stick.

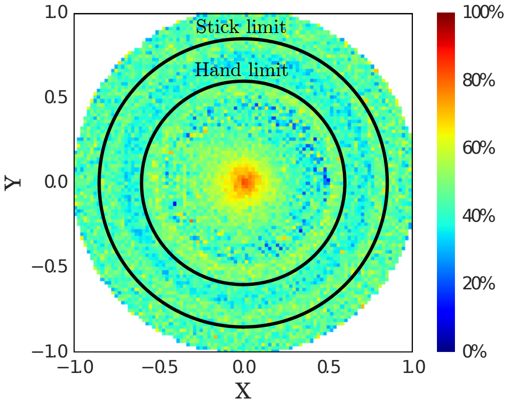


Figure 11.2: Competence to retrieve an object depending on the object's initial position, after 80000 iterations. 0 % means that competence to retrieve a toy there is as bad as with random agents, 100% says that agents perfectly retrieve a toy there.

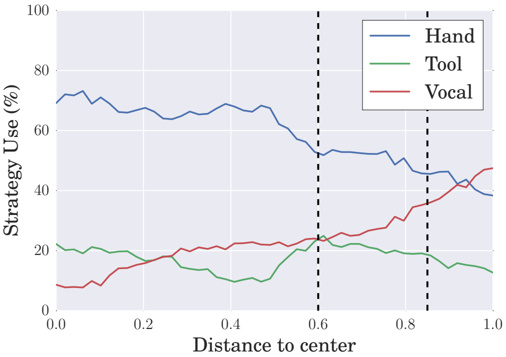


Figure 11.3: Strategy preferences depending on the distance of the toy. The two vertical bars shows the hand and stick limits.

Figures 11.2 and 11.3 summarize the motor and speech communication skills and strategies discovered by the learner at the end of a typical simulation. On figure 11.2 , one visualizes that the learner is able to retrieve and move objects relatively well in all areas of the table: within hand reach, within stick reach, and beyond hand and stick reach. This latter capability shows that it successfully uses the speech production strategy as a tool to manipulate its peers action and help it to retrieve objects: it has discovered a core aspect of the linguistic function of speech. This relative use of strategies is further depicted on figure 11.3 , where one observes that the vocal strategy is selected much more often to retrieve objects beyond the reach of hand or stick. Finally, figure 11.4 shows the distribution of errors for imitation speech sounds of the social peer: one observes that errors are in average smaller for speech sounds that correspond to objects manipulated in the scene than for distractor speech sounds. This is explained by the fact that once the speech strategy to retrieve objects is discovered, the learner targets speech sounds referring to objects much more often than distractors (as they are targeted not only for vocal imitation, but also as tools to retrieve objects), producing more varied local exploration of vocal tract movement which improves the internal model of how these object specific sounds are produced.

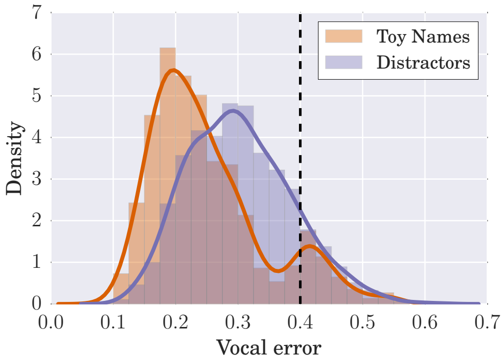


Figure 11.4: Distribution of accuracy of imitations of caregivers' sounds after 80000 iterations. Below 0.4 vocal error, sounds are recognized as imitations by the caregiver. Imitations of toy names are more accurate than imitations of distractors.

11.2 An evo-devo perspective: from curiosity to the evolution of language

In the models we discussed in this chapter and in the previous one, the developmental patterns exhibit a form of behavioral and cognitive epigenesis, as proposed by Gottlieb (1991). Developmental structures in these models are neither learnt from "tabula rasa" nor a predetermined result of an innate "program": instead, they self-organize out of the dynamic interaction between constrained cognitive mechanisms (including curiosity, learning, and abstraction), the morphological properties of the body, and the physical and social environment that itself is constrained and ordered by the developmental level of the organism (Oudeyer, 2011; Smith, 2013; Thelen and Smith, 1996). This self-organization includes the dynamic and automatic formation of behavioral and cognitive stages of progressively increasing complexity, sharing many properties with infant development (Piaget, 1952; Miller, 2002).

Such robotic models and experiments operationalize theories of epigenesis in behavioral development (e.g., Gottlieb, 1991; West and King, 1987; Lickliter and Honeycutt, 2009), as well as the dynamic systems conceptualization of development (Thelen and Smith, 1996). In particular, they allow us to see in detail how the interaction of heterogeneous mechanisms and constraints at several scales could form a dynamical system where developmental structures emerge. Such emergent developmental structures have deep implications for evolution. In particular, they constitute a reservoir of behavioral and cognitive innovations which can be later on recruited for functions not yet anticipated, and at both developmental and evolutionary scales: this is exaptation (Gould, 1991).

First, the results show that modality-independent generic mechanisms for curiosity-driven exploration of the body and its interactions can lead to the emergence of basic speech skills, vocal imitation, and finally the use of speech as a tool to influence other, i.e. language. This suggests that in principle the infant may develop initial language capabilities without an innate specific bias for learning speech and language, and without a teleological knowledge that such skills will be recruited later on, e.g. for achieving joint tasks with their social peers.

Second, such mechanisms should be related to the models we discussed in chapters 6,7, and 8, about the formation and evolution of shared vocalization systems in populations of individuals. Such population models have shown that when vocal learners are equipped with mechanisms of spontaneous vocal self-exploration and progressively tune their vocalizations to match those of their neighbors, conventionalized systems of sounds can be formed spontaneously at the group level from an initial state where each individual only produces random vocalizations. In these models, mechanisms for systematic vocal babbling were pre-programmed and specific to the vocal modality. Yet, computational experiments presented in chapters 10 and 11 showed that principled and modality-independent mechanisms for curiosity-driven exploration can drive learners to explore their vocal tract just as they explore the movement of their arms, observing the effects it produces on external objects, including social peers. Combining these models leads to a startling hypothesis: the interaction of individuals intrinsically motivated to learn about their body, including their vocal tract, and their affordances, can self-organize as a side effect shared speech systems at the group level. Similar analyses has been offered for the origins of joint visual attention (Yu and Smith, 2013; Deak, et al; 2014).

Further, arguments presented in Barto (2013) have shown that the evolutionary origins of such an intrinsic motivation to learn can be explained because it maximizes long-term evolutionary fitness in rapidly changing environmental conditions (e.g., due to human social and cultural structures, which can evolve much faster than the phylogenetic scale). In such a context, Singh et al. (2010) have shown with computer simulations that even if the objective fitness/reward function of an organism amounts to survive and reproduce, it may be more efficient to evolve a control architecture that encodes an innate surrogate reward function rewarding learning per se.

Mechanisms of information seeking can thus evolve independently of language, but yet, as argued above, may have spontaneously bootstrapped early forms of speech and language, both at the individual and population level. This may have opened the possibility for a later recruitment, selection and refinement of language functionalities that were not initially foreseen. This suggests a potentially strong role for curiosity-driven developmental process in the evolution of early stages of language.

11.3 What about other animals?

Such an evolutionary scenario raises interesting issues in the comparison between humans and other animals. As argued earlier, motivational mechanism of curiosity interacts with other motivational mechanisms like food or mate searching, and its weight in motivational arbitration may vary widely across species equipped with such curiosity. Hence, the high degree of competition for survival in many species can be expected to promote avoidance of risk, where aversive motivational systems overcome strongly the expression of curiosity-driven exploration. The multimodal systematicity and the extent to which open-ended free play and curiosity is expressed in humans, where children are comparatively highly protected for a long period, is actually unrivalled in the animal kingdom (Power, 1999). This might explain why many structures could emerge and be recruited out of curiosity in humans as opposed to other species. Yet, this remains an entirely open question, and points to an important research challenge related to the extension of robotic models presented in this chapter: while here the models focused on the self-organization of developmental structures produced by a single motivational mechanism (curiosity), it is of high importance to understand the consequences of coupling it with other forms of extrinsic, possibly aversive motivational systems, and study the impact on developmental dynamics.

Chapter 12

Understanding through Building

My aim in this book has been to contribute to our comprehension of the mechanisms that led to the development of speech in humans. Comprehension – a word whose Latin roots equate to the ideas of 'together' and 'seize' – means having a mental handle on these mechanisms, being able to point to their component parts and follow their logic. But if we are seeking to manipulate these mechanisms in our minds, what better way to start than to try manipulating them with our bodies, to 'grasp them in our hands' and to play with them, in the way that a child plays with the objects it encounters and so discovers how they work. In the same way, computer and robotic models that let us take these mechanisms apart and put them back together, playing with their separate elements as if they were pieces of Lego, can play several important roles. They give us a new formal, constructive scientific language for expressing theories of human speech and natural language that naturalizes these theories, anchoring them in their biological and physical substrates. They let us formulate new conjectures about humans that give rise to new studies and experiments to either support or refute them. Using them we can build in order to understand, delving *in silico* into the morphogenesis of speech, in constant interaction with other human scientists and social scientists. Our approach has involved linking things together: linking different disciplines, linking the different facets of the complex system represented by speech and the development of speech, linking using models that are not reductionist, but systemic.

Research into the origins of speech is only a few decades old, and so has only just begun to illuminate one corner of a very large, very dark room. Considerable work remains in filling out, organizing and making selections in the space of theories. Building computer models in collaboration with specialists in speech and the life sciences is one way of generating some more light.

The artificial systems presented in this book have shown how vocal structures, sharing crucial properties with systems of human speech, could arise spontaneously in a population of individuals in which the codes and their properties had not been pre-programmed, out of neural structures and sensorimotor interactions that are very simple. The individuals in these models do not even differentiate between sounds produced by other individuals and sounds that they themselves produce. If communication is defined as the sending of a signal by an individual intended to modify the internal state or behaviour of another individual, then the individuals in the computational models presented in chapters 6-9 cannot be said to communicate. Also, these models did not include any pressure for linguistic communication which would push the individuals to form repertoires of sounds contrasting with each other. Furthermore, indidivuals do not imitate each other in the sense that they do not immediately reproduce the vocalizations that they hear and do not store them explicitly in memory so as to be able to reproduce them later. Yet, thanks to the self-organizing properties of the complex system formed by the neural coupling between perceptual and motor modalities in each individual, and by the coupling between individuals due to the simple fact that they inhabit an environment where they hear each other, the simulations showed that an organized system of vocalizations emerged spontaneously. While at the start they only produce anarchic, holistic and inarticulate vocalizations, after several hundred interactions they produce discrete combinatorial vocalizations, with phonotactic rules, and conventionalized (all the individuals in the same simulation share the same system of vocalizations at the end, and individuals in different simulations generate different systems). Individuals are even subject to phenomena of acoustic illusions formed and acquired culturally, as are humans. Finally, a statistical study of the repertoires of vocalizations allows us to find the same regularities as those of human languages, in particular as regards the vowels (and when morpho-perceptual constraints similar to those of humans are used).

These models have shown how the existence of phonemes and phonemic coding is not necessarily a result of nonlinearities in the correspondence between vocal motor configuration and the sounds produced. Even when individuals were given abstract linear vocal tracts, discrete combinatorial vocalizations still appeared. This fuels an alternative hypothesis to Stevens's quantal theory of perception (Stevens, 1972), or to Carré and Mrayati's theory of distinctive regions (Carré and Mrayati, 1992). Stepping beyond the formation of phonemes, these models have shown how the interplay between neuronal circuits, at first random and competing to be activated, were able spontaneously to generate regular repertoires of vocalizations, organized according to elementary rules of syntax.

However, results from the models also indicated that these morpho-perceptual nonlinearities, in other words properties of the body, had a strong influence on the statistical distribution of phonemes. Using a model of the vocal tract and of the cochlea, the simulations generated systems of vocalizations with the same regularities as in human languages, particularly regarding vowels. Here we have a good illustration of the power of computer and robotic models: in systems that model all the components of an individual, it is possible to set up experiments where the different properties of the body and of the nervous system can be modified at will and independently of each other. This type of experiment, which would be impossible to conduct on living beings, allows the different properties to be considered separately and their respective roles to be better understood. The body becomes an experimental variable, thus opening up fascinating epistemological perspectives (Kaplan and Oudeyer, 2008).

In these computational models, systems of vocalizations are self-organized. Indeed, the components with which the individuals are initially endowed are of a lower level of complexity than those of the speech codes that are generated. The properties characterizing these components, as well as their local interactions, are qualitatively different from those characterizing the global structure formed by the speech code. The system's dynamics illustrates how the same mechanism can form a whole complex set of structures comprising the speech code, starting with assumptions of a lower order of complexity.

In moving the search for the origins of systems of vocalizations to the wider context of the evolution of forms in biology and in performing experiments *in silico*, we discover new maps that chart previously unknown paths leading towards the sources of speech. There already existed a number of proposed explanations for the structures of speech described here which relied on classical neo-Darwinian functionalist argumentation. For example, Lindblom

(1992) proposed that the statistical regularities in vowel inventories could be explained in terms of optimal perceptual distinctiveness, and thus in terms of their effectiveness in communication. Studdert-Kennedy (1998) proposed that phonemic coding enables transmission of information at a rate that drastically increases the power of communication, and thus that it is the result of an adaptation for communicating more effectively. But as D'Arcy Thompson began to show in *On Growth and Form*, when seeking to understand the evolution of a particular biological form, it is not enough simply to identify the adaptive criteria that allowed the natural selection of this form. It is also necessary to understand the mechanisms of generation, growth and morphogenesis. Alongside the mechanisms of genetic variation in the replication of organisms, there are ontogenetic and social development processes constraining and guiding the exploration of the space of forms. Just like in complex inorganic systems where organized structures form spontaneously, in the organic world the complex system of a developing organism has processes of self-organization that play a central role: they structure the Darwinian evolution of life forms.

Using computer models to explore and test this kind of process of morphogenesis and self-organization has thus given rise to several different evolutionary scenarios. These scenarios are predicated on the evolution of two specific prerequisites that are central to these models, and out of which speech codes form spontaneously: first, plastic neural circuits connecting the motor and the perceptual aspects of speech and, secondly, the existence of babbling, that is to say a spontaneous and systematic exploration of the vocal organ. One can hypothesize that these two elements evolved specifically in response to an evolutionary pressure towards linguistic communication. This is the classical adaptationist explanation: the mechanisms of simple, generic morphogenesis in our models are able to fill this explanation out by showing that the structures to be genetically "wired" do not need to be as complex as might have been imagined.

However, the generic nature of these two elements also suggests two complementary exaptationist scenarios in which the initial structures of speech emerged spontaneously as a collateral effect of the evolution of independent language capacities. In the first scenario the perceptuo-motor neural circuitry and the babbling behavior are both parts of the basic biological package required for adaptive vocal imitation, which could have adapted for reasons that have nothing to do with language and can be seen in several animal species that also manifest structured vocalizations.

In the second scenario, vocal babbling is simply a corollary of spontaneous "body babbling", driven by an intrinsic motivation to explore the body and its environment, by curiosity and the pure pleasure of learning, that is present in humans far more than in other animals. The robotic experiments presented in chapter 10 suggest that vocal interaction itself could emerge as one of the developmental structures resulting spontaneously from the dynamic interplay between curiosity-driven exploration, the learning system, the body of the robot and its physical and social environment. And these initially non-linguistic developmental mechanisms could lead to the morphogenesis of vocal structures that are discrete, combinatorial and shared within a particular group. Thus, at the sources of speech one may find this desire to learn, that we have in common with our ancestors, this internal force that makes babies explore their own bodies, their physical surroundings, and progressively lets them discover other people.

Aflalo TN and Graziano MSA (2006) Possible Origins of the Complex Topographic Organization of Motor Cortex: Reduction of a Multidimensional Space onto a Two-Simensional Array. J. Neurosci. 26: 6288-6297.

Ameisen J.C. (2000) La Sculpture du Vivant, Le Suicide Cellulaire ou la Mort Créatrice. Paris: Editions du Seuil.

Andry P., Gaussier P., Moga S., Banquet J.P., Nadel J. (2001) Learning and Communication in Imitation: An Autonomous Robot Perspective. IEEE Transaction on Systems, Man and Cybernetics, Part A: Systems and Humans, Volume 31, Number 5, pp431-44.

Arbib M. (1995) The Handbook of Brain Theory and Neural Networks. Cambridge, Mass: MIT Press.

Ashby W.R. (1952) Design for a Brain, Chapman Hall.

Bachelard G. (1865) La Formation de l'Esprit Scientifique, Paris, Vrin.

Bailly G., Laboissière R. and Galván A. (1997) Learning to speak: Speech production and sensory-motor representations. In Morasso P. and Sanguineti V., editors, Self-Organization, Computational Maps and Motor Control, pages 593–615. Elsevier, Amsterdam.

Balaban E. (1988) Bird song syntax: Learned intraspecific variation is meaningful, Proceedings of the National Academy of Sciences, Vol. 85, Nr. 10, p. 3657–3660.

Baldassarre G. (2011) What are intrinsic motivations? a biological perspective, in: Proceeding of the IEEE ICDL-EpiRob Joint Conference.

Baldassare G. and M. Mirolli (2013) Intrinsically motivated learning in natural and artificial systems, Berlin: Springer-Verlag.

Baldwin J. (1896) A new factor in evolution, American Naturalist, 30, pp. 441-451.

Ball P. (2001) The Self-Made Tapestry, Pattern Formation in Nature, Oxford University Press.

Baranes A., Oudeyer P-Y. (2013) Active Learning of Inverse Models with Intrinsically Motivated Goal Exploration in Robots, Robotics and Autonomous Systems, 61(1), pp. 49-73.

Baronchelli A., Loreto V. and Steels L. (2008) In-Depth Analysis of the Naming Game Dynamics: The Homogeneous Mixing Case. International Journal of Modern Physics C, 19(5):785-812.

Barto A. (2013) Intrinsic motivation and reinforcement learning, in G. Baldassarre and M. Mirolli, editors, Intrinsically Motivated Learning in Natural and Artificial Systems, pp. 17-47, Springer.

Barto A., Singh S., Chentanez N. (2004) Intrinsically motivated learning of hierarchical collections of skills, in Proc. 3rd Int. Conf. Develop- ment Learn., San Diego, CA, 2004, pp. 112-119.

Bellemare M., Srinivasan S., Ostrovski G., Schaul T., Saxton D. and Munos R. (2016). Unifying count-based exploration and intrinsic motivation. In Advances in Neural Information Processing Systems, page 14711479.

Berlyne D. (1960) Conflict, Arousal and Curiosity. New York: McGraw-Hill.

Bernard, C. (1945) Introduction à l'Etude de la Médecine Expérimentale, Editions du cheval ailé, Genève.

Bernheimer, H., Birkmayer, W., Hornykiewicz, Jellinger, K., and Seitelberger, F. (1973). Brain dopamine and the syndromes of parkinson and huntington: Clinical, morpho- logical and neurochemical correlations. J. Neurol. Sci. 20, 415-455.

Berrah A-R., Glotin H., Laboissière R., Bessière P., Boë (1996) From Form to Formation of Phonetic Structures: An Evolutionary Computing Perspective, in Proc. ICML 1996 Workshop on Evolutionary Computing and Machine Learning, pp. 23-29, Bari, Italy.

Berrah A. and R. Laboissière (1999) SPECIES : an Evolutionary Model for the Emergence of Phonetic Structures in an Artificial Society of Speech Agents, Advances in Articial Life, p. 674-678.

Berthier N. E. , Clifton R.K., McCall D.D., and D.J. Robin (1999) Proximodistal structure of early reaching in human infants, Experimental Brain Research, pp. 259-69, 127(3).

Binmore K. (1992) Fun and Games: a Text on Game Theory, Heath, Lexington, MA.

Boë L.J. (1999) Vowel spaces of newly-born infants and adults consequences for ontogenesis and phylogenesis, in 14th Internation Congress of Phonetic Sciences, pp. 2501-2504.

Boë L.J., Schwartz J.L., Vallée N. (1995) The Prediction of Vowel Systems: Perceptual Contrast and Stability, in Keller E., (ed.), Fundamentals of Speech Synthesis and Recognition, pp. 185-213, Chichester:John Wiley.

de Boer B. (2001) The Origins of Vowel Systems, Oxford Linguistics, Oxford University Press.

Boersma P. (1998) Functional Phonology: Formalizing the Interactions Between Articulatory and Perceptual Drives. The Hague: Holland Academic Graphics.

Browman C.P. and Goldstein L. (1986) Towards an Articulatory Phonology. In C. Ewan and J. Anderson (eds.) Phonology Yearbook 3. Cambridge: Cambridge University Press, pp. 219-252

Browman C.P. and Goldstein L. (2000) Competing Constraints on InterGestural Coordination and Self-Organization of Phonological Structures, Bulletin de la Communication Parlée, vol. 5, pp. 25-34.

Bruner J. (1962) On Knowing: Essays for the Left Hand (Cambridge, MA, Harvard University Press).

Bruner J., Jolly A., Sylva K. (1976) Play: Its Role in Development and Evolution, Basic Books, New York.

Caldwell CA, Whiten A (2002) Evolutionary perspectives on imitation: is a comparative psychology of social learning possible? Anim Cogn 5:193-208.

Calinon S. (2009) Robot Programming by Demonstration. EPFL Press, CRC Press.

Call J., and Carpenter M. (2002) Three Sources of Information in Social learning, in Imitation in animals and artifacts, eds C. L. Nehaniv and K. Dautenhahn (Cambridge, MA: MIT Press), pp. 211-228.

Camazine S., Deneubourg J-L., Franks N.R., Sneyd J., Theraulaz G., Bonabeau E. (2003) Self-Organization in Biological Systems, Princeton University Press.

Cameron J. and Pierce W. (2002). Rewards and Intrinsic Motivation: Resolving the Contoversy (Bergin and Garvey Press)

Cangelosi A., Metta G., Sagerer G., Nolfi S., Nehaniv C.L., Fischer K., Tani J., Belpaeme B., Sandini G., Fadiga L., Wrede B., Rohlfing K., Tuci E., Dautenhahn K., Saunders J., Zeschel A. (2010) Integration of Action and Language Knowledge: A Roadmap for Developmental Robotics. IEEE Transactions on Autonomous Mental Development, 2(3), 167-195.

Cangelosi, A., Schlesinger, M. (2015). Developmental robotics: From babies to robots. MIT Press.

Carlson, R., Granström, B., Fant, G. (1970). Some studies concerning perception of isolated vowels. STL-QPSR, 11(2-3), 019-035

Carré R., Mrayati M. (1992) Distinctive Regions in Acoustic Tubes. Speech Production modeling, Journal d'Acoustique 5, 141-159

Carroll, S.B. (2005) Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom, Norton.

Chalmers, A. F. (1990) Science and its fabrication. U of Minnesota Press.

Changeux J.P. (1983) L'Homme Neuronal, Fayard.

Changeux J.P., Courrège P., Danchin A. (1973) A Theory of the Epigenesis of Neuronal Networks by Selective Stabilization of Synapses, Proceedings of the National Academy of Sciences of the United States of America 70 (10): 2974-2978.

Chauvet G. (1995) La Vie dans la Matière, le Rôle de l'Espace en Biologie, Nouvelle Bibliothèque Scientifique, Flammarion.

Chomsky N. and M. Halle (1968) The Sound Pattern of English. Harper Row, New york.

Chomsky, N. (1975) Reflections on Language. Pantheon.

Christiansen, M. H., Chater, N. (2016) Creating language: Integrating evolution, acquisition, and processing. MIT Press.

Colas, C., Sigaud, O., Oudeyer, P. Y. (2018) GEP-PG: Decoupling Exploration and Exploitation in Deep Reinforcement Learning Algorithms. In Proceedings of International Conference on Machine Learning (ICML).

Cohn D. A., Ghahramani Z., Jordan M. I. (1996) Active learning with statistical models. Journal of artificial intelligence research, 4 : 129-145.

Coppens Y., Picq P. (2001) Aux Origines de l'Humanité, tome 1 : De l'Apparition de la Vie à l'Homme Moderne, Fayard.

Coupé C., Hombert J.M. (2005) Polygenesis of linguistic strategies: a scenario for the emergence of language, in Language Acquisition, Change and Emergence: essays in evolutionary linguistics, Minett , J. and Wang, W.S. (eds), Hong Kong, City University of Hong Kong Press, pp. 153-201.

Coupé C., Marsico E. et Pellegrino F. (2010) Les systèmes sonores des langues comme systèmes complexes, in Qu'appelle-t-on aujourd'hui les sciences de la complexité ? Langages, réseaux, marchés, territoires, Weisbuch, G. et Zwirn, H. (eds), Paris, Vuibert, pp. 37-65.

Coupé C. (2003) De l'origine du langage à l'origine des langues : modélisations de l'émergence et de l'évolution des systèmes linguistiques, Thèse de Doctorat, Sciences cognitives, Université Lyon 2.

Crothers J. (1978) Typology and Universals of Vowels Systems, in Greenberg, Ferguson, Moravcsik, eds., Universals in Human Language, Vol. 2, Phonology, pp. 93-152, Stanford University Press.

Csikszenthmihalyi M. (1991) Flow, the Psychology of Optimal Experience, New York: Harper Perennial.

D'arcy Thompson (1917) On Growth and Form, Cambridge University Press (édition: 1961).

Darwin C. (1859) On the Origins of Species, Signet Book (édition 1999).

Dayan P., Abbot L.F. (2001) Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, MIT Press.

Dawkins R. (1982) The Extended Phenotype, Oxford University Press.

Deak G., Krasno A., Triesch J., Lewis, J., Sepeta, L. (2014) Watch the hands: infants can learn to follow gaze by seeing adults manipulate objects. Developmental Science, early on-line. DOI: 10.1111/desc.12122.

Deci E. and Ryan R. (1985) Intrinsic Motivation and Self-Determination in Human Behavior. New York: Plenum.

Dempster A. P., Laird N. M., and Rubin D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B (Methodol.) 39, pp. 1-38.

Dessalles J. L. (2007) Why we talk: The evolutionary origins of language. Oxford University Press, USA.

Dessalles J. L., Ghadakpour, L. (2004) La construction cognitive du temps.In: D. Badariotti (Ed), Le temps dans les systèmes complexes naturels et artificiels - Actes des journées de Rochebrune. Paris: ENST 2004-S-001, 95-109.

Diehl R.L., Lotto A.J. and Holt L.L. (2004). Speech perception. Annual Review of Psychology, 74 , 431-461.

Dowek G. (2011) Une Deuxiéme Révolution Galiléenne ? http://www.lsv.fr/ dowek/Philo/galilee.pdf

Duda R., Hart P., Stork D. (2000) Pattern Classification, Wiley Publishers.

Edelman, G.M. (1993) Neural Darwinism: Selection and Re-entrant Signaling in Higher Brain Function. Neuron 10:115-125.

Einstein A. (1955) The Meaning of Relativity, Princeton University Press.

Einstein A., Infeld L. (1967) The Evolution of Physics, Simon and Schuster Clarion Book.

Eldredge, N., Gould, S. J. (1972) Punctuated Equilibria: an Alternative to Phylogenetic Gradualism, in: Models In Paleobiology, ed. T. J. M. Schopf.

Escudier P., Schwartz J-L. (eds.) (2000) La Parole, des Modèles Cognitifs aux Machines Communicantes, Hermès Sciences.

Fedorov V. (1972) Theory of Optimal Experiment, New York, NY: Academic.

Fernando C., Vasas V., Szathmáry E., Husbands P. (2011) Evolvable Neuronal Paths: A Novel Basis for Information and Search in the Brain. PLoS ONE 6(8).

Festinger, L. (1957) A theory of Cognitive Dissonance (Evanston, Row, Peterson).

Feyerabend, P. (1979) Contre la Méthode, Esquisse d'une Théorie Anarchiste de la Connaissance, Paris, Seuil.

Fiorillo, C. D. (2004). The uncertain nature of dopamine. Mol. Psychiatry, 122-123.

Fitch, T.W. (2011) Unity and Diversity in Human Language, Phil. Trans. R. Soc. B, vol. 366, no. 1563 376-388.

Flash T. and Hochner B. (2005) Motor Primitives in Vertebrates and Invertebrates, Current Opinion in Neurobiology, 15:1-7.

Forestier S., Oudeyer P-Y. (2017) A Unified Model of Speech and Tool Use Early Development. Proceedings of the 39th Annual Meeting of the Cognitive Science Society.

Forestier S., Oudeyer P-Y. (2016) Overlapping Waves in Tool Use Development: a Curiosity-Driven Computational Model. The Sixth Joint IEEE International Conference Developmental Learning and Epigenetic Robotics.

Forestier S., Mollard Y., Oudeyer P-Y. (2017) Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning. arXiv:1708.02190.

Frankel A.S. (1998) Sound production, Encyclopedia of Marine Mammals, 1998, pp. 1126-1137.

Freeman W. J. (1978) Spatial Properties of an EEG Event in the Olfactory Bulb and Cortex. Electroen- cephalogr. Clin. Neurophysiol. 44, 586-605.

French R.M., Messinger A. (1994) Genes, Phenes and the Baldwin Effect: Learning and Evolution in a Simulated Population, Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems, R. A. Brooks and P. Maes (Eds.) MIT, p. 277-282.

Friston K., Adams R. A., Perrinet L., Breakspear M. (2012) Perceptions as hypotheses: saccades as experiments. Frontiers in psychology, 3.

Friston K., Rigoli F., Ognibene D., Mathys C., Fitzgerald T., Pezzulo G. (2015) Active inference and epistemic value. Cognitive neuroscience, 1-28.

Gintis H., Alden E., Bowles S. (2001) Costly Signaling and Cooperation Journal of Theoretical Biology 213:103-119.

Gopnik A., Meltzoff A. N., Kuhl P. K. (1999) The scientist in the crib: Minds, brains, and how children learn. William Morrow and Co.

Glasersfeld E., (2001) The Radical Constructivist View of Science, in Foundations of Sciences, eds. Riegler A., vol. 6, no. 1-3, pp. 31-43.

Goldstein L. (2003) Emergence of Discrete Gestures, Proceedings of the International Congress of Phonetics Sciences, Barcelona.

Gold E. (1967) Language Identification in the Limit, Information and Control, 10, pp. 447-474.

Goldstein L. (2003b) Development of Phonology, http://www.ling.yale.edu:16080/ling165/.

Gottlieb G. (1991). Epigenetic Systems View of Human Development. Developmental Psychology, 27(1), 33-34.

Gottlieb J., Oudeyer P-Y., Lopes M., Baranes, A. (2013) Information Seeking, Curiosity and Attention: Computational and Neural Mechanisms, Trends in Cognitive Science, , 17(11), pp. 585-596.

Gottlieb J., Hayhoe M., Hikosaka O., Rangel A. (2014). Attention, Reward, and Information Seeking1. The Journal of Neuroscience, 34(46), pp. 15497-15504.

Gottlieb J., Oudeyer P-Y. (2018) Towards a Neuroscience of Active Samping and Curiosity, Nature Reviews Neuroscience, 19, pp. 758-770.

Gould S. J., Vrba E. S. (1982) Exaptation: A Missing Term in the Science of Form. Paleobiology 8:4-15.

Gould S. J. (1991) Exaptation: A crucial tool for an evolutionary psychology. Journal of social issues, 47(3), 43-65.

Gould S. J. (1997) The Exaptive Excellence of Spandrels as a Term and Prototype. Proceedings of the National Academy of Science USA 94:10750-755.

Gould S. J. (1982) Le Pouce du Panda, Grasset.

Gould S.J. (2006) La Structure de la Théorie de l'Evolution, Paris, Gallimard, coll. NRF Essais.

Guenther F.H., Gjaja M.N. (1996) The Perceptual Magnet Effect as an Emergent Property of Neural Map Formation. Journal of the Acoustical Society of America, 100, pp. 1111-1121.

Guenther, F. H., Hampson, M., and Johnson, D. (1998) A theoretical investigation of reference frames for the planning of speech movements. Psychol. Rev. 105, pp. 611-633. doi: 10.1037/0033-295X.105.4.611-633.

Guenther, F. H. (2006) Cortical interactions underlying the production of speech sounds. J. Commun. Disord. 39, pp. 350-365. doi: 10.1016/j.jcomdis.2006. 06.013

Guillaume P. (1925) L'Imitation chez l'Enfant, Paris: Alcan.

Gumperz J. J. and Levinson S. C. (1996) Introduction: Linguistic Relativity Re-Examined. In J. J. Gumperz, S. C. Levinson (Eds.), Rethinking linguistic relativity (pp. 1-20). Cambridge: Cambridge University Press.

Hagège C. (2006) Combat pour le Français : au Nom de la Diversité des Langues et des Cultures, Editions Odile Jacob.

Handel S, Todd S.K., Zoidis A.M. (2009) Rhythmic structure in humpback whale (Megaptera novaeangliae) songs: preliminary implications for song production and perception, J Acoust Soc Am. 2009 Jun;125(6):EL225-30.

Harnad S. (1990) The Symbol Grounding Problem, Physica D, vol. 42, pp. 335-346.

Hinton G. E. and Nowlan S. J. (1987) How learning can guide evolution. Complex systems, 1(1), 495-502.

Hombert, J.M. (ed) (2005) Aux origines des langues et du langage, Fayard.

Hombert, J.M. (2009) La diversité culturelle de l'Afrique est menacée, 429, La Recherche, pp. 36-39.

Hooks, M., and Kalivas, P. (1994). Involvement of dopamine and excitatory amino acid transmission in novelty-induced motor activity. J. Pharmacol Exp. Ther. 269, 976-988.

Howard I. and Messum P. (2011) Modeling the Development of Pronunciation in Infant Speech Acquisition, Motor Control, vol. 15(1), pp. 85-117.

Howard I. and Messum P. (2014) 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).

Hunt, J. M. (1965) Intrinsic motivation and its role in psychological development, Nebraska Symposium on Motivation, 13, 189-282.

Hurford J., Studdert-Kennedy M., Knight C. (1998) Approaches to the Evolution of Language, Cambridge, Cambridge University Press.

Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., Schaal, S. (2013) Dynamical movement primitives: learning attractor models for motor behaviors. Neural computation, 25(2), 328-373

Johnson M. H. (2011). Developmental Cognitive Neuroscience. (3rd Ed) John Wiley and Sons.

Kandel E.R., Schwartz J.H., Jessell T.M. (2000) Principles of Neural Science McGraw-Hill/Appleton and Lange.

Kaneko K., Tsuda I. (2000) Complex Systems: Chaos and Beyond, A Constructive Approach with Applications in Life Sciences, Springer.

Kagan J. (1972) Motives and development. J. Pers. Soc. Psychol. 22, pp. 51-66.

Kaplan F. (2001) La Naissance d' une Langue chez les Robots, Hermes Science.

Kaplan, F., Oudeyer, P-Y. (2008) Le corps comme variable expérimentale, Revue Philosophique de la France et de l'Etranger, pp. 287-298.

Kaplan F., Oudeyer P-Y., Bergen B. (2008) Computational Models in the Debate over Language Learnability, Infant and Child Development, 17(1), pp. 55-80.

Kaplan F. and Oudeyer P-Y. (2007) In Search of the Neural Circuits of Intrinsic Motivation, Frontiers in Neuroscience, 1(1), pp.225-236.

Kaplan, F. and Oudeyer, P.-Y. (2003) Motivational principles for visual know-how development. In Prince, C., Berthouze, L., Kozima, H., Bullock, D., Stojanov, G., and Balkenius, C., editors, Proceedings of the 3rd international workshop on Epigenetic Robotics : Modeling cognitive development in robotic systems, no. 101, pages 73-80. Lund University Cognitive Studies.

Kauffman, Stuart (1993) The Origins of Order: Self Organization and Selection in Evolution. Oxford University Press.

Kauffman S. (1996) At Home in the Universe: The Search for Laws of Self-Organization and Complexity, Oxford University Press.

Keefe A., Szostak J. (2001) Functional Proteins from a Random-Sequence Library, Nature 410, 715-718.

Kelley, L.A., Coe R.L., Madden J.R., Healy S.D. (2008) Vocal mimicry in songbirds, Animal Behaviour 76 (3): 521-528.

Kidd C., Piantadosi S. T., Aslin R. N. (2012) The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PLoS One, 7(5), e36399.

Kirby S. (1998) Syntax without Natural Selection: how Compositionality Emerges from Vocabulary in a Population of Learners, in Hurford, J., Studdert-Kennedy M., Knight C.

(eds.), Approaches to the Evolution of Language, Cambridge, Cambridge University Press.

Kirby S. and Hurford J. (2002) The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model. In Cangelosi, A. and Parisi, D., editors, Simulating the Evolution of Language, chapter 6, pages 121-148. Springer Verlag, London.

Kirby, S., Griffiths, T., Smith, K. (2014) Iterated learning and the evolution of language. Current opinion in neurobiology, 28, 108-114.

Kitano H. (2004) Cancer as a Robust System: Implications for Anticancer Therapy. Nature Reviews Cancer 4:227-235.

Kitano H. (2002) Systems Biology: A Brief Overview. Science. 295, 1662-1664.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Herwig, R. (2016) Systems biology: a textbook. John Wiley and Sons.

Knill D. and Pouget A. (2004) The Bayesian Brain: the Role of Uncertainty in Neural Coding and Computation, TRENDS in Neurosciences, Vol.27, No.12.

Kobayashi T., Kuroda T. (1987) Morphology of Crystals, ed. Sunagawa I., Terra Scientific.

Kobayashi, R. (1998) Modeling and Numerical Simulations of Dendritic Crystal Growth, Physica D: Nonlinear Phenomena, Volume 63, Issues 3-4, Pages 410-423.

Kohonen T. (1982) Self-Organized Formation of Topologically Correct Feature Maps, Biol. Cybern., no. 43, pp. 594-599.

Konishi M. (1989) Birdsong for Neurobiologists. Neuron , 3, 541-549.

Konishi, M. (2010) From central pattern generator to sensory template in the evolution of birdsong, Brain and Language, 15: 18-20.

Kroöger B. J., Kannampuzha J., and Neuschaefer-Rube C. (2009) Towards a neu- rocomputational model of speech production and perception. Speech Commun. 51, pp. 793-809. doi: 10.1016/j.specom.2008.08.002

Kuhn T.S. (1970) The Structure of Scientific Revolutions, University of Chicago Press.

Kuhl P. K., Williams K. A., Lacerda F., Stevens K. N., Lindblom, B. (1992) Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. Science, 255, 606-608.

Kuhl, P. K. (2004) Early language acquisition: cracking the speech code. Nat. Rev. Neurosci. 5, pp. 831-843. doi: 10.1038/nrn1533

Kupiec J-J., Sonigo P. (2000) Ni Dieu ni Géne, pour une autre Théorie de l'Hérédité, collection Science Ouverte, Seuil.

Krebs J.R., Ashcroft R., Weber M. (1978) Song repertoires and territory defense in the great tit, Nature 271, 539-542.

Labov W. (1994) Principles of Linguistic Change. Volume 1: Internal Factors. Oxford: Basil Blackwell.

Ladefoged, P. and I. Maddison (1996) The Sounds of the World's Languages. Blackwell Publishers, Oxford.

Langer, J.S. (1980) Instabilities and Pattern Formation in Crystal Growth, Rev. Mod. Phys. 52, 1-28.

Langton C. (1995) Artificial Life: an Overview. MIT Press.

Laversanne-Finot, A., Péré, A. and Oudeyer, P. Y.(2018) Curiosity Driven Exploration of Learned Disentangled Goal Spaces, in Proceedings of The 2nd Conference on Robot Learning, in PMLR 87:487-504.

Lehman J. and Stanley K. O. (2008). Exploiting open-endedness to solve problems through the search for novelty. In ALIFE, pages 329-336.

Levinson, S. (2003). Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press.

Libbrecht K. (2004) The Little Book of Snowflakes, Stillwater,MN:Voyageur.

Liberman A.M., Mattingly I.G. (1985) The Motor Theory of Speech Perception Revised, Cognition, 21, 1-36, 1985.

Lickliter R., and Honeycutt H. (2009) Rethinking Epigenesis and Evolution in Light of Developmental Science. Handbook of Behavioral and Comparative Neuroscience: Epigenetics, Evolution, and Behavior.

Liljencrants L., Lindblom (1972) Numerical Simulations of Vowel Quality Systems: The Role of Perceptual Contrast, Language 48, (1972), 839-862.

Lindblom, B. (1992) Phonological Units as Adaptive Emergents of Lexical Development, in Ferguson, Menn, Stoel-Gammon (eds.) Phonological Development: Models, Research, Implications, York Press, Timonnium, MD, pp. 565-604.

Lopes M., Melo F., Montesano L., and Santos-Victor J. (2010) Abstraction Levels for Robotic Imitation: Overview and Computational Approaches, in From Motor Learning to Interaction Learning in Robots. Vol. 264, eds O. Sigaud and J. Peters (Berlin; Heidelberg: Springer), 313-355.

Lopes M., Oudeyer P-Y. (2010) Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges (Guest editorial), IEEE Transactions on Autonomous Mental Development, 2(2), pp. 65-69.

Lopes M., and Oudeyer P.-Y. (2012) The strategic student approach for life-long exploration and learning, in IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL), (San Diego, CA). doi: 10.1109/DevLrn.2012.6400807.

Lopes M., Montesano L. (2014) Active Learning for Autonomous Intelligent Agents: Exploration, Curiosity, and Interaction. arXiv preprint arXiv:1403.1497.

Lopes M., Oudeyer P-Y. (2010) Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges (Guest editorial) IEEE Transactions on Autonomous Mental Development, 2(2), pp. 65-69.

Lopes M., Lang T., Toussaint M. and P-Y. Oudeyer (2012) Exploration in Model-based Reinforcement Learning by Empirically Estimating Learning Progress, Neural Information Processing Systems (NIPS), Tahoe, USA, 2012.

Lopes M., Oudeyer P-Y. (2012) The Strategic Student Approach for Life-Long Exploration and Learning, in Proceedings of IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-Epirob), San Diego, USA.

Loreto V., Mukherjee A., Tria, F. (2012) On the Origin of the Hierarchy of Color Names, Proc. Natl. Acad. Sci. USA 109 (18), 6819.

Lowenstein G. (1994). The psychology of curiosity: a review and reinterpretation, Psychological Bulletin 116(1): 75-98.

Maeda S. (1989) Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model, Speech production and speech modelling, pp. 131-149.

Markey K. L. (1994) The Sensorimotor Foundations Of Phonology: A Computational Model of Early Childhood Articulatory and Phonetic Development. PhD thesis, University of Colorado at Boulder.

Martius G., Der R., and Ay N. (2013) Information driven self-organization of complex robotic behaviors. PloS one, 8(5):e63400.

McGeer T. (1993) Dynamics and Control of Bipedal Locomotion, J. Theoret. Biol., vol. 16, no. 3, pp. 277-314.

MacNeilage P.F. (1998) The Frame/Content Theory of Evolution of Speech Production. Behavioral and Brain Sciences, 21, 499-548.

Maddieson I. (1984) Patterns of Sound, Cambridge University Press.

Marler P., Slabbekoorn H.W. (2004). Nature's music: the science of birdsong. Academic Press.

Mataric M., Williamson M., Demiris J., Mohan A. (1998) Behavior-Based Primitives for Articulated Control, Proc. Fifth Interntional Conference of the Society for Adaptive Behavior, pp. 165–170., MIT Press, Cambridge, Mass., pp. 165-170.

McGeer T. (1993) Dynamics and Control of Bipedal Locomotion, J. Theoret. Biol., vol. 16, no. 3, pp. 277-314.

Mehler J., Christophe A., Ramus F. (2000) What we Know about the Initial State for Language. In A. Marantz, Y. Miyashita, W. O'Neil (Eds.), Image, Language, Brain: Papers from the first Mind-Brain Articulation Project symposium (pp. 51-75). Cambridge, MA: MIT Press.

Mercado E 3rd, Herman LM, Pack AA. (2005) Song copying by humpback whales: themes and variations, Anim Cognition, 8(2):93-102.

Miller P. (2002) Theories of developmental psychology, Worth Publishers.

Miura K., Yoshikawa Y., Asada M. (2012) Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver, Advanced Robotics 26, pp. 23-44.

Morgan C. L. (1896) On modification and variation. Science 4: 733-740.

Moulin-Frier C. (2011) Rôle des Relations Perception-Action dans la Communication Parlée et l'Emergence des Systèmes Phonologiques : Etude, Modélisation Computationnelle et Simulations. Thèse de doctorat de l'Universit/'e de Grenoble, France.

Moulin-Frier C., Diard J., Schwartz J.-L., Bessière P. (2015) COSMO ("Communicating about Objects using Sensory-Motor Operations"): a Bayesian modeling framework for studying speech communication and the emergence of phonological systems, Journal of Phonetics, in press.

Moulin-Frier C., Laurent R., Bessière P., Schwartz J.-L., and Diard J. (2012) Adverse Conditions Improve Distinguishability of Auditory, Motor and Perceptuo-Motor Theories of Speech Perception: an Exploratory Bayesian Modeling Study. Language and Cognitive Processes, 27(7-8):1240-1263.

Moulin-Frier C., Nguyen S.M., Oudeyer P-Y. (2014) Self-organization of early vocal development in infants and machines: the role of intrinsic motivation, Frontiers in Psychology (Cognitive Science), 4(1006).

Moulin-Frier C. and Oudeyer P-Y. (2012) Curiosity-Driven Phonetic Learning, Proceedings of IEEE International Conference on Development and Learning and Epigenetic Robotics, San Diego, US.

Moulin-Frier C., and Oudeyer P.-Y. (2013a) Exploration strategies in develop- mental robotics: a unified probabilistic framework, in International Conference on Development and Learning, Epirob, Osaka.

Moulin-Frier C., and Oudeyer P.-Y. (2013b) The role of intrinsic motivations in learning sensorimotor vocal mappings: a developmental robotics study, in Proceedings of Interspeech, (Lyon).

Moulin-Frier C., Schwartz J., Diard J. et Bessière P. (2008) Emergence of a Language through Deictic Games within a Society of Sensori-Motor Agents in Interaction, dans 8th International Seminar on Speech Production, ISSP'08, Strasbourg France.

Nakanishi A. (1998) Writing Systems of the World. Alphabets. Syllabaries. Pictograms, Ed. Charles E., Tuttle Company.

Nakaya U. (1954) Snow Crystals: Natural and Artificial, Cambridge: Harvard University Press.

Nicolis G., Prigogine I. (1977) Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations, Wiley.

Noble, D. (2006) The Music of Life: Biology beyond the Genome. Oxford: Oxford University Press.

Nowak M. A., Komarova N. L. and Niyogi P. (2002) Computational and Evolutionary Aspects of Language. Nature 417, 611-617.

Nguyen S. M., and Oudeyer P.-Y. (2012) Active choice of teachers, learning strate- gies and goals for a socially guided intrinsic motivation learner. Paladyn J. Behav. Robot. 3, pp. 136-146. doi: 10.2478/s13230-013-0110-z

Nguyen, M., Oudeyer, P-Y. (2013) Active Choice of Teachers, Learning Strategies and Goals for a Socially Guided Intrinsic Motivation Learner, Paladyn Journal of Behavioural Robotics, 3(3):136:146. Oller, D.K. (2000) The Emergence of the Speech Capacity, Lawrence Erlbaum and Associates, Inc.

van Ooyen A., van Pelt J., Corner M.A., Kater S.B. (2003) Activity-Dependent Neurite Outgrowth: Implications for Network Development and Neuronal Morphology, In: Van Ooyen, A. (ed.) Modeling Neural Development. The MIT Press, Cambridge, Massachusetts.

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. science, 314(5796), 91-94.

Oudeyer P-Y. (2001a), Coupled Neural Maps for the Origins of Vowel Systems. in the Proceedings of ICANN 2001, International Conference on Artificial Neural Networks, pp. 1171-1176, LNCS 2130, eds. G. Dorffner, H. Bischof, K. Hornik, Springer Verlag.

Oudeyer P-Y (2001b), The Epigenesis of Syllable Systems : a Computational Model. Proceedings of ORAGE 2001, Orality and Gestuality conference, Aix-en-Provence, France,, 2001.

Oudeyer P-Y (2001c), The Origins Of Syllable Systems : an Operational Model. in the Proceedings of the 23rd Annual Conference of the Cognitive Science society, COGSCI'2001, pp. 744-749, eds. J. Moore, K. Stenning, Laurence Erlbaum Associates.

Oudeyer P-Y. (2002a) Phonemic Coding Might be a Result of Sensori-Motor Coupling Dynamics, in the Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior, pp. 406-416, eds. B. Hallam, D. Floreano, J. Hallam, G. Hayes, J-A. Meyer, MIT Press.

Oudeyer P-Y. (2002b) A Unified Model for the Origins of Phonemically Coded Syllables Systems, in the Proceedings of the 24th Annual Conference of the Cognitive Science Society, Laurence Erlbaum Associates.

Oudeyer P-Y. (2003a) The Social Formation of Acoustic Codes with "Something Simpler", in the Proceedings of the Second International Conference on Imitation in Animals and Artefacts, eds. Dautenham K., Nehaniv C., Aberystwyth, Wales. acrobat document.

Oudeyer P-Y. (2003b) From Analogue to Digital Vocalization, in Evolutionary Pre-Requisistes for Language, ed. Tallerman M., Oxford University Press.

Oudeyer P-Y (2005a) How Phonological Structures can be Culturally Selected for Learnability, Adaptive Behavior, 13(4), pp. 269-280.

Oudeyer P-Y. and Kaplan F. (2007) Language Evolution as a Darwinian Process: Computational Studies, Cognitive Processing, 8(1), pp. 21-35.

Oudeyer P-Y. (2005b) The Self-Organization of Combinatoriality and Phonotactics in Vocalization Systems, Connection Science, 17(3-4), pp. 325-341

Oudeyer P-Y. (2005c) The Self-Organization of Speech Sounds, Journal of Theoretical Biology, 233(3), pp. 435-449.

Oudeyer P-Y. (2006a) Self-Organization in the Evolution of Speech, Studies in the Evolution of Language, Oxford University Press.

Oudeyer P-Y., Kaplan F. (2006b) Discovering Communication, Connection Science, 18(2), pp. 189-206.

Oudeyer, P-Y. and Kaplan, F. (2007) Language Evolution as a Darwinian Process: Computational Studies, Cognitive Processing, 8(1), pp. 21-35.

Oudeyer, P-Y. and Kaplan, F. (2007b) What is Intrinsic Motivation? A Typology of Computational Approaches, Frontiers in Neurorobotics, 1:6, doi: 10.3389/neuro.12.006.2007.

Oudeyer P-Y, Kaplan , F. and Hafner, V. (2007) Intrinsic Motivation Systems for Autonomous Mental Development, IEEE Transactions on Evolutionary Computation, 11(2), pp. 265-286.

Oudeyer P-Y. (2009) Sur les Interactions entre la Robotique et les Sciences de l'Esprit et du Comportement, in Informatique et Sciences Cognitives: Influences ou Confluences?, eds. C. Garbay et D. Kaiser, série Cogniprisme, Presses Universitaires de France.

Oudeyer P-Y. (2010) On the Impact of Robotics in Behavioral and Cognitive Sciences: from Insect Navigation to Human Cognitive Development, IEEE Transactions on Autonomous Mental Development, 2(1), pp. 2-16.

Oudeyer P-Y. (2011) Developmental Robotics, Encyclopedia of the Sciences of Learning, N.M. Seel ed., Springer Reference Series, Springer.

Oudeyer P-Y. and Smith. L. (2016) How Evolution may work through Curiosity-driven Developmental Process, Topics in Cognitive Science, 1-11.

Oudeyer P-Y., Gottlieb J., and Lopes, M. (2016) Intrinsic motivation, curiosity and learning: theory and applications in educational technologies, Progress in Brain Research, 229, pp. 257-284.

Oudeyer P-Y. (2018) Computational Theories of Curiosity-driven Learning, in "The New Science of Curiosity", ed. G. Gordon, Nova publishing.

Oudeyer P-Y., Kachergis G., Schueller W. (2018) Commutational and Robotic Models of Early Language Development, a Review, in International Handbook on Language Development, ed. Horst J., von Koss Torkildsen J., Taylor and Francis.

Pfeifer R., Lungarella M. et Iida F. (2007) Self-Organization, Embodiment, and Biologically Inspired Robotics, Science 318, pp. 1088-1093.

Peirce C.S (1958) Collected Papers. (CP). Band I-VI. (Hrsg.) Charles Hartshorne und Paul Weiss. Harvard University Press 1931-1935. Band VII, VIII. (Hrsg.) Arthur W. Burks.

Piaget J. (1952) The Origins of Intelligence in Children, New York: International University Press.

Pinker S., Bloom P. (1990), Natural Language and Natural Selection, The Brain and Behavioral Sciences, 13, pp. 707-784.

Popper, K. (1984) La Logique de la Découverte Scientifique, Paris, Payot.

Power T.G. (1999) Play and Exploration in Children and Animals, Lawrence Erlbaum Associates.

Prigogine I. and Nicolis G. (1977) Self-Organization in Non-Equilibrium Systems. Wiley.

Redford M.A., Chen, C.C., Miikkulainen R. (2001) Constrained Emergence of Universals and Variation in Syllable Systems. Language and Speech 44:27-56.

Richards E.J. (2006) Inherited Epigenetic Variation, Revisiting Soft iInheritance. Nat Rev Genet 7(5): 395-401.

Rizzolatti G, Fadiga L, Gallese V, Fogassi L (1996) Premotor Cortex and the Recognition of Motor Actions. Cognitive Brain Res 3: 131-141.

Rizzolatti G., Arbib M.A. (1998) Language within our Grasp. Trends Neurosci 21: 188-194.

Rolf M., and J.J. Steil (2013) Efficient exploratory learning of inverse kinematics on a bionic elephant trunk, IEEE Trans. Neural Networks and Learning Systems 25(6), pp. 1147-1160.

Salinas E., Abbott L.F. (1994) Vector reconstruction from firing rates. Journal of Computational Neuroscience 1: 89-107.

Schembri M., Mirolli M., Baladassarre G. (2007) Evolution and learning in an in- trinsically motivated reinforcement learning robot, in: Springer (Ed.), Proceedings of the 9th European Conference on Artificial Life (2007), Berlin, pp. 294-333.

Schlesinger M., Parisi D., and Langer J. (2000) Learning to reach by constraining the movement search space, Developmental Science, 3:67-80.

Schmidhuber J. (1991) Curious model-building control systems, in Proc. Int. Joint Conf. Neural Netw., Singapore, 1991, vol. 2, pp. 1458-1463.

Schmidhuber J. (2011) Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem, in: Report arXiv:1112.5309.

Schueller, W., Loreto, V., Oudeyer, P. Y. (2018). Complexity Reduction in the Negotiation of New Lexical Conventions. Proceedings of 40th Annual Cognitive Science Society Meeting (CogSci 2018).

Schultz, W. (1998) Predictive reward signal of dopamine neurons. J. Neurophysiol. 80, 1-27.

Schulz L. (2012) The origins of inquiry: Inductive inference and exploration in early childhood. Trends in Cognitive Sciences, 16, 382-389.

Schwartz J.L., Boë L.J., Vallée N., Abry C. (1997) The Dispersion/Focalization Theory of Vowel Systems, Journal of phonetics, 25:255-286.

Schwartz J.L., Boë L.J., Vallée N., Abry C. (1997b) Major Trends in Vowel Systems Inventories, Journal of Phonetics, 25, pp. 255-286.

Sekuler R. and Blake R. (1994) Perception, McGrawHill.

Shannon C., (1948) A Mathematical Theory of Communication, Bell System Technical Journal, vol. 27, pp. 379-423 and pp. 623-656, July and October.

Sigismund, B. (1971) Child Language: A Book of Readings, Chapter Kind und Welt. Englewood Cliffs, NJ: Prentice-Hall. (Original work published in 1856).

Singh S., Lewis R. L., Barto A.G., Sorg J.(2010) Intrinsically motivated reinforcement learning: An evolutionary perspective, Autonomous Mental Development, IEEE Transactions on 2(2): 70-82.

Smith L. B. and Thelen E. (1993) A Dynamic Systems Approach to Development, MIT Press.

Smith L. B., Breazeal, C. (2007) The dynamic lift of developmental process. Developmental Science, 10(1), 61-68.

Smith L. B. (2013) It's all connected: Pathways in visual object recognition and early noun learning. American Psychologist, 68(8), 618.

Smith, L. B., Jayaraman, S., Clerkin, E., Yu, C. (2018) The developing infant creates a curriculum for statistical learning. Trends in cognitive sciences.

Spranger M. and Steels L. (2012) Emergent Functional Grammar for Space. In Steels, L., editor, Experiments in Cultural Language Evolution, Advances in Interaction Studies (vol. 3), pages 207-232, John Benjamins. Amsterdam.

Stahl A. E. and Feigenson, L. (2015) Observing the unexpected enhances infants' learning and exploration. Science, 348(6230), 91-94.

Steels, L. (2016) Agent-based models for the emergence and evolution of grammar. Phil. Trans. R. Soc. B, 371(1701), 20150447.

Steels L. and Hild M. (2012) Language Grounding in Robots. Springer, New York.

Steels L. (2012) Experiments in Cultural Language Evolution. Advances in Interaction Studies (vol. 3), John Benjamins, Amsterdam.

Steels L, Kaplan F. (2001) AIBO's first words: the Social Learning of Language and Meaning. Evolution of Communication, 4(1):3-32.

Steels L, Belpaeme T. (2005) Coordinating Perceptually Grounded Categories through Language: a Case Study for Colour. Behavioral and Brain Sciences, 28:469-529.

Steels L, Loetzsch M. (2008) Perspective Alignment in Spatial Language. In: Coventry KR, Tenbrink T, Bateman JA, editors. Spatial Language and Dialogue. Oxford University Press.

Steels, L. (2003) Evolving Grounded Communication for Robots. Trends in Cognitive Science, 7(7):308-312

Steels, L. (2001) The Methodology of the Artificial. Behavioral and brain sciences, 24(6).

Steels L. (1999) The Talking Heads Experiment. Volume 1. Words and Meanings. Laboratorium, Antwerpen.

Steels L. (1997) The Synthetic Modeling of Language Origins. Evolution of Communication, 1(1):1-35.

Steels L. and Oudeyer P-Y. (2000) The Cultural evolution of Phonological Constraints in Phonology, in Bedau, McCaskill, Packard and Rasmussen (eds.), Proceedings of the 7th International Conference on Artificial Life, pp. 382-391, MIT Press.

Stevens, K.N. (1972) The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data, in David, Denes (eds.), Human Communication: a Unified View, pp. 51-66, New-York:McGraw-Hill.

Studdert-Kennedy M., Goldstein L. (2002) Launching Language: The Gestural Origin of Discrete Infinity, in Christiansen M. and Kirby S. (eds.), Language Evolution: The States of the Art, Oxford University Press.

Studdert-Kennedy M. (1998) The Particulate Origins of Language Generativity: From Syllable to Gesture. In Hurford, J. R., Studdert-Kennedy, M. and Knight C., editors, Approaches to the Evolution of Language - Social and Cognitive Bases. Cambridge: Cambridge University Press.

Studdert-Kennedy M. (2005) How did Language Go Discrete ? In : Tallerman, M. (ed.), Evolutionary Prerequisites of Language, pp. 48-67, Oxford University Press.

Stulp F. and Oudeyer P. Y. (2018) Proximodistal exploration in motor learning as an emergent property of optimization. Developmental science, 21(4), e12638.

Taine H. (1971) Child Language: A Book of Readings, Chapter Acquisition of Language by Children. Englewood Cliffs, NJ: Prentice-Hall. (Original work published in 1856).

Thelen E. (1994). Three-month-old infants can learn task-specific patterns of interlimb coordination. Psychological Science, 5(5), 280-285.

Thelen E. S., Smith L. B. (1996). Dynamic systems approach to the development of cognition and action. MIT press.

Turing A. (1952) The Chemical Basis of Morphogenesis, Philosophical Transactions of the Royal Society of ndon. Series B, Biological Sciences, 237(641), pp. 37-72.

Vallée N. (1994) Systèmes Vocaliques: de la Typologie aux Prédictions, Doctorat en Sciences du Langage, Université Stendhal, Grenoble, France.

van Geer, P. (1991) A Dynamic Systems Model of Cognitive and Language Growth. Psychol. Rev. 98, 3. doi: 10.1037/0033-295X.98.1.3

Vihman, M. (1996). Phonological Development: The Origins of Language in the Child. Cambridge, MA: Blackwell.

Vihman M. M., Ferguson, C. A., and Elbert, M. (1986) Phonological development from babbling to speech: common tendencies and individual differences. Appl. Psycholinguist. 7, pp. 3-40. doi: 10.1017/S0142716400007165

von Frisch, K. (1974) Animal Architecture. London: Hutchinson.

Waddington, C. H. (1946). How animals develop. London : George Allen and Unwin Ltd.

Waldrop, M. (1990) Spontaneous Order, Evolution, and Life, Science, 247, pp. 1543-1545. Williams, G. (1996) Adaptation and Natural Selection: a Critique of some Current Evolutionary Thought, Princeton University Press.

Warlaumont A. (2013a) Salience-based reinforcement of a spiking neural network leads to increased syllable production, in IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL), (Osaka). doi: 10.1109/DevLrn.2013.6652547

Warlaumont A. S. (2013b) Prespeech motor learning in a neural network using reinforcement. Neural Netw. 38, pp. 64-95.

Weng J., McClelland J., Pentland A., Sporns O., Stockman I., Sur M., and Thelen E. (2001) Autonomous mental development by robots and animals, Science, vol. 291, pp. 599-600, 2001.

Werner, S., Vu, H. T. K., Rink, J. C. (2017). Self-organization in development, regeneration and organoids. Current opinion in cell biology, 44, 102-109.

West-Eberhard, M-J. (2003) Developmental plasticity and evolution, New York: Oxford University Press.

West M. J., and King A. P. (1987) Settling nature and nurture into an ontogenetic niche. Developmental Psychobiology, 20(5), 549-562. PMID: 3678619

White R. (1959) Motivationreconsidered:Theconceptofcompetence, Psychol. Rev., vol. 66, pp. 297-333.

Wilkinson, D. (2011) Stochastic Modelling for Systems Biology, Chapman and Hall/CRC.

Williams G. (1996) Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought, Princeton University Press.

Wolfram S. (2002) A New Kind of Science, Wolfram Media.

Yu C. and Smith L. B. (2012) Embodied Attention and Word Learning by Toddlers, Cognition, 125(2):244-62.

Yu C. and Smith L. B. (2013) Joint Attention without Gaze Following: Human Infants and Their Parents Coordinate Visual Attention to Objects through Eye-Hand Coordination. PloS one, 8(11), e79659.

Zuidema, W. and de Boer, B. (2009) The Evolution of Combinatorial Phonology, Journal of Phonetics 37(2) 125-144.

Zuidema W. (2002) How the Poverty of the Stimulus Solves the Poverty of the Stimulus, in: Suzanna Becker, Sebastian Thrun, and Klaus Obermayer (eds.), Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02), MIT Press, Cambridge, MA, pp. 51-58, 2003.