Connection Science
Vol. 17, No. 3–4, September–December 2005, 325–341

Taylor & Francis
Taylor & Francis Group

The self-organization of combinatoriality and phonotactics in vocalization systems

PIERRE-YVES OUDEYER*

Sony CSL Paris, 6 Rue Amyot, 75005 Paris, France

This paper shows how a society of agents can self-organize a shared vocalization system that is discrete, combinatorial and has a form of primitive phonotactics, starting from holistic inarticulate vocalizations. The originality of the system is that: (1) it does not include any explicit pressure for communication; (2) agents do not possess capabilities of coordinated interactions, in particular they do not play language games; (3) agents possess no specific linguistic capacities; and (4) initially there exists no convention that agents can use. As a consequence, the system shows how a primitive speech code may bootstrap in the absence of a communication system between agents, i.e. before the appearance of language.

Keywords: Origins of speech; Self-organization; Evolution; Phonetics; Phonology; Combinatoriality

1. The complexity of human vocalizations

Human vocalizations have a complex organization. They are characterized by a number of properties that need to be explained.

Discreteness and combinatoriality. Speech sounds are phonemically coded as opposed to holistically coded. This implies two aspects: (1) in each language, the continuum of possible vocalizations is broken into discrete units (this is discreteness) and (2) these units are systematically reused to build higher level vocalization structures like syllables (this is combinatoriality). For example, in articulatory phonology (Browman and Goldstein 1986), 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 harnesses several articulators (e.g. the jaw, the tongue) to produce 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 (e.g. 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. Furthermore, gestures and their combinations, which may be called 'phonemes', are systematically reused in the gestural scores that specify the syllables of each language. This is what we call combinatoriality. Some researchers call the combination of discreteness and combinatoriality 'phonemic coding'.

*Email: py@csl.sony.fr

Connection Science
ISSN 0954-0091 print/ISSN 1360-0494 online © 2005 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/09540090500217145

Phonotactics and patterns. The way phonemes are combined is also very particular: (1) only certain phoneme sequences are allowed to form a syllable in each language, the set of which defines the phonotactics of the language (e.g. 'spink' is a possible syllable in English, but 'npink' and 'ptink' are not possible; in Tashliyt Berber, 'tgzmt' and 'tkSmt' are allowed, but are impossible in French) and (2) the set of allowed phoneme combinations is organized into patterns. This organization into patterns means that, for example, one can summarize the allowed phoneme sequences of Japanese syllables by the patterns 'CV/CVN/VN', where 'CV', for example, defines syllables composed of two slots, and in the first slot only the phonemes belonging to a group that we call 'consonants' are allowed, while in the second slot, only the phonemes belonging to the group that we call 'vowels' are allowed (and N stands for 'nasals').

Universal tendencies. Reoccurring 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 three, four, five or six 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% of languages), while some other vowels are very rare, like [y], [oe] and [ui] (5% of languages). Also, there are structural regularities. For example, if a language contains a front unrounded vowel of a certain height, for example the/e/in 'pet', it will also usually contain the back rounded vowel of the same height, which here would be the/o/in 'pot'. There are also regularities concerning the allowed sequences of phonemes. For example, all languages allow 'CV' syllables, but many disallow clusters of consonants at the beginning of syllables.

Sharing. 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. They also share the same phonotactics.

Diversity. 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' of 'read' as identical. Different languages may also have very different phonotactics.

Where does this organization come from? There are two complementary kinds of answers that must be given (Oudeyer 2003). The first kind is a functional answer that makes a hypothesis about the function of systems of speech sounds, and then shows that systems having the organization described are efficient for achieving this function. This has been proposed, for example, by Lindblom (1992), who showed that discreteness and statistical regularities can be predicted by searching for the most efficient vocalization systems in terms of compromise between perceptual distinctiveness and articulatory cost. This kind of answer is necessary, but not sufficient: it does not say how evolution (genetic or cultural) might have found this optimal structure. In particular, naïve Darwinian search with random mutations (i.e. plain natural selection) might not be sufficient to explain the formation of this kind of complex structure: the search space is just too large (Ball 2001). This is why there needs to be a second kind of answer stating how evolution might have found these structures. In particular, this amounts to showing how self-organization might have constrained the search space and helped natural selection. This can be done by showing that a much simpler system can spontaneously self-organize into the more complex structure that we want to explain.

Self-organization is a phenomenon that is complicated to understand. The computer happens to be the tool most suited for its exploration and understanding (Steels 1997). It is now an essential tool in the domain of human sciences and, in particular, for the study of the origins of language (Cangelosi and Parisi 2002). One of the objectives of this paper is to illustrate

how it can help to develop our intuition about the role of self-organization in the origins of language, and speech in particular.

Examples of works using this methodology have already been developed, e.g. Browman and Goldstein (2000), de Boer (2001) and Oudeyer (2001) concerning speech, and Steels (1997), Kirby (2001), Kaplan (2001) or Cangelosi (2003) concerning lexicons and syntax. As far as speech is concerned, Browman and Goldstein (2000) showed how the continuum of gestures could be discretized, de Boer (2001) showed how a society of agents could develop a shared vowel system, and Oudeyer (2001), building upon the work of de Boer, showed how a society of agents could develop a shared syllable system with basic phonotactic rules. Works like those of Steels (1997), Kirby (2001), Kaplan (2001), de Boer (2001) and Oudeyer (2001) provide an explanation of how a convention like the speech code can be established and propagated in a society of contemporary human speakers. They show how self-organization helps in the establishment of society-level conventions only with local cultural interactions between agents; but they share a number of strong assumptions as far as the capabilities of agents are concerned. Indeed, the interactions between their agents follow the rules of a game that is a complex set of structured conventions. This game is called the 'imitation game' in de Boer (2001) and Oudeyer (2001). It includes, for example, the ability to play changing roles, to understand when one is being imitated or given feedback, or to understand the meaning of a feedback signal. They also share the assumption that agents are provided with the motivation to communicate and form a large repertoire of distinctive vocalizations (there are repulsive forces between the items of their repertoires). These assumptions are interesting and already permit to show a number of crucial results; but they imply that these models deal rather with the cultural evolution of languages than with the origins of language. Indeed, if one wants to understand the origins of language and speech sounds in particular, one needs to understand how the capabilities of the agents that these models assume could have appeared, which is not obvious since they are evolutionarily complex (Oudeyer 2003).

A way to attack this question of the origins of language (speech in particular) is to show how speech codes with the above-mentioned properties could be formed without such complex assumptions. The work described in Browman and Goldstein (2000) was a step in this direction, showing how agents who attuned the distributions of their vocalizations to each other could come to a shared discretization of the articulatory continuum. Yet, it did study static vocalizations (these were points in an abstract one-dimensional space) and involved only two agents that self-organized a repertoire of two different vocalizations. Furthermore, the discretization of the articulatory continuum required the presence of non-linearities in the function that mapped articulatory configurations to perceptions.

Oudeyer (2005) presented another system with evolutionarily simple assumptions, based on the coupling of generic neural devices that were innately randomly wired and implanted in the head of artificial agents. He showed how this system could self-organize so that the agents develop a shared vocalization system with discreteness and statistical regularities, starting from holistic inarticulate vocalizations. The originality of the system was that: (1) it did not include any explicit pressure for communication (e.g. there was no pressure to keep sounds distinctive from each other, as opposed to de Boer (2001) and Oudeyer (2001)); (2) agents did not possess capabilities of co-ordinated interactions, in particular they did not play language games (as opposed to Kaplan (2001), Steels (1997), de Boer (2001) and Oudeyer (2001)); (3) agents did not possess any specific linguistic capacities; (4) initially, there does not exist any convention that agents can use (as opposed to Kaplan (2001) and Kirby (2001) where agents already share a system of strings or literals that they can pass to each other with no ambiguity); and (5) there was no need for non-linearities in the function that maps articulatory configurations to perceptions in order to account for the discretization of the articulatory continuum (as opposed to Browman and Goldstein (2000)).

This system addressed the questions of discreteness, universal tendencies of phoneme repertoires, sharing and diversity. In particular, it predicted the major statistical tendencies characterizing the vowel systems of human languages. However, it addressed the question of combinatoriality only superficially, and did not address at all the questions related to phonotactics and phonological patterns. The goal of this paper is to present an extension of this system, which gives an account of the systematic reuse of speech sounds in the building of complex vocalizations, and of the formation of cultural rules and patterns of sound combination. The extension is based on the addition of a map of neurons with temporal receptive fields. These are initially randomly pre-wired, and control the sequential programming of vocalizations. They evolve with local adaptive synaptic dynamics.

2. The system

In this paper, a summary of the architecture presented in detail in Oudeyer (2005) will be made before presenting the extension. The system is composed of agents that are themselves composed of an artificial brain connected to an artificial vocal tract and an artificial ear. Agents can produce and hear vocalizations. As described in Oudeyer (2005), one can model each component from the most abstract to the most realistic manner. In this paper, the goal is to explore the principles of the formation of phonotactics and of phonological patterns, rather than to build a realistic predictive model. Thus, the most abstract version of the components presented in Oudeyer (2005) will be used. In particular, this means that agents produce two-dimensional vocalizations (one articulatory dimension and one temporal dimension). Only one space is used to represent vocalizations: the perceptual space is bypassed and only the motor space is used. So, we presuppose that agents can translate a vocalization from the perceptual space to the motor space, which is acceptable, since in Oudeyer (2005) it was shown how this mapping could be learnt by the agents. The articulatory dimension that we use is also abstract, but one could imagine that it represents the place or the manner of constriction, for example. Finally, the agents are put in a virtual space in which they wander randomly, and at random times they generate vocalizations that are heard by themselves as well as the closest agent.

The brain of the agent is organized into two neural maps: (1) one 'spatial' neural map coding for static articulatory configurations; and (2) one 'temporal' neural map coding for the sequences of activations of the neurons in the static neural map (this constitutes the extension of the system presented in Oudeyer (2005)).

2.1 *The spatial neural map*

The spatial neural map contains neural units N_i, which have broadly tuned Gaussian receptive fields. We denote by $\mathbf{v}_{i,t}$ the centre of the Gaussian related to N_i, which we call its 'preferred vector' since it corresponds to the stimulus that activates maximally the neural unit. If we do note $G_{i,t}$ the tuning function of N_i at time t, $\boldsymbol{s}$ one input vector, $\mathbf{v}_{i,t}$ the preferred vector of N_i at time t, then:

$$G_{i,t}(\boldsymbol{s}) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{\frac{1}{2}(\mathbf{v}_{i,t}\boldsymbol{s})^2/\sigma^2}.$$

The parameter σ determines the width of the Gaussian, so if it is large the neurons are broadly tuned (a value of 0.05, as used below, means that a neuron responds substantially to 10% of the input space).

All the spatial neural units have initially a random preferred vector, following a uniform distribution. Each neural unit codes for an articulatory configuration, defined by the value of its preferred vector. If the neural unit is activated by the agent and a GO signal is sent to the neural map, then there is a low-level control system that drives the articulators continuously from the current configuration to the configuration coded by the activated neuron.[†] A vocalization is thus here a continuous trajectory in the articulatory space, produced by the successive activation of some neural units in the spatial neural map, combined with a GO signal. As will be seen later on, this activation is controlled internally by the temporal neurons.

As explained earlier, only one space will be used to represent vocalizations. Thus, when an agent produces a vocalization, defined by its trajectory in the articulatory space, the agent that can perceive this vocalization has direct access to the trajectory in the articulatory space. The perception of one vocalization produces changes in the spatial neural map. The continuous trajectory is segmented in small samples corresponding to the cochlea time resolution, and each sample serves as an input stimulus to the spatial neural map. The receptive fields of neural units adapt to these inputs by changing their preferred vector (the width of the Gaussian does not evolve). For each input, the activation of each N_i is computed, and their receptive field updated so that if the same stimulus comes again next time, it will respond a little bit more (this is weighted by their current activation). Basically, adaptation is an increase in sensitivity to stimuli in the environment. The formula is:

$$G_{i,t+1}(\boldsymbol{s}) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{-\frac{1}{2}(\mathbf{v}_{i,t+1}-\boldsymbol{s})^2/\sigma^2},$$

where $G_{i,t+1}$ is the tuning function of N_i at time $t+1$ after the update due to the perception of $\boldsymbol{s}_t$ at time t, and $\mathbf{v}_{i,t+1}$ the updated preferred vector of N_i:

$$\mathbf{v}_{i,t+1} = \mathbf{v}_{i,t} + 0.001 \cdot G_{i,t}(\boldsymbol{s}_t) \cdot (\boldsymbol{s}_t - \mathbf{v}_{i,t}).$$

From a geometrical point of view, the preferred vector of each neural unit is shifted towards the input vector, and the shift is higher for units that respond a lot than for units that do not respond very much.[‡]

2.2 *The temporal neural map*

In Oudeyer (2005), the production of vocalizations was realized by activating randomly neurons in the spatial map. There was no possibility of encoding the order in which the neurons were activated, and as a consequence agents ended up producing vocalizations in which all phoneme combinations were allowed (but of course only the phonemes that appeared as a result of the self-organization of the neural map were used). On the contrary, a temporal neural map will be used that can encode the order of activations of spatial neurons, and is also used to activate the spatial neurons.

Each temporal neuron is connected to several spatial neurons. A temporal neuron can be activated by the spatial neurons through these connections. The tuning function of temporal neurons has a temporal dimension: their activation depends not only on the amplitude of the

[†]There is always only one spatial neuron activated at a time when an agent *produces* a vocalization, as will be explained later on. When a vocalization is *perceived* by the agent, all spatial neurons are activated, but a GO signal is used in that case to trigger a response to the perceived vocalization.

[‡]The neural network used here is technically similar to self-organizing feature maps (Kohonen 1982). In this case, the input space is of the same dimensionality as the output space, so it is not used to make dimensionality reduction. Feature maps are normally used to extract some regularities in high dimensional input data. Here, there is no regularity in the input data initially. Input data are generated by other neural networks of the same kind. Regularities are, rather, created through self-organization as explained in the 'dynamics' section.

activation of the spatial neurons to which they are connected, but also on the order in which they are activated, which itself depends on the particular vocalization that is being perceived. The mathematical formula to compute the activation of the temporal neuron i is:

$$GT_i = \sum_{t=0}^{T} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{\|t\, T_j\|^2/\sigma^2} \cdot \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{\|G_{j,t}\|^2/\sigma^2},$$

with T denoting the duration of the perceived vocalization, N the number of spatial neurons to which it is connected (which is two here, and each temporal neuron is initially connected to two randomly chosen spatial neurons), T_j a parameter that determines when the temporal neuron i is sensitive to the activation of the spatial neuron j, and $G_{j,t}$ the activation of the spatial neuron, j at t. Here, the T_j values are such that the temporal neuron that they characterize is maximally activated for a sequence of spatial neuron activation in which two neurons are never maximally activated at the same time and for which the maximal activation is always separated by a fixed time interval. In brief, this means that rhythm is not taken into account in this simulation: we just consider order. Mathematically,

$$T_1 = 0, T_2 = \tau, t_3 = 2 \cdot \tau, \ldots, T_N = (N-1) \cdot \tau,$$

where τ is a time constant.

As stated earlier, the temporal neurons are also used to activate the spatial neurons. The internal activation of one temporal neuron, coupled with a GO signal, provokes the successive activation of the spatial neurons to which it is connected, in the order specified by the T_j parameters. This implies that the temporal pattern is regular, and only one neuron is activated at the same time. In this paper, each temporal neuron will be connected to only two spatial neurons, which means that a temporal neuron will code for a sequence of two articulatory targets ($N = 2$). This will allow us to represent easily the temporal neural map, but this is not crucial for the results. When an agent decides to produce a vocalization, which it does at random times, it activates one temporal neuron chosen randomly and sends a GO signal.

Initially, a high number of temporal neurons is created (500), connected randomly to the spatial map with random values of their internal parameters. Using many neurons means that basically all possible sequences of activations of spatial neurons are encoded in the initial temporal neural map. The plasticity of the temporal neurons is different from the plasticity of spatial neurons.[†] The parameters of temporal neurons stay fixed during the simulations, but the neurons can die. As a consequence, what changes in the temporal neural map is the number of surviving neurons. The neuronal death mechanism is inspired from apoptosis (Ameisen 2000), and fits with the theory of neural epigenesis developed by 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 are provided to the neurons that are often activated, and prevent automatic suicide Ghosh (1996). We apply this principle of generation and pruning to our temporal neurons, and depending on their mean activity level. The mean

[†]Yet, some recent experiments that are not described in this paper because they were not conducted with the same systematicity, indicate that it is possible to use, for both neural maps, the same neural dynamics and still obtain results similar to those presented here. In these experiments, the common neural dynamics was the same as the one used here for the temporal neural map.

activity of a temporal neuron j is computed with the formula:

$$MA_{j,t} = \frac{MA_{j,t-1} \cdot (window - 1) + GT_{j,t}}{window},$$

where *window* has the initial value 50 (the value of the window size influences the speed of convergence, but the system is rather robust in terms of end result if we change it). The initial value $MA_{j,0}$ is equal to $2 \cdot vitalThreshold$. 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 value of this threshold is chosen so that there is not enough potential activity for all the neurons to stay alive: stability arises at the map level only after a certain number of neurons have been pruned.

2.3 *The coupling of perception and production*

The crucial point of this architecture is that the same neural units are used both to perceive and to produce vocalizations, both in the spatial and in the temporal neural map. As a consequence, the distribution of targets that are used for production is the same as the distribution of receptive fields in the spatial neural map, which themselves adapt to inputs in the environment. This implies, for example, that if an agent hears certain sounds more often than others, he will tend to produce them also more often than others. The same phenomenon applies also to the order of the articulatory targets used in the vocalizations. If an agent hears certain combinations often, then this will increase the mean level of activation of the corresponding temporal neurons, which in turn increases their chance of survival and so increases the probability that they will be used to produce the same articulatory targets combinations. These coupling create positive feedback loops that are the basis of the self-organization that will now be described.

One has to note that this is not realized through explicit imitation, defined as the repetition of a sound that has just been perceived, or of a sound that has been perceived before and has been stored explicitly in memory.[†] This is rather a side-effect of an increase of the selectivity of neurons, and of the competition for neurotrophins between the temporal neurons, which are very generic local low-level neural mechanisms. Additionally, agents do not play any language game in the sense used in the literature (Steels 1997). In fact, they have no capacity for co-ordinated protocol-based social interactions. They are just in a world in which they wander around and sometimes produce sounds and adapt to the sounds they hear around them.

3. The dynamic formation of phonotactics and patterns of combinations

In these simulations, a population of 10 agents was used. As initially the preferred vectors of the spatial neurons are random, and as there is a massive number of random temporal neurons, agents produce vocalizations that are holistic and inarticulate: the continuum of possible articulatory targets is used, and nearly all possible sequences of targets are produced. The initial state of both neural map in two agents is represented on figure 1: the spatial map is represented on the x-axis, which shows the preferred vectors, and is also represented on the y-axis, which shows the same information. The temporal map is represented by the

[†]Yet, the existence of a neural structure that allows the mapping between articulation and perception, which might correspond to the so-called 'mirror neurons' (Rizzolatti *et al.* 1996), might still be the result of a phylogenetic evolution that happened under a selective pressure for imitation capabilities. We just say that these structures, which are only a part of a complete imitation machinery, are not used here for imitation, and their existence has the side-effect of participating in the formation of a shared discrete combinatorial speech code.

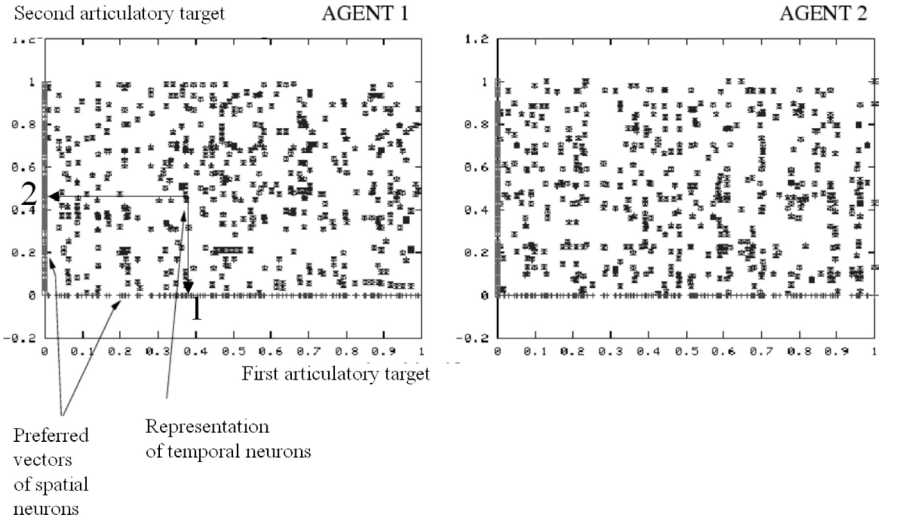

Figure 1. The neural maps of two agents at the beginning of the simulation. The neural map of one agent is represented on the left, and the neural map of the other agent is represented on the right. The spatial neurons are represented by their preferred vectors plotted on the x-axis and also plotted on the y-axis. The temporal neurons are represented by small segments (which nearly appear as points here due to their low level of neurotrophins) whose centre has its x and y corresponding to preferred vectors of the spatial neurons. The x coordinate of a temporal neuron corresponds to the first target that it encodes, and the y coordinate corresponds to the second target that it encodes.

small segments in the middle of the figure, which all correspond to a point (x, y) for which x corresponds to an existing preferred vector in the spatial map, and y to another existing preferred vector in the spatial map. The x coordinate of a temporal neuron corresponds to the first articulatory target of the vocalization that it encodes, and the y coordinate corresponds to the second target that it encodes. The length of the segment represents the level of neurotrophins that each neuron possess.

After several hundred time steps, as shown and explained in detail in Oudeyer, (2005), a clustering of the preferred vectors of the spatial map was observed. Figures 2 and 3 show examples of the neural maps after 1000 interactions in two agents (taken randomly among the

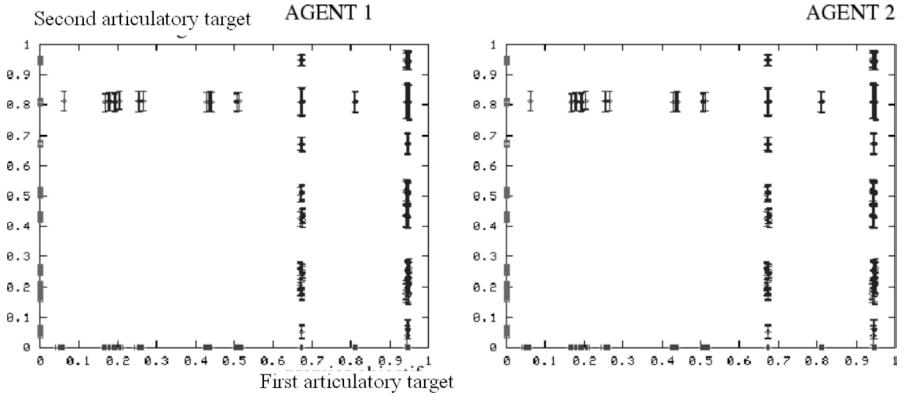

Figure 2. The neural maps of the same two agents after 1000 interactions. We observe: (1) that the preferred vectors of the spatial neural map are now clustered, which means that vocalizations are now discrete: the articulatory continuum has been broken; and (2) that many temporal neurons have died and the surviving ones are organized into lines and columns: this means that phonotactic rules have appeared, that the repertoire of vocalization can be organized into patterns, and that some phonemes are reused systematically for building vocalizations, i.e. vocalizations are now combinatorial.

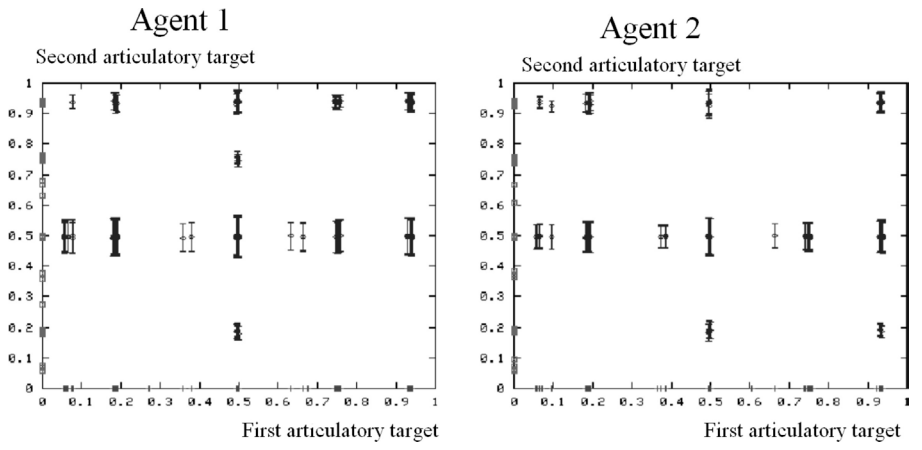

Figure 3. Another example of neural maps of two agents after 1000 interactions in another simulation.

10 agents). Moreover, the clusters are the same for all the agents of the same simulation, and different for agents of different simulations. This shows that now the vocalizations that they produce are discrete: the articulatory targets that they use belong to one of several well-defined clusters, so the continuum of possible targets has been discretized.

Moreover, if we observe the temporal map, we discover that there remain only temporal neurons coding for certain articulatory target sequences. This means that some sequences of targets belonging to the spatial clusters are not produced any more. All the agents of the same population share not only the same clusters in the spatial map, but also the same surviving groups of temporal neurons, as figures 2 and 3 show. This means that rules of phoneme sequencing have appeared, which are shared by all the population. In brief, this is the self-organization of a primitive form of phonotactics. Yet, this is not all that we can observe from the temporal neural map. We also see that the surviving temporal neurons are organized into lines and columns. This means that the set of allowed phoneme sequences can be summarized by patterns. If we call the phonemes associated with the eight clusters of the spatial map on figure 2

$$p_1, p_2, \ldots, p_8,$$

then we can summarize the repertoire of allowed sequences by:

$$(p_6, *), (p_8, *), (*, p_7),$$

where * means 'any phoneme in $p_1, \ldots, p_8$'. This implies that the system of vocalizations that the agents are producing is now combinatorial: some phonemes are reused systematically for the building of different complex vocalizations. The repertoire is thus organized into patterns. Yet, one has to remark that the types of patterns that appear are quite different from the types of patterns of real human languages, such as, for example, the 'CV/CVN/VN' organization of syllables in Japanese. Indeed, in human languages, patterns define slots in which the set of phonemes that can appear are often disjunct: in particular, the consonants set (C) and the vowels set (V) have intrinsic properties that determine their valences and thus their privilege of occurrence in certain slots. So the complexity of the patterns that form in the simulations has not yet reached that of human languages.

The states shown in figures 2 and 3 are convergence states. Indeed, both the states of the spatial map and of the temporal neural map crystallize after a certain amount of time. In Oudeyer (2005), it was explained in detail why the spatial map practically converged into a

set of clusters for a wide range of values of the parameter σ, which determines the dynamics of spatial neurons.

It will now be explained why there is a convergence in the dynamics of the temporal neural map, as figures 4 and 5 show (the evolution of the number of surviving neurons within the temporal maps of two agents has been plotted). As explained above, the initial level of activity ($MA_{j,0}$) of the temporal neurons is set to a constant (2.*vitalThreshold*) that is higher than the mean level of activity that will actually be computed for each neuron at the beginning of the simulation when they are all still alive. As a consequence, the mean level of activity of all neurons is going to go down at the beginning of a simulation. Since there is stochasticity in the system, due to the random choice of temporal neurons when a vocalization is produced, and also due to the fact that not all uniform distributions of preferred vectors are exactly the same in different agents, not all the $MA_{j,t}$'s will decrease exactly in the same manner. In particular, $MA_{j,t}$ of certain temporal neurons will the go below the vital threshold (*vitalThreshold*) before the others and die (indeed, *vitalThreshold* is chosen so that it is higher than the mean level of activity of neurons if they are all alive). The survival of one temporal neuron in a cluster of the temporal map of one agent *ag* depends on the number of neurons in the corresponding cluster in other agents, whose survival depends in return on the number of neurons in the cluster of the agent *ag*. This creates positive feedback loops: sometimes, and by chance, a number of neurons die in the same cluster of one agent, which favours the death of similar neurons in other agents, because having fever neurons in one cluster or area of the space decreases the probability of producing a vocalization coded by the neurons of this cluster and so decreases the mean level of activity of the corresponding cluster in the other agents. Conversely, clusters composed of neurons with a high mean level of activity will favour the survival of similar clusters in other agents. This interaction between the competition and the co-operation in the clusters of temporal neurons of all agents will push a number of neurons, and a number of clusters of neurons, below the vital threshold, until there remain few enough clusters so that the neurons that compose them are activated often enough to survive and 'live' together. This explains the stabilization observed in figures 4 and 5, where we see the two phases: a first phase of initial and rapid pruning of neurons, and a second phase of stabilization.

The 'co-operation'/positive reinforcement can happen between clusters of temporal neurons coding for the same phonemic sequence, but also between clusters of temporal neurons sharing only one articulatory target at the same location within the vocalization. This is due to the mode of activation of temporal neurons, as detailed in the formula given earlier. For example, let us denote p_1, p_2, p_3 and p_4 as four distinct articulatory targets belonging to four distinct clusters.

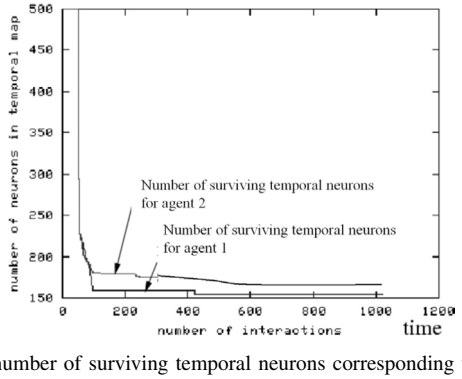

Figure 4. Evolution of the number of surviving temporal neurons corresponding to the temporal neural map of the two agents of figure 2. Observe that there is a first phase of massive pruning, followed by a stabilization that corresponds to a convergence of the system.

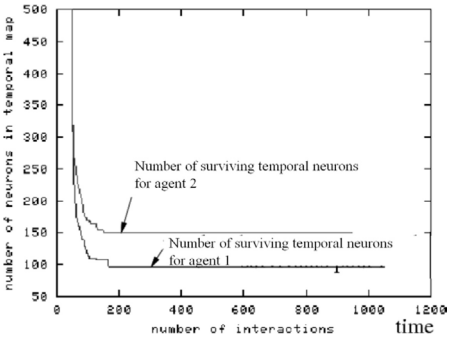

Figure 5. Another example of the evolution of the number of surviving temporal neurons, corresponding to the final neural maps of figure 3. It can be observed that here the two agents do not possess exactly the same number of surviving neurons: this is due to the intrinsic stochasticity of the system. Nevertheless, as figure 3 indicates, they share the same phonotactics and the same patterns.

If the similarity of two vocalizations with the same sequence of phonemes is about unity, 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 zero. 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 agents that 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 agents that 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 neurons map. These are the lines and the columns that we observed in figures 2 and 3, and this explains why we observe the formation of phonological patterns in the phonotactics developed by the agents. To summarize, the interactions between competition and co-operation among individual clusters explains the formation of shared and stable repertoires of allowed phoneme sequences, and the interaction between competition and co-operation among groups of clusters explains the formation of phonological patterns.

4. The influence of articulatory and energetic constraints on statistical preferences in phonotactics

The mechanism presented in the previous section is such that if we run a large number of simulations, there will not be any statistical preference in the localization of clusters of spatial neurons and in the localization of clusters of temporal neurons. We shall now study how an articulatory bias can introduce preferences. As detailed in Oudeyer (2005), a typical articulatory bias is due to the non-linearities of the mapping between the articulatory space, the acoustic space and the perceptual space. Some small changes in the articulatory configuration of the human vocal tract can produce large changes in the acoustic and perceptual image, and vice versa. If one uses an integrated architecture with one articulatory neural map and one acoustic neural map, as in Oudeyer (2005), then even if the preferred vectors of all the neurons of both maps are initially randomly and uniformly spread across the space, their distribution quickly becomes biased by the non-linearities of the mapping (this happens if the two maps are connected so that changes in the distribution of one map are propagated to the other map; see Oudeyer (2005)). In Oudeyer (2005), it was shown how the use of a realistic model of vowel perception and production could implement such a constraint and introduce statistical

preferences in the repertoires of vowels formed by the societies of agents. In particular, it was possible to predict the most frequent vowel systems in human languages.

Here, as only the articulatory representation and its associated neural map are used, this kind of bias will be modelled simply by initially generating a biased distribution of initial random preferred vectors. A distribution was chosen in which there were more preferred vectors close to unity than to zero. This is illustrated by two examples in figure 6. On the one hand, and as explained in detail in Oudeyer (2005), it is easy to see that this will lead to a statistical preference for clusters of spatial neurons with a preferred vector close to unity, and so for phonemes corresponding to articulatory targets close to unity. On the other hand, this bias will also influence the statistical preference of certain kinds of phoneme sequences: as there are more preferred vectors near unity in the spatial neural map, the associated temporal neurons will be activated more often, so their mean level of activity will be higher, which implies that they have a greater chance of surviving. As a consequence, there will be a statistical preference for sequences of phonemes whose articulatory configurations of all targets are close to unity.

Using only this kind of bias is nevertheless too simplistic if one wants to grasp the principles that explain the statistical preferences for certain kinds of phonotactics over other kinds of phonotactics in human languages. Indeed, this kind of bias suggests that phonotactics preferences can be directly derived from phonemic preferences; but this is not at all the case in human languages: vowels like 'a/e/i/o/u' or consonants like 't/m/n' are statistically preferred, but not all syllable sequences composed of these phonemes are statistically frequent in human languages (e.g. 'ta' or 'me' are very frequent, but 'tmn' or 'aet' are very rare). Indeed, the statistical preferences are certainly the outcome of the interaction of several constraints.

This point will be illustrated by introducing another constraint in the system. This is an energetic constraint. In humans, each vocalization involves the displacement of organs, which requires muscular energy: certain vocalizations are easier to pronounce from an energetical point of view than some others. Several researchers (Lindblom 1992, Redford *et al.* 2001) have already proposed that this kind of energy cost is an important component in the formation of human vocalization systems. The energy cost of one vocalization will be modelled here as the amount of displacement of the articulator from a rest position defined as the articulatory configuration of value zero (this is a variant of the energy cost used by Redford *et al.* (2001), which measures the articulatory difference between subsequent phonemes). As the speed of the articulator when it moves is constant here, there is a simple way to compute the energy associated with the vocalization, composed of the targets p_1 and p_2:

$$e(p_1, p_2) = p_1^2 + p_2^2.$$

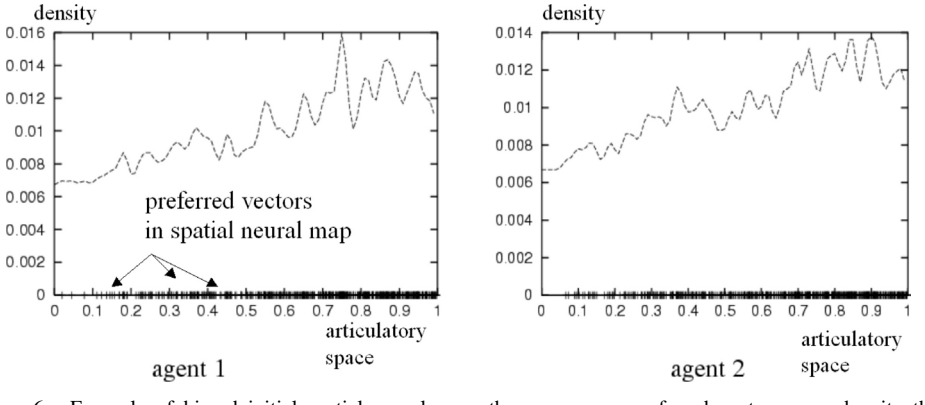

Figure 6. Example of biased initial spatial neural map: there are more preferred vectors around unity than around zero.

This energy will influence the survival of the temporal neurons. Indeed, I explained earlier that the survival of temporal neurons depends on the level of neurotrophins that they receive. A neuron could receive neurotrophin in proportion to its level of activation. The stress associated with the spending of energy can in reverse prevent the reception of neurotrophins (Ghosh 1996). In particular, temporal neurons coding for vocalizations with targets close to zero will be favoured by this constraint as compared with the temporal neurons coding for vocalizations with targets close to unity. Denote $Nt_{i,t}$ as the level of neurotrophins received by the temporal neuron N_i at time t. Then we can compute:

$$Nt_{i,t} = MA_{i,t} - c_1 \cdot e(p_{1,N_i}, p_{2,N_i}),$$

where c_1 is a normalizing constant so that both the terms of activation and of energy have the same ranges, and where p_{1,N_i} and p_{2,N_i} are respectively, the first and second articulatory target encoded by temporal neuron N_i. Again, there is a constant *vitalThreshold* such that if the level of neurotrophins $Nt_{i,t}$ becomes smaller, then the temporal neuron N_i is pruned. This constant is chosen so that not all temporal neurons can survive. Here, $MA_{i,0} = 0.06$, $vitalThreshold = 0.03$, $c_1 = 15$ and there are 150 spatial neurons and 500 initial temporal neurons. Figure 7 gives two examples of initial spatial and temporal neural maps. The segments on the representation of temporal neurons represent here the initial value of their neurotrophin level $Nt_{i,0}$. As the $MA_{i,0}$ component is the same for all temporal neurons, this also gives a representation of the energy cost associated with the vocalizations coded by the temporal neurons (the higher the segment, the lower the energy cost). Only 100 temporal neurons are represented here, instead of 500, for better visibility. This figure shows that there are more temporal neurons with associated targets close to unity, but that these neurons with targets close to unity have individually the lowest level of neurotrophins.

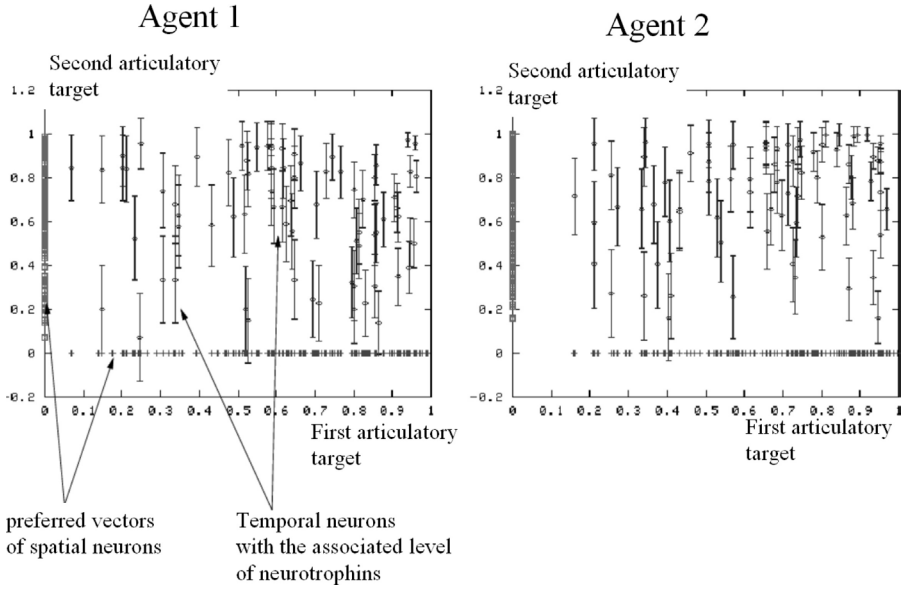

Figure 7. Example of a biased initial temporal neural map. The initial level of neurotrophins associated with temporal neurons, represented by the length of the segments, is shown here. Observe that temporal neurons close to (0, 0) have the largest initial level of neurotrophins, but that the temporal neurons close to (1, 1) are more numerous and so will be activated more often initially, which means that they will receive more neurotrophins than those close to (0, 0).

Let us now run the system and observe how the combination of these two constraints can lead to the formation of phonotactic systems whose statistical properties cannot be deduced from each constraint studied independently. Five hundred simulations were run, a database was made of all the surviving temporal neurons after convergence of the system, and the results plotted in figure 8. It can be observed that there is a clear statistical preference for vocalizations composed of targets located in the centre of the space, and not near zero, as would result from the energetical constraint alone, or near unity, as would result from the non-linearity articulatory constraint alone.

This shows how crucial it is to understand in detail all the constraints influencing the formation of repertoires of vocalizations, as well as the interaction among these constraints, if one wants to understand why, for example, human languages prefer CV syllables to CCVC syllables. This result is positive in the sense that it illustrates the kind of dynamics that can give rise to apparently idiosyncratic phonotactics regularities. This helps us develop our intuition of the self-organized processes that shape vocalization systems; but this result is also negative in the sense that it shows how far we are from being able to predict human languages' statistical preferences in phonotactics. Indeed, our knowledge of the physiological, energetical and representational dimensions of human speech is extremely limited. There are a few areas for which there are probably good models, such as the perception and production of vowels, which allow the use of realistic constraints in a predictive model of the statistical regularities of vowel systems (de Boer 2001, Oudeyer 2005). But, for example, we know very little about the energetical cost of vocalizations, and the existing models of the brain representations of speech signals, which are crucial for the understanding of the articulatory/perceptual non-linearities, are still very speculative. We are not even able to make a list of all the possible constraints that might influence the process of creation of vocalizations. This also explains why, instead of building a system based on very speculative models of realistic constraints, we chose to build a system with completely abstract representations and constraints, which facilitates the understanding of the dynamics.

Finally, it should be said that another constraint that would be very interesting to integrate is the functional constraint. Indeed, for reasons explained in the Introduction, we developed a system free of functional constraint: the agents had no motivation for building a communication system with a repertoire of distinctive vocalizations. It has been shown (Oudeyer 2005) that, even without this motivation and with no repulsive force, still the system self-organized a shared repertoire of vocalizations that could be categorized distinctively. Yet, if we imagine that this system actually describes a process that took place in the evolution of humans before they had language, it was certainly recruited later on in order to communicate. This

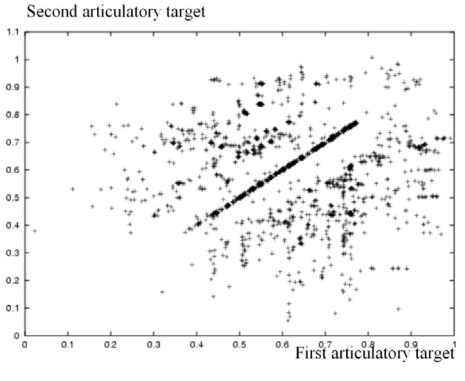

Figure 8. Distribution of surviving temporal neurons in 500 simulations.

means that a functional pressure came in and added new constraints, such as the perceptual distinctiveness between similar vocalizations, which typically would disfavour sequences of identical phonemes like 'aaa' or 'mmm'. This case could be studied by coupling the system described in this paper with the imitation game invented by de Boer (2001) and extended to syllables by Oudeyer (2001).

5. Conclusion

In Oudeyer (2005), a system was presented showing how a society of agents could self-organize a discrete speech code shared by all speakers of the same community, and different in different communities. Also shown was how it allowed the prediction of certain statistical regularities characterizing the repertoires of phonemes in human languages. The originality of the system was that: (1) it did not include any explicit pressure for communication; (2) agents did not possess capabilities of co-ordinated interactions, in particular they did not play language games like the 'imitation game'; (3) agents did not possess any specific linguistic capacities; (4) initially no convention existed that agents could use; and (5) there was no need for non-linearities in the function that maps articulatory configurations to perceptions in order to obtain the discretization of the articulatory continuum.

This made the system a good tool to think about and develop our intuitions about the bootstrapping of speech, and to attack the problem of the origins of language, as opposed to the problem of the formation of languages, which has already been studied extensively in the computer modelling literature (e.g. Kaplan 2001, Kirby 2001, de Boer 2001, Oudeyer 2001, Cangelosi and Parisi 2002). Indeed, by making evolutionarily simpler assumptions than existing models, it allows us to understand how natural selection, in an environment favouring the reproduction of individuals capable of communication, could have been guided by self-organization to establish the first and primitive forms of conventions, such as the speech codes that our agents generate. In this paper, a natural and crucial extension to our earlier work has been presented, introducing a mechanism that takes into account the order of articulatory targets both in production and in perception of vocalizations. This has allowed us to show that, similarly, combinatoriality as well as a primitive form of phonotactics can self-organize in a population of agents. Diversity was again a feature: different populations of agents developed different phonotactics systems. Moreover, the set of allowed phonemic sequences could always be organized into patterns. Yet, these patterns are quite different from the types of patterns of real human languages in which there are phonological categories such as consonants and vowels, which possess disjunctive valences and privileges for the occurrence in certain syllabic slots. Searching for mechanisms that could account for the formation of such phonological categories will be the subject of future work.

Also studied theoretically herein is how the addition of constraints such as non-linearities due to the articulatory/acoustic mapping or such as the energetic cost of vocalizations could influence the statistical preferences of populations of agents for certain kinds of phonotactics. This has shown that if one wants to be able to predict the actual phonotactics preferences in the human languages, then it is crucial to take into account all the constraints as well as their interactions. Unfortunately, the speech sciences are too young and our knowledge of these constraints is today either speculative or not detailed enough. Whereas it is possible to make relatively realistic models of the production and the perception of vowels, which allows one to build predictive models of human vowels systems (e.g. Lindblom 1992, Schwartz *et al.* 1997b, de Boer 2001, Oudeyer 2005), existing models of the production and perception of consonants,

and models of the production and perception of sequenced articulatory targets, can hardly be used in a predictive model of human phonotactics because they would introduce too much *ad hoc* and speculative biases. Indeed, let us take the example of the 'Frame-Content Theory' developed by MacNeilage (1998), which states that vocalizations consist of the deformation of 'frame' cycles of opening and closing the jaw, and thus that vocalizations are subject to the articulatory cost of the deformation of these default cycles. This theory, even if it provides interesting insights into the understanding of speech, does not specify operationally how this cost is computed by the motor system and the neuronal networks to which it is connected. As a consequence, the potential modeller is left with the obligation to invent cost functions, and this will necessarily introduce assumptions that will have a strong impact on the result of a simulation, as has been shown in this paper. There is thus a risk that these assumptions, not founded on real observations, distort the initial qualitative theory (e.g. the 'Frame-Content Theory') and destroy the potential benefits of using it in a simulation.

This is also why, in the work presented in this paper, we preferred to stay at an abstract and theoretical level, which has the advantage of allowing one to understand better the biases that are programmed in, but also to understand the biases that could be introduced by for example, a so-called 'realistic' model of the vocal tract. Owing to these considerations, we believe that the priority in the possible continuation of this work is not to introduce realistic models of the human perceptual and production apparatus for complex vocalizations, but to study the incorporation of a functional pressure for communication. Indeed, it has been shown here that one can already go a long way without functional pressure for communication, but if one wants to bridge the gap with the formation and evolution of contemporary speech systems, it is a necessity to use such a pressure. Indeed, some phenomena can only be accounted for with it, like the existence of large vowels systems (more than 10 vowels), which requires a mechanism of active phonemic creation and repulsive forces among the different phonemic categories. This study could be done by coupling the system presented in this paper with higher level systems like the ones described in de Boer (2001) or Oudeyer (2001).

Acknowledgements

I would like to thank the anonymous reviewers for their valuable comments on this paper. This research was partially supported by the ECAGENTS project founded by the Future and Emerging Technologies programme (IST-FET) of the European Community under EU R&D contract IST-2003-1940.

References

J. Ameisen, *La Sculpture du Vivant. Le Suicide Cellulaire ou la Mort Cratrice*, Paris: Seuil, 2000.
P. Ball, *The Self-made Tapestry, Pattern Formation in Nature*, Oxford: Oxford University Press, 2001.
C. Browman and L. Goldstein, "Towards an articulatory phonology", *Phonology Yearb.*, 3, pp. 219–252, 1986.
C. Browman and L. Goldstein, "Competing constraints on intergestural coordination and self-organization of phonological structures", *Bull. Commun. Parle*, 5, pp. 25–34, 2000.
A. Cangelosi, "Neural network models of category learning and language", *Brain Cognit.*, 53, pp. 106–107, 2003.
A. Cangelosi and D. Parisi, *Simulating the Evolution of Language*, Berlin: Springer, 2002.
J. Changeux and A. Danchin, "The selective stabilization of developing synapses: a plausible mechanism for the specification of neuronal networks", *Nature*, 264, p. 705, 1976.
B. de Boer, *The Origins of Vowel Systems. Oxford Linguistics*, Oxford: Oxford University Press, 2001.
A. Ghosh, "Cortical development: with an eye on neurotrophins", *Curr. Biol.*, 6, pp. 130–133, 1996.
F. Kaplan, *La Naissance d'une Langue chez les Robots*, Paris: Hermes Science, 2001.
S. Kirby, "Spontaneous evolution of linguistic structure—an iterated learning model of the emergence of regularity and irregularity", *IEEE Trans. Evolut. Comput.*, 5, pp. 102–110, 2001.
T. Kohonen, "Self-organized formation of topologically correct feature maps", *Biol. Cybernet.*, 43, pp. 59–69, 1982.

B. Lindblom, "Phonological units as adaptive emergents of lexical development", in *Phonological Development: Models, Research, Implications*, C. Ferguson, L. Menn and C. Stoel-Gammon, Eds, Timonnium, MD: York Press, 1992, pp. 565–604.

P.F. MacNeilage, "The frame/content theory of evolution of speech production", *Behav. Brain Sci.*, 21, pp. 499–511, 1998.

P.-Y. Oudeyer, "The origins of syllable systems: an operational model", in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society, COGSCI'2001*, J. Moore and K. Stenning, Eds, London: Laurence Erlbaum Associates, 2001, pp. 744–749.

P.-Y. Oudeyer, "L'auto-organisation de la parole", PhD thesis, Université Paris VI (2003).

P.-Y. Oudeyer, "The self-organization of speech sounds", *J. Theor. Biol.*, 233, pp. 435–449, 2005.

M.A. Redford, C.C. Chen and R. Miikkulainen, "Constrained emergence of universals and variation in syllable systems", *Language Speech*, 44, pp. 27–56, 2001.

G. Rizzolatti, L. Fadiga, V. Gallese and L. Fogassi, "Premotor cortex and the recognition of motor action", *Cognitive Brain Res.*, 3, pp. 131–141, 1996.

J. Schwartz, L. Boë, N. Vallée and C. Abry, "Major trends in vowel systems inventories", *J. Phonet.*, 25, pp. 255–286, 1997a.

J. Schwartz, L. Boë, N. Vallée and C. Abry, "The dispersion/focalization theory of vowel systems", *J. Phonet.*, 25, pp. 255–286, 1997b.

L. Steels, "The synthetic modeling of language origins", *Evolut. Commun.*, 1, pp. 1–34, 1997.