From Analogue to Digital Speech Sounds

Pierre-Yves Oudeyer
Sony Computer Science Lab, Paris
e-mail: py@csl.sony.fr

Abstract

Sound is a medium used by humans to carry information. The existence of this kind of medium is a pre-requisite for language. It is organized into a code, called speech, which provides a repertoire of forms that is shared in each language community. This code is necessary to support the linguistic interactions that allow humans to communicate. How then may a speech code be formed prior to the existence of linguistic interactions?

Moreover, the human speech code is characterized by several properties: speech is digital and compositional (vocalizations are made of units re-used systematically in other syllables); phoneme inventories have precise regularities as well as great diversity in human languages; all the speakers of a language community categorize sounds in the same manner, but each language has its own system of categorization, possibly very different from every other. How can a speech code with these properties form?

These are the questions we will approach in the paper. We will study them using the method of the artificial. We will build a society of artificial agents, and study what mechanisms may provide answers. This will not prove directly what mechanisms were used for humans, but rather give ideas about what kind of mechanism may have been used. This allows us to shape the search space of possible answers, in particular by showing what is sufficient and what is not necessary.

The mechanism we present is based on a low-level model of sensory-motor interactions. We show that the integration of certain very simple and non language-specific neural devices allows a population of agents to build a speech code that has the properties mentioned above. The originality is that it pre-supposes neither a functional pressure for communication, nor the ability to have coordinated social interactions (they do not play language or imitation games). It relies on the self-organizing properties of a generic coupling between perception and production both within agents, and on the interactions between agents.

1 The speech code

Sound is a medium used by humans to carry information to speak to each other. The existence of this kind of medium is a pre-requisite for language[1]. It is organized into a code, called speech, which provides a repertoire of forms that is shared in each language community and allows its users to encode content information. This code is mainly conventional, and thus intuitively requires coordinated interaction and communication to be established. How then may a speech code be formed prior to the existence of communication and of language-like interaction patterns?

Moreover, the human speech code is characterized by several properties which we have to explain. Here are some of them:

Property 1 (discreteness and compositionality): speech sounds are phonemically coded. This implies two aspects: 1) in each language, the continuum of possible sounds is broken into digital units; 2) these units are systematically re-used to build higher level structures of sounds, like syllables.

For example, in articulatory phonology (Studdert-Kennedy and Goldstein, 2002), a vocalization is viewed as multiple tracks in which gestures are performed in parallel (the set of tracks is called the gestural score). A gesture is the combination of several articulators (e.g. the jaw, the tongue) to perform a constriction somewhere in the mouth. The constriction is defined by the place of obstruction of the air as well as the manner. While for example, given a subset of organs, the space of possible places of constrictions is a continuum (for example the vowel continua from low to high, executed by the tongue body) each language uses only a few places to perform gestures. This is what we call discreteness [2]. Furthermore, gestures and their combinations, that may be called "phonemes", are systematically re-used in the gestural scores who specify the syllables of each language. This is what we call compositionality. Some researchers call this "phonemic coding".

Property 2 (universal tendencies): re-occurring units of vocalization systems are characterized by universal tendencies. For example, our vocal tract makes it possible to produce hundreds of different vowels. Yet, each particular vowel system uses most often only 3, 4, 5 or 6 vowels, and extremely rarely more than 12 (Schwartz et al. 1997a). Moreover, there are vowels that appear much more often than others. For example, most languages contain the vowels [a], [i] and [u] (87 percent of languages) while some others are very rare, like [y], [oe] and [ui] (5 percent of languages). Also, there are structural regularities: for example, if a language contains a front unrounded vowel of a certain height, for example the /ϵ/ in "bet", it will also usually contain the back rounded vowel of the same height, which would be here the /aw/ in "hawk".

[1] Other examples of mediums that human use are manual signs or writing signs. This paper is concerned with sounds.

[2] By the way, the fact that the audible speech stream is continuous and produced by a mixture of articulatory movements is not incompatible with "discreteness": "discreteness" applies to the command level, which specifies articulatory targets in time, which are then sequentially and continuously reached by the articulators under the control of a low-level motor controller

Property 3 (diversity): the speakers of a particular language use the same phonemes and they categorize speech sounds in the same manner. Yet, they do not necessarily pronounce each of them exactly the same way.

Property 4 (sharing): At the same time, each language categorizes speech sounds in its own way, and sometimes does it very differently from other languages. For example, Japanese speakers categorize the "l" of "lead" and the "r" or "read" as identical.

This paper will approach the question of how a speech code with these properties may have formed from non-speech prior to the ability to have linguistic interactions.

2 Existing approaches

2.1 The reductionist approach

One of the approaches is "reductionist": it tries to reduce properties of the speech system to properties of some of its parts. In other words, this approach hopes to find a physiological or neural structure whose characteristics are sufficient to deduce the properties of speech.

For example, cognitive innatism (Chomsky and Halle, 1968; Pinker and Bloom, 1990) defends the idea that the brain features a neural device specific to language (the Language Acquisition Device) which knows at birth the properties of speech sounds. This knowledge is supposed to be pre-programmed in the genome. A limit of this approach is that its defenders have remained rather imprecise on what it means for a brain to know innately the properties of language. In other words, this hypothesis is not naturalized. Also, no precise account of the origins of these innate devices has ever been provided.

Other researchers focus on the vocal tract physics as well as on the cochlea electro-mechanics. For example, they claim that the categories that appear in speech systems reflect the non-linearities of the mapping from motor commands to percepts. Phonemes would correspond to articulatory configurations for which small changes lead to small changes in the produced sound. (Stevens, 1972) defends this idea. There is no doubt that the morpho-perceptual apparatus influences the shape of speech sounds. Yet, this reductionist approach has straightforward weaknesses. For example, it does not explain the large diversity of speech systems in the world's languages (Maddieson, 1984). Also, there are many experiments which show that the zones of non-linearity in perception of some languages are not compatible with those of some other ones (e.g. Japanese do not make any perceptual difference between the "l" of "lead" and the "r" of "read").

Another example of this type of explanation is that of (Studdert-Kennedy and Goldstein, 2002; Studdert-Kennedy, this volume) for the origins of discreteness, or "particulate speech" in his terms. Studdert-Kennedy and Goldstein remark that the vocal apparatus is physiologically composed of discrete independent articulators like the jaw, the tongue, the lips, the velum, etc. This implies that there is some digital re-use in complex utterances due to the independent articulators that move. We completely agree with this remark. Yet, some other aspects of discreteness are not accounted. Indeed, for example, as (Studdert-Kennedy and Goldstein, 2002) note, once you have chosen to use a given set of articulators, there remains the problems of how the continuous space of possible constrictions or timings between gestures is discretized. (Goldstein, 2003) proposed a solution to this question that we will review later in the paper (since it is not reductionist but is a mixture of self-organization and functionalism).

One has to note that this "reductionist" approach proposes answers to the questions concerning the presence of properties 1) and 2) of the speech code, but they address neither the diversity of speech sounds nor the fact that they are shared across communities of agents. They also do not provide answers to the chicken-and-egg problem of the formation of a code. But one has to say that this was certainly not their goal either.

2.2 The functionalist approach

The functionalist approach attempts to explain the properties of speech sounds by relating them to their function. Basically, it answers the "why" question by saying "the system has property N because it helps to achieve function F". It answers the "how" question by saying "systems with property N were formed through Darwinian evolution (genetic or cultural) under the pressure to achieve function F". This approach could also be called "adaptationist"[3]: systems with property N were designed for ("ad") their current utility ("apt"). Note that most often the functionalist explanations take into account the constraints due to brain structure, perceptual and vocal systems.

Typically, in the case of the four properties of speech sounds we are interested in, this function is "communication". This means that the sounds of a speech code may be perceptually distinct enough so that they are not confused and communication can take place. The constraints which are involved typically include a cost of production, which evaluates how much energy is to be spent to produce the sounds. So, in this view, speech sounds are a reservoir of forms which is (quasi-)optimal in terms of perceptual distinctiveness and energy production.

For example, (Lindblom, 1992) showed that if we search for vowel systems which are a good compromise between perceptual distinctiveness and energy cost of articulation, then we find the most frequent vowel systems in human languages. (Lindblom, 1992) also showed similar results concerning the re-use of units to form syllables.

Operational scenarios describing how cultural Darwinian evolution formed these systems have also been described. For example, (de Boer, 2001) built a computer simulation which showed how cultural evolution might have worked,

[3] Here, we use the term adaptationism in its general form: the adaptation may be achieved through genetic or cultural evolution

through processes of imitations among agents. In this simulation, the same mechanism explains both the acquisition of vowels and its formation; this mechanism is imitation. As a consequence, he also proposes an answer to the question: "How are vowel systems acquired by speakers?".

One has to note that the model of de Boer does not deal with questions concerning discreteness (which is built in) and compositionality (indeed, his agents produce only simple static vowel sounds, and compositionality is a property of complex dynamic sounds). Yet, this model is very interesting since it shows a process of formation of a convention, i.e. a vowel system, within a population of agents. This really adds value to the work of Lindblom for example, since it provides a mechanism of (implicit) optimization which Lindblom assumed.

Yet, one has also to remark that the imitation game that agents play is quite complex and requires a lot of assumptions about the capabilities of agents. Each of the agents maintains a repertoire of prototypes, which were associations between a motor program and its acoustic image. In a round of the game, one agent, called the speaker, chose an item of its repertoire, and uttered it to another agent, called the hearer. Then the hearer would search in its repertoire for the closest prototype to the speaker's sound, and produce it (he imitates). Then the speaker categorizes the utterance of the hearer and checks if the closest prototype in its repertoire is the one he used to produce its initial sound. Then he tells the hearer whether it was "good" or "bad". Each item in the repertoires has a score which is used to promote items which lead to successful imitations and prune the others. In case of bad imitations, depending on the scores of the prototype used by the hearer, either this prototype is modified so as to better match the sound of the speaker, or a new prototype is created, as close as possible to the sound of the speaker.

From the description of the game, it is clear that to perform this kind of imitation game, a lot of computational/cognitive power is needed. First of all, agents need to be able to play a game, involving successive turn-taking and asymmetric changing roles. Second, they need to be able to voluntarily try to copy the sound production of others, and be able to evaluate this copy. Finally, when they are speakers, they need to recognize that they are being imitated intentionally, and give feed-back/re-inforcement to the hearer about the success or not. The hearer has to be able to understand the feedback, i.e. that from the point of view of the other, he did or did not manage to imitate successfully.

It seems that the level of complexity needed to form speech sound systems in this model is characteristic of a society of agents which has already some complex ways of interacting socially, and has already a system of communication (which allows them for example to know who is the speaker and who is the hearer, and which signal means "good" and which signal means "bad"). The imitation game is itself a system of conventions (the rules of the game!), and agents communicate while playing it. It requires the transfer of information from one agent to another, and so requires that this information be carried by some shared "forms". So it pre-supposes that there is already a shared system of forms. The vowel systems that appear do not really appear "from scratch". This does not mean at all that there is a flaw in de Boer's model, but rather that it deals with the evolution of language[4] rather than with the origins (or, in other terms it deals with the formation of languageS - "les langues" in French - rather than with the formation of language - " le langage" in French). Indeed, de Boer presented interesting results about sound change, provoked by stochasticity and learning by successive generations of agents. But the model does not address the bootstrapping question: how the first shared repertoire of forms appeared, in a society with no communication and language-like interaction patterns? In particular, the question of why agents imitate each other in the context of de Boer's model (this is programmed in) is open.

Another model in the same spirit was proposed by (Browman and Goldstein, 2000; Goldstein, 2003). This model is very interesting since it is the only one we know, except the work presented in the present paper, which tries to approach the question of the origins of the discretization of the continuum of gestures (they call this "emergence of discrete gestures")[5] They built a simulation in which two agents could produce two gestures, each parameterized by a constriction parameter taken in a continuous one-dimensional space (this space is typically the space of possible places of constrictions, or the continuous temporal interval between two gestures). Agents interacted following the rules of the "attunement game". In one round of the game, both agents produced their two gestures, using for each of them a parameter taken in the continuum with a certain probability. This probability was uniform for both gestures at the beginning of the simulation: this meant that a whole continuum of parameters was used. Then, agents recovered the parameter of the other agent's first gesture, and compared it to the parameter they used themselves. If this matched, then two things occurred: the probability to use this parameter for the first gesture was increased, and the probability to use the same value for the second gesture is decreased. This simulated the idea that agents are attempting to produce both of their gestures differently (so that they are contrasted and can be differentiated), and the idea that they try to produce each of them similarly to the corresponding one of the other agent (so that a convention is established). At the end of the simulations, agents converged to a state in which they used only one value for each gesture, so the space was digitalized, and these pairs of values were the same for the two agents in the same simulation and different in different simulations. Goldstein made simulations using and not using non-linearities of the articulatory to acoustic mapping. Not using it led to the uniform use of all parameters across all simulations, while using it led to the statistical preference of parameters falling in the stable zones of the mapping.

Like de Boer's simulation, in this model agents have coordinated interactions: they follow the rules of a game. Indeed, they both need to produce their gestures together in one round of the game. Secondly, as in the "imitation game", a pressure for differentiating sounds is programmed in, as well as a pressure to copy the parameters of the other agent. This means that it is supposed, as

[4] the evolution of the speech sounds of language to be more precise

[5] There is also the work of (Studdert-Kennedy, this volume), but as explained earlier, it focuses on another kind of discreteness in speech, i.e. the one related to the independent and parallel use of different sets of organs to perform gestures.

in de Boer's work, that agents already live in a community in which complex communication exists. Yet, this was certainly not a concern of the author who proposed this simulation in the context of research in phonology, while we are in this paper in the context of research on the bootstrapping of language. Thus, it remains to be seen how digital speech, which has been argued to be crucial for the rise of language (Studdert-Kennedy and Goldstein, 2002), may have been there without supposing that complex communication has already risen! More precisely, how digital speech may appear without a pressure to contrast sounds? This is one of the issues we propose to solve in the paper, but we will come back to that later on. Also, in the model of Goldstein, one assumption is that agents directly exchange the targets that they used to produce gestures (there is noise, but they are still given targets). Yet, the vocalizations of humans are continuous trajectories, first in the acoustic space, and then in the organ relation space. So what a human gets from the gesture of another is not the target, but the realization of this target which is a continuous trajectory from the start position to the target. And because targets are sequenced, vocalizations do not stop at targets but continue their "road" towards the next target. The task of recovering the targets from the continuous trajectory is very difficult, and at least has not been solved by human speech engineers. Maybe the human brain is equipped with an innate ability to detect events corresponding to targets in the stream, but this is a strong speculation and so incorporating it in a model is a strong (but yet interesting!) assumption. In the present paper, we will not make this assumption: agents will produce complex continuous vocalizations specified by sequences of targets, but will not be able initially to retrieve any kind of "event" that may help them to find out where the targets were. Instead, they will use a time resolution filter which will make that each of the points on the continuous trajectory is considered as a target (while only very few of them actually are targets!). This introduces a huge amount of noise (which is not white, but has a particular structure). Yet, we will show that our society of agents converges to a state in which agents have broken the continuum of possible targets into a discrete repertoire which is shared by the population. Using the structure of the activation of the neural maps of agents, at the end it is possible to retrieve where the targets were (but this will be a result rather than an assumption).

3 The "blind snow-flake maker" approach

Functionalist models have their strengths and weaknesses, which we are not going to detail in this paper (for a discussion, see Oudeyer, 2003). Rather, we will propose another track of research, which we think has been left nearly unexplored in the field of the origins of language. This is what we may call the blind snow-flake maker approach (by analogy with the "blind watch-maker" of (Dawkins, 1986) which illustrated the functionalist approach).

Indeed, there are mechanisms in nature which shape the world, like the one which governs the formation of snow-flakes, which are quite different from Darwinism (Ball, 2001). They are characterized by the property of self-organization,

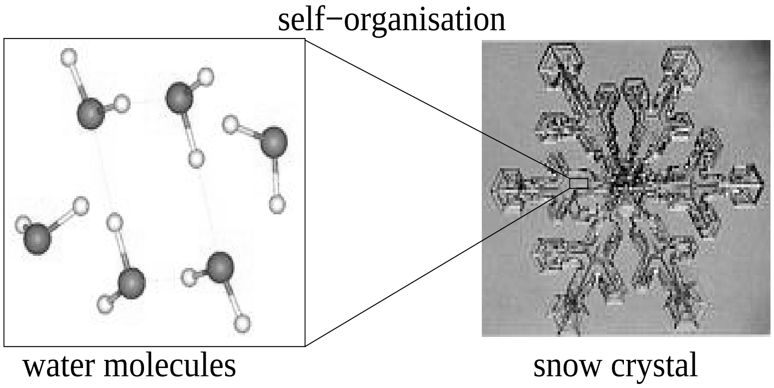

Figure 1: The properties of water molecules and the way they interact are qualitatively different from the symmetrical large-scale structure of snow-crystals: this is self-organization.

like Darwinism, but do not include any concept of fitness or adaptation. Self-organization is here defined as the following property of a system: the local properties which characterize the system are qualitatively different from the global properties of the system. [6]

The formation of snow-crystals is illustrated on figure 1). The local mechanism here at play is the physical and chemical interactions between water molecules. If one looks at these physical and chemical properties, one never finds something that looks like the structure of snow crystals. Yet, if one lets these molecules interact at the right temperature and pressure, marvelous symmetrical structures, with a great diversity in exact shapes, form (Kobayashi et al., 1987). This is an example of a mechanism that shows self-organization, and builds very complex shapes which are not adaptive or functional (it would be hard to claim that it helps the water to survive!). There is no reason why this kind of "free" formation of structures would not appear in the biological world. This idea has been defended by (D'arcy Thompson, 1961; Kauffman, 1995; Gould and Vrba, 1982). D'arcy Thompson presented the example of the formation of the hexagonal honeycomb of the honey bee. Indeed, honey bees build walls of wax made by regular hexagonal cells which tile the whole plane. This is remarkable because 1) there are only three ways to tile the plane with

[6] Note that this definition of self-organization is "non-magical" and different from the one saying that this is a property of systems in which the operations of the higher level cannot be accounted for solely by the laws governing the lower-order level, i.e. can not be predicted from nor reducible to the constituents. We do not include any dimension of surprise in the concept of self-organization (when we will say that the system described in the paper self-organizes, it does not mean at all that its behavior is surprising or unpredictable from the components, but that its global behavior has qualitative properties different from the properties of its constituents).

regular shapes (squares, triangles and hexagons), and 2) hexagons are optimal since it takes less material to cover the same area with hexagons than with triangles or squares. There are two possible ways to account for this. First of all, one could think that honeycombs were designed as an adaptation of the honey bees to minimize their metabolic cost: this is the Darwinist functionalist approach. The honey bees would have tried out many possible shapes until they bumped on hexagons, which they would have found out to be less energy consuming. This would imply that honey bees would have acquired sophisticated instincts that allow them to build perfect hexagons without compasses and set-squares. This explanation is plausible but elaborate, and requires a time-consuming search in the space of forms by the honey bees. A second explanation, proposed by D'Arcy Thompson, is much more straightforward for the honey bees. For him, hexagonal forms are the consequence of purely physical forces: if the wax of the comb is made soft enough by the body heat of the bees, then it is reasonable to think of the compartments as bubbles surrounded by a sluggish fluid. And physics make the bubbles pack together in just the hexagonal arrangement of the honeycomb! (See figure 2) So, the pattern of hexagonal tiling appears spontaneously without pressure to economize energy consumption for the honey bees. Thus, while hexagonal cells are very useful to honey bees, their formation is not explained by this use, but by physical laws which are remote from this use. This is an example of pattern involved in biology which is not due to Darwinism. [7]

The goal of this paper is to present an approach to the formation of speech codes which is very similar in spirit to the approach of D'Arcy Thompson to the formation of honeycombs. We will propose that the formation of sound systems with the properties of discreteness, compositionality, universal tendencies, diversity and sharing, may be a result of the self-organization happening in the interactions of modules which were not necessarily selected for communication. The mechanism is not based on the manipulation of the genetic material, but results from the interaction of agents and from a number of generic neural and physical modules (that may have a function by themselves not related to speech communication) during the lifetime of agents. Note that the scenario that we propose explains how sound systems with the four properties formed before being related to communication, but says nothing about how it could have been recruited later to be used as an information carrier in communication. To come back to the story of bird feathers, it is like explaining how the feathers came up with the thermoregulation pressure, but not saying how the feathers were recruited to fly.

[7] Yet, examples of structures in the biological world which are purely "order for free" are difficult to identify: Darwinism builds other mechanisms which might show self-organization, and these formed mechanisms shape/constrain the space within which Darwinism takes place, so there is a two-way interaction.

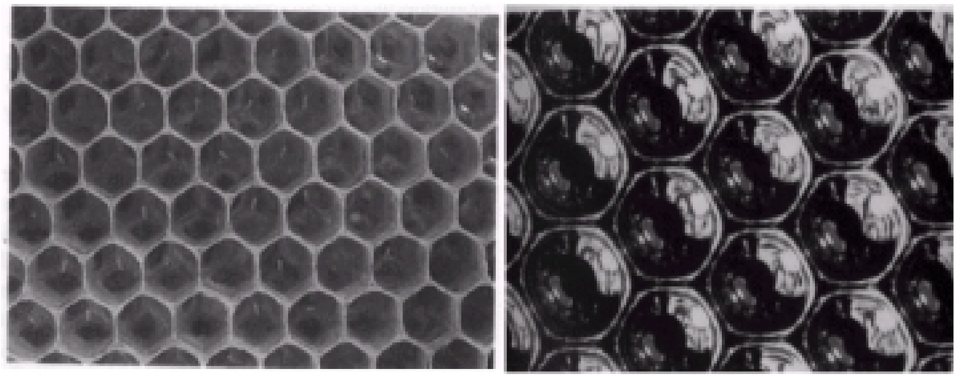

Figure 2: The figure on the left shows the regular hexagonal tiling of the honey comb. This tiling is optimal in terms of wax economy. Hence, a Darwinian explanation might be proposed: the shape of the honeycomb is the result of a long evolution of shapes tried out by honey bees, which finally found out the hexagon and kept it for its optimality. Yet, the figure on the right shows the pattern which is taken by a raft of water bubbles: this is exactly the honey comb pattern! Hence, D'Arcy Thompson proposed that the hexagonal pattern of the honey comb might have a much simpler explanation than the Darwinian functional explanation: it is a "free" result of the physical forces applied to the walls of wax (if honey bees build small cells of rather random shapes, their heat will make them have properties similar to those of water bubbles and hexagons appear naturally).

4 The mechanism

The model is a generalization of the one described in (Oudeyer 2001a), which was used to model a particular phenomenon of acoustic illusion, called the perceptual magnet effect. This model was itself a generalization and unification of the previously existing models of (Damper and Harnad, 2000) and (Guenther and Gajda, 1996).

It is based on the building of an artificial system, composed of robots/agents endowed with working models of the vocal tract, of the cochlea and of some parts of the brain. The complexity and degree of reality of these models can be varied to investigate which aspects of the results are due to which aspects of the model. We also have to repeat again that while some parts of the model are inspired from knowledge in neuroscience, we are not trying to reproduce faithfully what is in the human brain. Rather, we try to build an artificial world in which we can study the phenomenon that we described at the beginning of the paper (i.e. the speech code). Because we know exactly what is happening in this artificial world, in particular what are the assumptions, we hope this will help in our understanding of speech. The way it helps is that it allows us to give sufficient conditions for the appearance of a speech code, and also can tell us what is not necessary for it (e.g. we will show that imitation or feed-back are not necessary). Because the mechanisms that formed speech involve the interaction of many components and complex dynamics, artificial systems are a crucial tool for studying them and it helps to get intuitive understanding of them. Our artificial system aims at proving the self-coherence and logical plausibility of the concept of the "blind snow-flake maker" applied to the origins of digital speech sounds. For more details on this methodology of the artificial, see (Steels, 2001; Oudeyer, 2003).

4.1 The architecture of the artificial system

Here is a summary of the architecture of the system, and in particular the architecture of agents. Technical details can be found in the appendix. Each agent has one ear which takes measures of the vocalizations that it perceives, which are then sent to its brain. It also has a vocal tract, whose shape is controllable and allows it to produce sounds. The ear and the vocal tract are connected to a brain, which is basically a set of interconnected neurons. There are two sets of neurons. A first set is called the "perceptual map", and gets input from the measures taken by the ear. Then, the neurons of the perceptual map can send their output to the second set of neurons, which is called the "motor map" (this could also be called "articulatory map"). These motor neurons can send signals to a controller which drives the vocal tract. These signals should be viewed as commands specifying articulatory targets to be reached in time. The articulatory targets are typically relations between the organs of the vocal tract (like the distance between the lips or the place of constriction). They correspond to what is called a "gesture" in the articulatory phonology literature (Browman and Goldstein, 2000; Studdert-Kennedy, this volume). Figure 3 gives an overview of this architecture. In this paper, the space of organ relations will be two-dimensional (place and manner of constriction) or three-dimensional (place, manner of articulation and rounding).

What we call here a neuron is a box which receives several inputs/measures, and integrates them to compute its activation, which is propagated through output connections. Typically, the integration is made by first weighting each input measure (i.e. multiplying the measure by a weight), then summing these numbers, and applying to the sum a function called the "tuning function". The tuning function is here a gaussian curve, whose width is a parameter of the simulation. A weight is attached to every connection between neurons. In the model, all weights are initially random.

Also, the neurons in each neural map are all interconnected. This means that they receive inputs from all other neurons in the map. When a stimulus is perceived, this gives an initial activation of all the neurons in the two maps. Then, the activation of each neuron, after the update of all the weights, is updated according to the new activation of the neurons to which it is connected. This is repeated until the activations stabilize. This is what is called an attractor in dynamical systems language. This attractor, i.e. a set of neuron activations which is stabilized, is the same for a number of different stimuli, called its basin of attraction. This models categorization behavior. There are as many

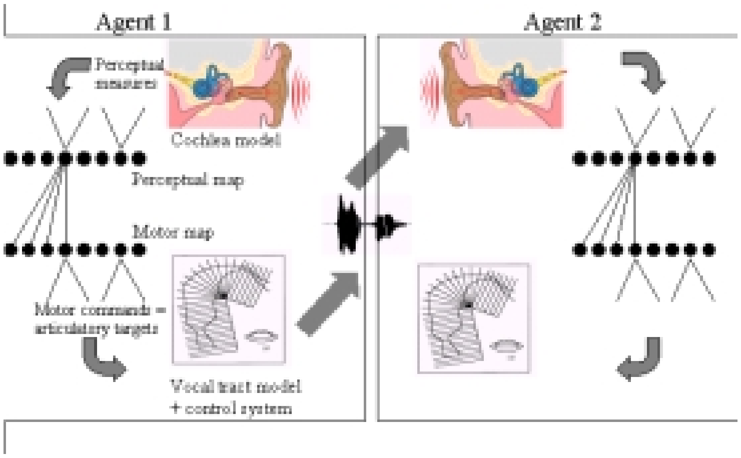
Agent 1
Agent 2
Perceptual measures
Cochlea model
Perceptual map
Motor map
Motor commands = articulatory targets
Vocal tract model + control system

Figure 3: Overview of the architecture

categories as there are attractors.

The production of a vocalization consists in choosing a set of articulatory targets. To choose these targets, one activates sequentially and randomly neurons in the motor map of agents. This activation is a command which we take as the definition of a gesture in this paper. One target is specified by the weights of the output connections of the activated motor neurons. Then there is a control system which executes these commands by pulling continuously and sequentially the organs towards the targets. [8] Here, the control system simply amounts to generating a continuous trajectory in the organ relation space which passes through the targets. This is achieved in the paper through simple spline interpolation, which is basically a polynomial interpolation. Because initially the weights of the connections are random, agents produce vocalizations whose articulatory targets are uniformly spread across the space of possible targets. This implies that their vocalizations are initially analog as far as the commands are concerned (the whole continuum of physically possible commands is used).

[8] It is important to note that this way of producing complex articulation already contains some discreteness. We assume that syllables are specified as a sequence of targets. This is in fact in line with the literature on motor control in mammals (Kandel et al., 2001), which describes it as being organized in two levels: a level of high level discrete commands (our targets), and a low-level which takes care of the execution of these motor commands. So this level of discreteness at the level of commands may not be a feature to be explained in the context of research on the origins of language since it is already present in the motor control architecture of mammals. Yet, we do not suppose that initially these targets are organized: the set of commands used to define targets is taken in a continuum of possible commands and there is no re-use of targets from one syllable to another one; discreteness and compositionality are a result of the simulations. Also, we do not assume that there is discreteness at the perceptual level: agents are not able to detect “events” in the acoustic stream (Yet, at the end they are able to identify the categories of targets which were used to produce the sound).

They are not phonemically coded.

Agents produce vocalizations which are not produced by a static configuration of the vocal tract, but rather a continuous movement of the vocal tract. This implies that agents get a continuous trajectory (in the acoustic space) from the vocalizations of other agents. Now, we explain how this trajectory is processed, and how it is used to change the weights of the connections between the neurons.

First of all, agents are not able to detect high-level events in the continuous trajectory which would allow them for example to figure out which points were the targets that the other agents used to produce it. Rather, they segment the trajectory in very small parts, that corresponds to the time resolution of perception (this models the time resolution of the cochlea). Then, each of these small parts is averaged, giving a value in the acoustic space, which is sent to the perceptual neurons. Each perceptual neuron is then activated.

The weights change each time the neurons to which they are connected are activated. The input connections of the perceptual neurons are changed so that the neurons become more sensitive to the stimuli that activated them, and the change is bigger for neurons whose activation is high than for neurons whose activation is low (this is a sensitization of neurons). Then the activation of the perceptual neurons is propagated to the motor neurons. Then, two possibilities: 1) the motor neurons were already activated because the vocalization was produced by the agent itself, and the weights of the connections between the perceptual and the motor neurons are re-inforced if they correspond to a link between two neurons whose activation is correlated, and weakened if they correspond to a link between neurons whose activation is not correlated (this is hebbian learning). This learning rule allows the agent to learn the mapping between percepts and motor commands during babbling. 2) if the motor neurons were not already activated (the sound comes from the vocalization of another agent), then the weights of the connections between the two maps are not changed. Then, the weights of the connections between the motor neurons and the control system are changed. The neuron with the highest activation in the neural map is selected, and its output weights, which specify one organ relation, are used as a reference to update the other weights: they are changed so that the organ relation they specify looks a little more like it, and this change is weighted by the current activation of each motor neuron.

A crucial point is the coupling between the production process and the perception process. Let us just call the weights of the input connections of the perceptual neurons the preferred vectors of these neurons. This name comes from the fact that the set of weights of a neuron forms a vector, and that the stimulus which has the same values as the weights will activate maximally the neuron. We also call the output weights of the motor neurons their preferred vector. Now, the set up and the dynamics of the two neural maps ensures that the distribution of preferred vectors in the motor map corresponds to the distribution of preferred vectors in the perceptual map: if one activates randomly many times all the neurons in the motor map to produce sounds, then this gives a distribution of sounds which is the same as the one coded by the neurons of the perceptual map. The distribution of the preferred vectors of the neurons in the perceptual map change when sounds are perceived: this implies that if an agent hears certain sounds more often than others, he will tend to produce them also more often than others (here, a “sound” refers to one small sub-part of a vocalization, generated by the time resolution filter described earlier). It is important to see that this process of attunement is not realized through imitation, but is a side effect of an increase of sensitivity of neurons, which is a very generic local low-level neural mechanism (Kandel et al., 2001).

Now, agents are put together in a world in which they will wander randomly. At random times, they produce a vocalization, and agents next to them hear the sound and adapt their neural maps. Each agent also hears its own sounds, using it to learn the mapping from perception to motor commands.

While at the beginning every agent produces sounds whose targets are randomly spread across the continuum, we will show that their neural maps self-organize and synchronize so that after a while they produce complex sounds whose targets belong to a small number of well defined clusters (so the continuum has been digitalized, and because the number of clusters is small compared to the number of vocalizations they produce during their life-time, there is systematic re-use, i.e. compositionality), and that these clusters are the same for all the agents (the code is shared). Moreover, because there are few clusters, and because sounds are made by combining them, targets get re-used automatically and systematically (so compositionality appears). Finally, in each simulation run, the set of clusters which appears is different (so there is diversity).

We use two kinds of models for mapping from motor configurations to sounds and then perception. The first kind is abstract and trivial: this is a random linear mapping from one space to the other space. This allows us to see what we can already get without special properties of the mapping, in particular without non-linearities. In fact, we will show that we get quite far without them: discreteness, compositionality, sharing, diversity. Then, we use a model of the mapping which is more realistic, using three motor parameters: tongue height, tongue body and lip rounding. The formants corresponding to any configurations are then calculated using the model of (de Boer, 2001), which is based on human data. This model allows us to predict the vowel systems which appear in human languages. So it allows us to account for some universal tendencies in human vowel systems.

4.2 Non-assumptions

Agents do not play a language game in the sense used in the literature (Hurford et al., 1998), and in particular do not play the “imitation game” which is for example used in (de Boer, 2001). Their interactions are not structured, there are no roles and no coordination. In fact, they have no social skill at all. They do not distinguish between their own vocalizations and those of others. They do not communicate. Here, “communication” refers to the emission of a signal by an individual with the intention of conveying information which will modify the state of at least one other agent, which does not happen here. Indeed, agents do not even know that there are other agents around them, so it would be difficult to say that they communicate.

5 The dynamics

5.1 Using the abstract linear articulatory/perceptual mapping

The present experiment used a population of 20 agents. Let us describe first what we obtain when agents use the linear articulatory synthesizer. In the simulations, we use 500 neurons per neural map, and $\sigma = 0.05$ (width of their tuning function). The acoustic space and the articulatory space are both two dimensional, with values in each dimension between 0 and 1. These two dimensions can be thought of as the place and the manner of articulation.

Initially, as the preferred vectors of neurons are randomly and uniformly distributed across the space, the different targets that compose the productions of agents are also randomly and uniformly distributed. Figure 4 shows the preferred vectors of the neurons of the perceptual map of two agents. We see that they cover the whole space uniformly. They are not organized. Figure 5 shows the dynamic process of relaxation associated with these neural maps, and due to their recurrent connections. This is a representation of their categorizing behavior. Indeed, each little arrow represents the overall change of activation pattern after one iteration of the relaxation (see the appendix). The beginning of an arrow represents a pattern of activations at time t (generated by presenting a stimulus whose coordinates correspond to the coordinates of this point; this is possible because the population vector is also a decoding scheme which computes the stimulus which activated a neural map). The end of the arrow represents the pattern of activations of the neural map after one iteration of the relaxation. The set of all arrows allows one to visualize several iterations: start somewhere on the figure, and follow the arrows. At some point, for every initial point, you get to a fixed point. This corresponds to one attractor of the network dynamic, and the fixed point to the category of the stimulus that gave rise to the initial activation. The zones defining stimuli which fall in the same category are visible on the figure, and are called basins of attractions. With initial preferred vectors uniformly spread across the space, the number of attractors as well as the boundaries of their basins of attractions are random.

The learning rule of the acoustic map is such that it evolves so as to approximate the distribution of sounds in the environment (but this is not due to imitation!). All agents produce initially complex sounds composed of uniformly distributed targets. Hence, this situation is in equilibrium. Yet, this equilibrium is unstable, and fluctuations ensure that at some point, symmetry breaks: from time to time, some sounds get produced a little more often than others, and these random fluctuations may be amplified through positive feedback loops. This leads to a multi-peaked distribution: agents get in a situation like that of figure 6 (for the unbiased case) which corresponds to figure 4 after 2000 interactions in a population of 20 agents. Figure 6 shows that the distribution of prefered vectors is no longer uniform but clustered. Yet, it is not so easy to visualize the clusters with the representation in figure 6, since there are a few neurons which have preferred vectors not belonging to these clusters. They are not statistically significant, but introduce noise into the representation. Furthermore, in the clusters, basically all points have the same value so that they appear as one point. Figure 7 allows us to visualize better the clusters by showing the attractor landscape that is associated with them. We see that there are now three well-defined attractors or categories, and that there are the same in the two agents represented (they are also the same in the 18 other agents in the simulation). This means that the targets the agents use now belong to one of several well-defined clusters, and moreover can be classified automatically as such by the relaxation of the network. The continuum of possible targets has been broken, sound production is now digital. Moreover, the number of clusters that appear is low, which automatically brings it about that targets are systematically re-used to build the complex sounds that agents produce: their vocalizations are now compositional. All the agents share the same speech code in any one simulation. Yet, in each simulation, the exact set of modes at the end is different. The number of modes also varies with exactly the same set of parameters. This is due to the inherent stochasticity of the process. We will illustrate this later in the paper.

It is very important to note that this result of crystallization holds for any number of agents (experimentally), and in particular with only one agent which adapts to its own vocalizations. This means that the interaction with other agents - i.e. the social component -, is not necessary for discreteness and compositionality to arise. But what is interesting, is that when agents do interact, then they crystallize in the same state, with the same categories. To summarize, there are so far two results in fact: on the one hand discreteness and compositionality arise thanks to the coupling between perception and production within agents, on the other hand shared systems of phonemic categories arise thanks to the coupling between perception and production across agents.

We also observe that the attractors that appear are relatively well spread across the space. The prototypes that their centers define are thus perceptually quite distinct. In terms of Lindblom's framework, the energy of these systems is high. Yet, there was no functional pressure to avoid close prototypes. They are distributed in that way thanks to the intrinsic dynamic of the recurrent networks and their rather large tuning functions: indeed, if two neuron clusters just get too close, then the summation of tuning functions in the iterative process of relaxation smoothes their distribution locally and only one attractor appears.

5.2 Using the realistic articulatory/acoustic mapping

In the previous paragraph, we supposed that the mapping from articulations to perceptions was linear. In other words, constraints from the vocal apparatus due to non-linearities were not taken into account. This was interesting because it showed that no initial asymmetry in the system was necessary to get discreteness

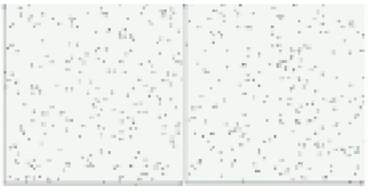

Figure 4: Acoustic neural maps at the beginning. As with all other figures, the horizontal axis represents the first formant (F1), and the vertical axis represents the second effective formant (F2'). The unit is the bark, and they are oriented from low-values to high values.

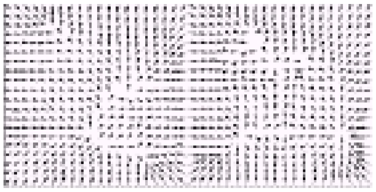

Figure 5: Representation of the two agent's attractor field initially

(which is very asymmetrical). In other words, this shows that there is no need to have sharp natural discontinuities in the mapping from the articulations to the acoustic signals and to the perceptions in order to explain the existence of discreteness in speech sounds (we are not saying that the non-linearities of the mapping do not help, just that they are not necessary!).

Yet, this mapping has a particular shape which introduces a bias into the pattern of speech sounds. Indeed, with the human vocal tract, there are articulatory configurations for which a small change gives a small change in the produced sound, but there are also articulatory configurations for which a small change gives a large change in the produced sound. While the neurons in the motor map have initially random preferred vectors with a uniform distribution, this distribution will soon become biased: the consequence of non-linearities will be that the learning rule will have different consequences in different parts of the space. For some stimuli, a lot of motor neurons will have their preferred vectors shifted a lot, and for other, very few neurons will have their preferred vectors shifted. This will very quickly lead to non-uniformities in the distribution of preferred vectors in the motor map, with more neurons in the parts of the space for which small changes give small differences in the produced sounds, and with less neurons in the parts of the space for which small changes give large differences in the produced sounds. As a consequence, the distribution of the targets that compose vocalizations will be biased, and the learning of the neurons in the perceptual maps will ensure that the distributions of the preferred vectors of these neurons will also be biased.

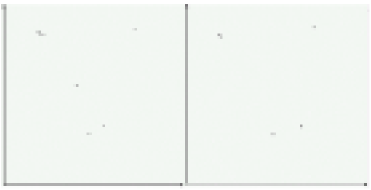

Figure 6: Neural maps after 2000 interactions, corresponding to the intial state of figure 1. The number of points that one can see is fewer than the number of neurons, since clusters of neurons have the same prefered vectors and this is represented by only one point.)

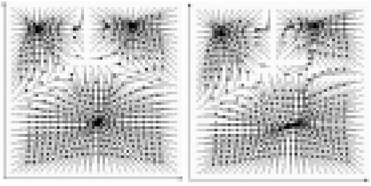

Figure 7: Representation of the attractor fields of 2 agents after 2000 interactions. The number of attractors is fewer that the number of points in the last figure. This is because in the last figures, some points corresponded to clusters and other to single points. The broad width of the tuning function make that the landscape is smoothed and individual point which are not too far from clusters do not manage to form their own basin of attraction.

The articulatory synthesizer used in this part is the one used in (de Boer, 2001). This models only the production of vowels. The fact that agents produce only vocalizations composed of vowel sounds does not imply that the model does not hold for consonants. We chose this articulatory synthesizer because this is the only one which is at the same time fast enough and realistic enough for our computer simulations. The articulatory space (or organ relation space) is here 3-dimensional: tongue height (i.e. manner of articulation), tongue position (i.e. place of articulation), and lip rounding. Each set of values of these variables is then transformed into the first four formants, which are the poles of the vocal tract shaped by the position of the articulators. Then, the effective second formant is computed, which is a non-linear combination of the first 4 formants. The first and effective second formants are known to be good models of our perception of vowels (de Boer, 2001). To get an idea of it, figure 8 shows the state of the acoustic neural maps of one agent after a few interactions between the agents (200 interactions). This represents the bias in the distribution of preferred vectors due to the non-linearities.

A series of 500 simulations was run with the same set of parameters, and each time the number of vowels as well as the structure of the system was checked.

Each vowel system was classified according to the relative position of the vowels, as opposed to looking at the precise location of each of them. This is inspired by the work of (Crothers, 1978) on universals in vowel systems, and identical to the type of classification performed in (de Boer, 2001). The first result shows that the distribution of vowel inventory sizes is very similar to that of human vowel systems (Ladefoged and Maddieson, 1996): figure 10 shows the 2 distributions (in plain line the distribution corresponding to the emergent systems of the experiment, in dotted line the distribution in human languages), and in particular the fact that there is a peak at 5 vowels, which is remarkable since 5 is neither the maximum nor the minimum number of vowels found in human languages. The prediction made by the model is even more accurate than the one provided by (de Boer, 2001) since his model predicted a peak at 4 vowels. Then the structure of the emergent vowel systems was compared to those ones in human languages as reported in (Schwartz et al. 1997b). More precisely, the distributions of structures in the 500 emergent systems was compared to the distribution of structure in the 451 languages of the UPSID database (Maddieson, 1984). The results are shown in figure 11. We see that the predictions are rather accurate, especially in the prediction of the most frequent system for each size of vowel system (less than 8). Figure 9 shows an instance of the most frequent system in both emergent and human vowel systems. In spite of the predictions of one 4-vowels system and one 5-vowels system which appear frequently (9.1 and 6 percent of systems) in the simulations and never appear in UPSID languages, these results compare favorably to those obtained by (de Boer, 2001). In particular, we obtain all this diversity of systems with the appropriate distributions with the same parameters, whereas de Boer had to modify the level of noise to increase the sizes of vowel systems. Yet, like de Boer, we are not able to predict systems with many vowels (which are admittedly rare in human languages, but do exist!). This is certainly a limit of our non functional model. Functional pressure to develop efficient communication systems might be necessary here. As a conclusion of this part, one can say that the model defends the idea that the particular phonemes which appear in human languages are under the influence of the articulatory/perceptual mapping, but that their existence, which means the phenomenon of phonemic coding, is not due to this mapping but to the sensory-motor coupling dynamics.

6 Discussion

A crucial assumption in the artificial system presented in the paper is the fact that there are connections between the motor vocal neural map and the perceptual acoustic map which allow the agents to learn the mapping between the two spaces. How may these connections have appeared ?

First of all, it is possible that they appeared through genetic darwinian evolution under a pressure for language. But the simplicity and genericity of the neural architecture allows other ways of explaining their origins which do not necessitate a pressure for language. These scenarii truly illustrate the "blind

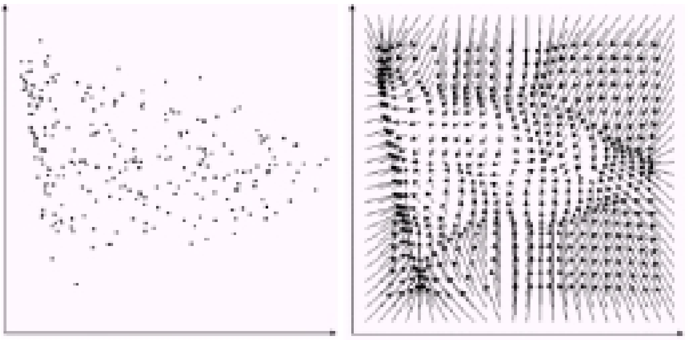

Figure 8: Initial neural map and attractor field of one agent within a population of 20 agents. Here the realistic articulatory synthesizer is used

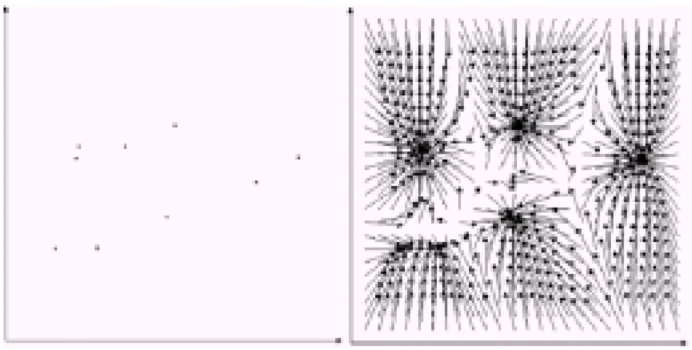

Figure 9: neural map and attractor field of the agent of figure 4 after 2000 interactions with other 20 agents. The corresponding figures of other agents are nearly identical, as in figures 2 and 3. The produced vowel system corresponds to the most frequent 5 vowel system in human languages.

snow-flake maker" approach.

A first alternative scenario is that these connections evolved for imitation. Imitation may have appeared for purposes very different from language: for example, it might have evolved to maintain social cohesion. The copying of behaviors may have been used to mark friendship for example, as in some species of birds. Interestingly, this kind of imitation does not require a system of sounds made of digital units that can be re-used to produce infinite combinations. Also, in this kind of imitation, agents do try to copy behaviours or sounds, but do not try to discriminate sounds. This means that there is no pressure to develop a system of sounds which are different from each other and categorized as such. There is no need to have a system of categories as a whole if the only thing useful is just evaluating the similarity of the sound produced by yourself and the one

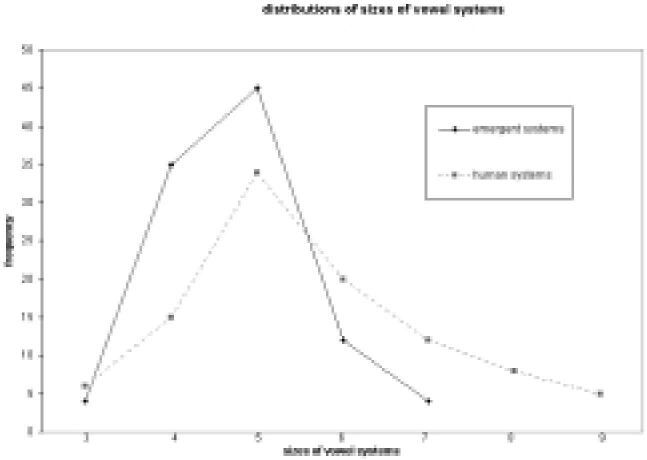

Figure 10: Distribution of vowel inventories sizes in emergent and UPSID human vowel systems

produced by someone else at a given moment. But discreteness, re-use and a system of differences and categories are necessary for speech. The artificial system of this paper precisely shows that with just a simple neural system that may very well have evolved for ”imitation for social cohesion” (making one which is simpler for this task is difficult !), you get freely, through self-organisation, a system of sounds which is shared by a population, digital, with systematically re-used units and a system of categorization. In other words, you get exactly what speech needs without the need for speech. So, even is these neural devices that are assumed in the paper evolved for imitation, they produce speech sounds systems without a functional pressure for speech (as used in the context of language). The simulations of de Boer and Browman and Goldstein do suppose this pressure for speech through the fact that their agents do try to produce the sounds of their repertoire differently (and in the case of de Boer, they try to make the repertoire as big as possible). On the contrary, the agents here do not try to distinguish sounds. The fact that they get to a system of categories which allows them to distinguish sounds is a self-organised result. Also, the notion of repertoire is not pre-programmed, but appears as a result.

A second alternative scenario for the origins of this neural structure which allows to learn the mapping between sounds and articulatory configurations is possible. The system just needs initially random neurons which are sensible to sounds, random neurons which are sensible to motor commands, and random connections between these two sets of neurons. Then it needs that the random connections between these two sets of neurons adapt by following a very general dynamics: hebbian learning. Then, activating randomly and uniformly the motor neurons leads to a movement of the vocal tract which produces

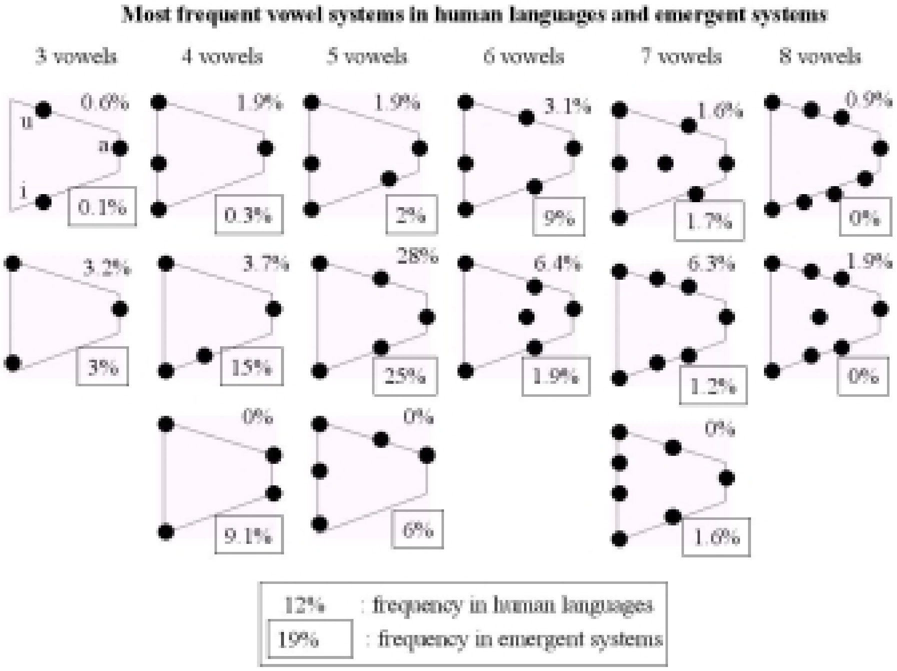

Figure 11: Distribution of vowel inventories structures in emergent and UPSID human vowel systems. This diagram uses the same notations than the one in (Schwartz et al., 1997). Note that here, the vertical axis is also F2, but oriented from high values to low values.

sounds, which in turn acivate the perceptual neurons, and then the connections between the two maps self-organize so that after a while the mapping is effectively learnt. This is what we call babbling. Crucially, this architecture does not require precise pre-wiring during ontogeny that is pre-programmed by the genes. The neurons in the perceptual map and in the motor map certainly existed well before speech (in fact they are there since there are ears and mouths ! So the question of course is how did they come to be connected ? It is quite possible that these connections are a side-effect of general architectural design constraints of the brain ((Gould and Vrba, 1982) give many examples of other features of the bodies and brains of animals which appeared in a similar manner). Indeed, it is obvious that the connections between certain modalities like for example vision and the motor control of arms are necessary and thus existed very early in mammalian evolution. Then, it might very well be that the most efficient strategy to make these connections which are useful for each individual is to connect all modalities rather than to precisely connect particular modalities with other particular modalities. It might be more efficient because it requires fewer specifications for the growth process, and thus might be more robust, and the advantage of this robustness might be superior to the cost of having a priori unnecessary connections. By the way, this way of shaping the brain by initial generation of many random neural structures and then a pruning phase is accepted by a large part of the neuroscience community (see (Changeux, 1981)). But then all mammals should have these connections between neurons that perceive sounds and neurons that control the movements of the mouth. So, why are we the only one to have a system of speech sounds like we have ? And in particular, why monkeys or chimps do not have speech sounds like us ? It is probable that they DO have the connections between the neurons that perceive sounds and those that control the mouth, but that the key is somewhere else. The key is in the BABBLING. Precisely, one of the assumptions that we make and that monkeys or chimps do not seem to implement, is the fact that the agents activate spontaneously, often and randomly the neurons of their motor map. This means that they spontaneously try out many articulatory configurations and repeat these trials. In other terms, they practice. Monkeys of chimps do not practice any motor activity to our knowledge. For example, once they have thrown a stone towards an objective, they will not try to do it again repeatedly. And it seems that a major evolution which gave rise to primitive humans with increased skills as compared to their ancestors is the ability to practice. Primitive humans were babbling all sorts of motor activities. There was certainly a general drive to explore all motor activities available to the body. This drive, obviously beneficial for the learning of many skills useful for the individual, pushed them also to babble with their vocal apparatus. And then we come to the beginning of the simulation presented in the paper, which shows that self-organisation takes place and generates "freely" a system of sounds shared by the agents who live in the same area and which is phonemically coded/digital. The monkeys or chimps may have the connections, but because they do not practice, the neural structures connecting the two modalities certainly die (through the pruning process of activity-dependent neural epigenesis (Changeux, 1981)). On the contrary, the humans do practice, which allows not only to keep the neural system alive, but also to generate a shared speech code.

7 Conclusion

This paper presented a mechanism which provides a possible explanation of how a speech code may form in a society of agents which does not already possess means to communicate and coordinate in a language-like manner. Indeed, as opposed to other computational models of the origins of language (see Hurford et al., 1998), the agents do not play language games. They have in fact no social skills at all. We believe the value of the mechanism we presented is as an example of the kind of mechanism that might solve the language bootstrapping problem. We show how one crucial pre-requisite, i.e. the existence of an organized medium which can carry information in a conventional code shared by a population, may appear without linguistic features being already there.

Furthermore, this same mechanism allows us to account for properties of the speech code like discreteness, compositionality, universal tendencies, sharing and diversity. We believe that this account is original because 1) only one mechanism is used to account for all these properties and 2) we need neither a pressure for efficient communication nor innate neural devices specific to speech (the same neural devices used in the paper can be used to learn hand-eye coordination for example).

Models like the one of de Boer (de Boer, 2001) are to be seen as describing phenomena occuring later in the evolutionary history of language. More precisely, de Boer's model, as well as for example the one presented in (Oudeyer 2001b) for the formation of syllable systems, deals with the recruitement of speech codes like those that appear in this paper, and studies how they are further shaped and developed under functional pressure for communication. Indeed, if we have here shown that one can already go a long way without such pressure, some properties of speech can only be accounted for with it. An example is the phenomenon of chain shifts, in which the prototypes of sounds of a language are all moved around the space.

Yet, in (de Boer, 2001) and (Oudeyer, 2001b), the recruitement of the speech code is pre-programmed. How this could have happened in the origins of language is a problem which remains to be solved. A particular instantiation of the problem is: How do agents come to have the idea to use a speech code to name objects? In fact, the problem of the recruitement of features not initially designed for a certain linguistic function is present at all the levels of language, ranging from sounds to grammar. The question of how recruitement comes about is a major challenge for research on the origins of language. In this paper, we have shown one example of recruitement: individual discrete sounds were systematically re-used in the building of complex vocalizations, and this was not pre-programmed.

8 Acknowledgements

I would like to thank Michael Studdert-Kennedy for his very helpful comments which helped to improve this paper. Note that he does not necessarily agree with what I presented here. I would also like to thank Luc Steels for supporting the research presented in the paper.

9 Appendix: Technical details of the mechanism

The neurons have a gaussian tuning function. If we note $tune_{i,t}$ the tuning function of n_i at time t, s one stimulus vector, v_i the preferred vector (the weights) of n_i, then the form of the function is:

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

[9] The parameter σ determines the width of the gaussian, and so if it is large the neurons are broadly tuned (a value of 0.05 means that a neuron responds substantially to 10 percent of the input space).

When a neuron in the perceptual map is activated because of an input s, then its preferred vector is changed. The mathematical formula of the new tuning 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 input, and $v_{i,t+1}$ the preferred vector of n_i after the processing of s:

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

Also, when a sound is perceived and through propagation activates the motor neurons, the weights of the output connections of these neurons also change. The preferred vector of the most active neuron is taken as a reference: the other preferred vectors are changed so that they get closer to this preferred vector. The change is made with exactly the same formula as for the neurons in the perceptual map, except that s is the preferred vector of the most active neuron.

When an agent hears a vocalization produced by itself, the motor neurons are already activated when the perceived sound activates the neurons in the perceptual map. Then, the weights of the connections between the two neural maps change. A hebbian learning rule is used. If i is a neuron of the perceptual map connected to a neuron j of the motor neural map, then the weight $w_{i,j}$ changes:

$$\delta w_{i,j} = c_2 * (tune_{i,s_i} - < tune_{i,s_i} >)(tune_{j,s_j} - < tune_{j,s_j} >) \text{ (correlation rule)}$$

where s_i and s_j are the input of neurons i and j, $< tune_{i,s_i} >$ the mean activation of neuron i over a certain time interval, and c_2 a small constant. All neurons between the two maps are connected.

[9] the notation $v_1 * v_2$ denotes the scalar product between vector v_1 and vector v_2

Both the perceptual and the motor neural map are recurrent. Their neurons are also connected to each other. The weights are symmetric. This gives them the status of a dynamical system: they have a Hopfield-like dynamics with point attractors, which are used to model the behavior of categorization. The weights are supposed to represent the correlation of activity between neurons, and are learnt with the same hebbian learning rule:

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

These connections are used to relax each neural map after the activations have been propagated and used to change the connections weights. The relaxation is an update of each neuron's activation according to the formula:

$$act(i, t+1) = \frac{\sum_j act(i,t) * w_{i,j}}{\sum_i act(i,t)}$$

where $act(i)$ is the activation of neuron i. This is the mechanism of competitive distribution, together with its associated dynamical properties.

To visualize the evolution of the activations of all neurons during relaxation, we use the "population vector". The activation of all the neurons in a neural map can be summarized by the "population vector" (see Georgopoulos et al., 1988): it is the sum of all prefered vectors of the neurons weighted by their activity (normalized as here we are interested in both direction and amplitude of the stimulus vector):

$$pop(v) = \frac{\sum_i act(n_i) * v_i}{\sum_i act(n_i)}$$

The normalizing term is necessary here since we are not only interested in the direction of vectors.

10 References

Ball P. (2001) The self-made tapestry, Pattern formation in nature, Oxford University Press.

de Boer, B. (2001) The origins of vowel systems, Oxford Linguistics, Oxford University Press.

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

Changeux J.P. (1983) L'homme neuronal, Fayard, 1983.

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

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.

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

Dawkins R. (1986) The Blind Watch-Maker, Penguin Books.

Damper R. and Harnad S. (2000) Neural network modeling of categorical perception. Perception and Psychophysics, 62 p.843-867.

Georgopoulos, Kettner, Schwartz (1988), Primate motor cortex and free arm movement to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. Journal of Neurosciences, 8, pp. 2928-2937.

Guenther and Gjaja (1996) Magnet effect and neural maps, Journal of the Acoustical Society of America, vol. 100, pp. 1111-1121.

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

Gould, S. J., Vrba, E. S. (1982) Exaptation: A missing term in the science of form. Paleobiology 8:4-15.

Hurford, J., Studdert-Kennedy M., Knight C. (1998), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Kandel E.R., Schwartz J.H., Jessell T.M. (2001) Principles of Neural Science McGraw-Hill/Appleton and Lange.

Kauffman S. (1996) At Home in the Universe: The Search for Laws of Self-Organization and Complexity, Oxford University Press.

Kobayashi T., Kuroda T. (1987) Morphology of Crystals, ed. Sunagawa I., Terra Scientific.

Ladefoged, P. and I. Maddieson (1996) The Sounds of the World's Languages. Blackwell Publishers, Oxford.

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.

Maddieson I. (1984) Patterns of sound, Cambridge university press.

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 Origins Of Syllable Systems: an Operational Model. to appear in proceedings of the International Conference on Cognitive science, COGSCI'2001, Edinburgh, Scotland., 2001.

Oudeyer, P-Y. (2003) The Bootstrapping of Speech Sounds in a Society of Artificial Agents, Submitted to Language and Speech.

Pinker, S., Bloom P., (1990), Natural Language and Natural Selection, The Brain and Behavioral Sciences, 13, pp. 707-784.

Schwartz J.L., Boe L.J., Valle N., Abry C. (1997a) The Dispersion/Focalisation theory of vowel systems , Journal of phonetics, 25:255-286, 1997.

Schwartz J.L., Bo L.J., Valle N., Abry C. (1997b) Major trends in vowel systems inventories, Journal of Phonetics, 25, pp. 255-286.

Steels, L. (2001) The methodology of the artificial. Behavioral and brain sciences, 24(6).

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.

Studert-Kennedy M., Goldstein L., (2002) Launching Language: The Gestural Origin of Discrete Infinity, to appear in Christiansen M. and Kirby S. (eds.), Language Evolution: The States of the Art, Oxford University Press.