The self-organization of combinatorial vocalization systems

Pierre-Yves Oudeyer
*Sony CSL Paris, 6, rue Amyot, 75005 Paris, France
http://www.csl.sony.fr/ py

Abstract

In previous papers, we presented a system which showed how a society of agents could self-organize a shared discrete vocalization system, starting from holistic inarticulate vocalisations. The originality of the system was that: 1) it did not include any pressure for communication; 2) it did not include any social capacity (agents did not play a language game for example); 3) it pre-supposed neither linguistic capacities nor the existence of conventions. We present here an extension of the system which shows how rules of sound combination as well as patterns of combinations can self-organize and be shared by the society of agents. This illustrates how phonotactics might have bootstrapped.

1 Introduction

Human vocalizations have a complex organization. They are discrete and combinatorial: they are built through the combination of units, and these units are systematically re-used from one vocalization to the other. These units appear at multiple levels (e.g. the gestures, the coordination of gestures, the phonemes, the morphemes). While the articulatory space that defines the physically possible gestures is continuous, each language only uses a discrete set of gestures. While there is a wide diversity of the repertoires of these units in the world languages, there are also very strong regularities (for example, the high frequency of the 5 vowel system /e,i,o,a,u/). The way the units are combined is also very particular : 1) not all sequences of phonemes are allowed in a given language (this is its phonotactics), 2) the set of allowed phoneme combinations is organized into patterns. This organization into patterns means that for example, one can summarize the allowed phonemes of Japanese by the patterns “CV/CVC/VC”, 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 ”consonant” are allowed, while in the second slot, only the phonemes belonging to the group that we call ”vowels” are allowed.

It is then obvious to ask where this organization comes from. There are two complementary kinds of answers that must be given (Oudeyer, 2003). The first kind is a functional answer stating which is the function of systems of speech sounds, and then showing that systems having the organization that we described are efficient for achieving this function. This has for example been proposed by (Lindblom, 1992) who showed that discretenesss and statistical regularities can be predicted by searching for the most efficient vocalization systems. 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, naive Darwinian search with random mutations (i.e. plain natural selection) might not be sufficient to explain the formation of this kind of complex structures : the search space is just too large (Ball, 2001). This is why there needs a second kind of answer stating how evolution might have found these structures. In particular, this amounts to show how self-organization might have constrained the search space and helped natural selection. This can be done by showing that a much simpler system spontaneously self-organizes into the more complex structure that we want to explain.

(Oudeyer, 2005) has shown how a system of this kind, based on the coupling of generic neural devices which were innately randomly wired and implanted in the head of artificial agents, could self-organize so that the agents develop a shared vocalization system with discreteness, combinatoriality and statistical regularities. The originality of the system was that: 1) it did not include any pressure for communication; 2) it did not include any social capacity (agents did not play a language game for example); 3) it pre-supposed neither linguistic capacities nor the existence of conventions. We present now an extension of this system which gives an account of the formation of rules of sound combination as well as of patterns of sound combinations. This amounts to the formation of phonotactics. 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

We are going to make a summary of the architecture presented in details in (Oudeyer, 2005), before presenting the extension. The system is composed of agents which 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, our goal is to explore the principles of the formation of phonotactics and of phonological patterns, rather than to build a realistic predictive model. Thus, we will use the most abstract version of the components presented in (Oudeyer, 2005). In particular, this means that agents produce two-dimensional vocalizations (one articulatory dimension and one temporal dimension). We use only one space to represent vocalizations: the perceptual space is bypassed and only the motor space is used. So, we pre-suppose that agents can translate a vocalization from the perceptual space to the motor space, which is acceptable since in (Oudeyer, 2005) we showed 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 which 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; 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 $f_{p,i}$ the centre of the gaussian, which we call its "preferred vector" since it corresponds to the stimulus which activates maximally the neural unit. All the 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 which 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 we will see later on, this activation is controlled internally by the temporal neurons.

As we explained earlier, we use only one space to represent vocalizations. Thus, when an agent produces a vocalization, defined by its trajectory in the articulatory space, the agents that can perceive this vocalization gets directly 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 each 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.

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 to encode the order in which the neurons were activated, and as a consequence agents ended by 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, we will use here a temporal neural map which can encode the order of activations of spatial neurons, and is 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 activation of the spatial neurons to which they are connected, but depends also on the

order in which they are activated, which itself depends on the particular vocalization which is being perceived.

As stated in the first paragraph, 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. Here, 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 (the order is coded by some internal parameters of the temporal neuron). This will allow us to represent easily the temporal neural map (but this is not crucial to the result). When an agent decides to produce a vocalization, it activates randomly one temporal neuron and sends a GO signal.

Initially, a high number of temporal neurons are created (500), and are connected randomly to the spatial map with random values of their internal parameters. Using many neurons makes 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 they can die. As a consequence, what changes in the temporal neural map is the number of surviving neurons. The neuron death mechanism is inspired from apoptosis (Ameisen, 2000), and fits with the theory of neural epigenesis developed by (Changeux, 1983). 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 neurotrophines they receive. Neurotrophins are provided to the neurons which are often activated, and prevent them from automatic suicide. We apply this principle of generation and pruning to our temporal neurons, and depending on their mean activity level. There is a $vitalThreshold$ constant which 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 amount 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 which are used for production is the same than 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 feed-back loops which are the basis of the self-organization that we will now describe.

3 The dynamic formation of phonotactics and patterns of combinations

In these simulations, we use a population of 20 agents. 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 which 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 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, and y to another existing preferred vector. 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. The length of the segment represents the level of neurotrophins that each neuron possess.

After several hundred time steps, as we have shown and explained in details in (Oudeyer, 2005), we observe a clustering of the preferred vectors of the spa-

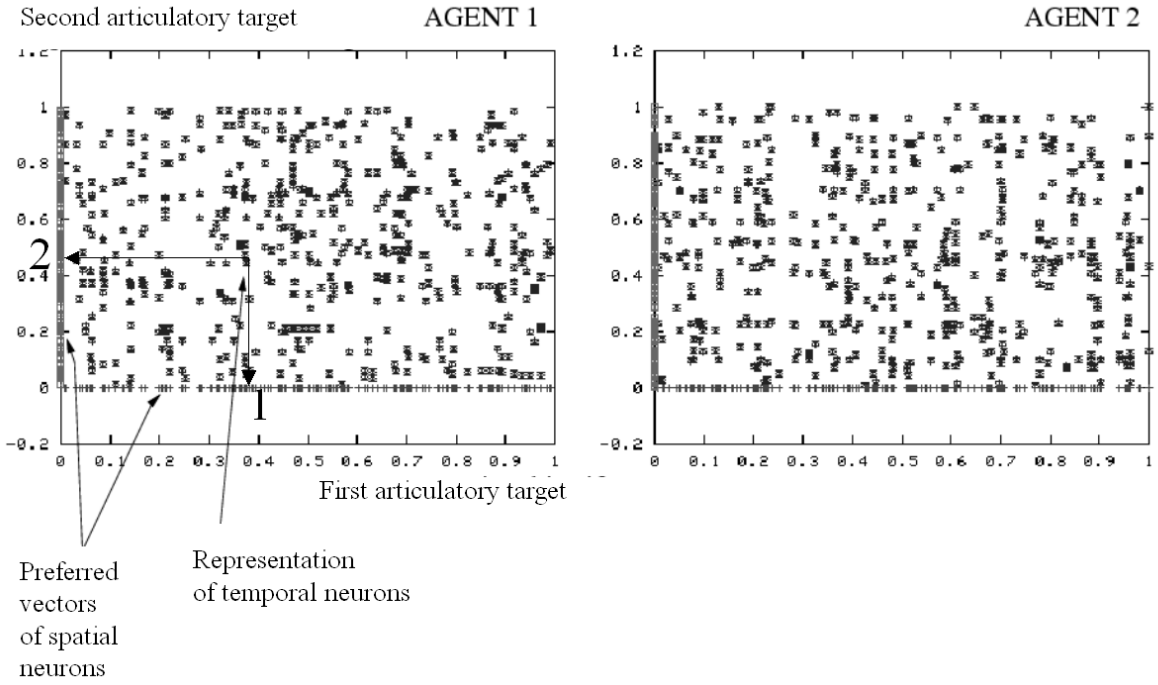

Figure 1: The neural maps of two agents at the beginning of the simulation. The neural maps of one agent is represented on the left, and the neural maps of the other agent are represented on the right. The spatial map is represented by its preferred vectors plotted on the x-axis and also plotted on the y-axis. The temporal neural map is represented by small segments whose center has its x and y corresponding to preferred vectors of the spatial neural map. 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.

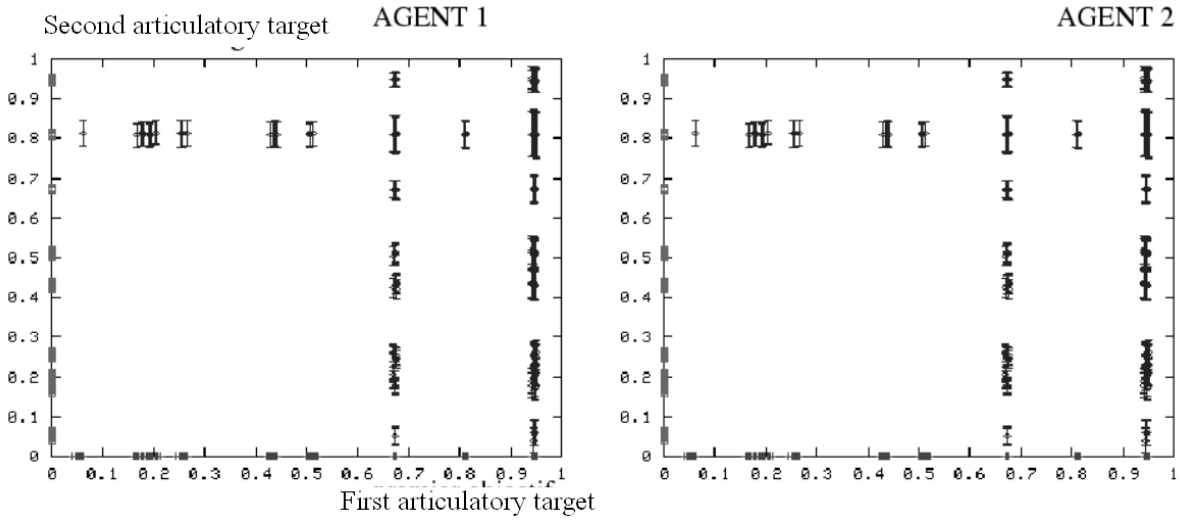

Figure 2: The neural maps of the same two agents after 1000 interactions. We observe: 1) that the preferred vectors of the spatial neural map are now clustered, which means that vocalizations are now discrete: phonemic coding has appeared; 2) that many temporal neurons have died and the surviving ones are organized into lines and columns: this means that phonotactic rules have appeared and moreover that the repertoire of vocalization can be organized into patterns.

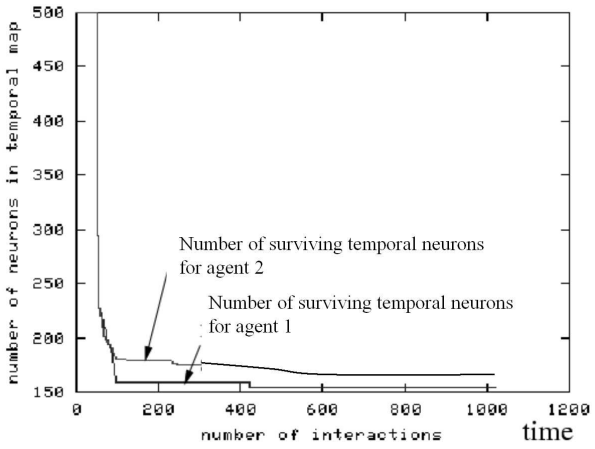

Figure 3: Evolution of the number of surviving temporal neurons corresponding to the temporal neural map of the two agents in figure 2. We observe that there is a first phase of massive pruning, followed by a stabilization which corresponds to a convergence of the system.

tial map. Figure 2 shows an example of the neural maps after 1000 interactions in two 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, and so the continuum of possible targets has been discretized.

Moreover, if we observe the temporal map, we discover that there remains only temporal neurons coding for certain articulatory target sequences. This means that some sequences of targets made possible by the clusters are not produced any more. All the agents of the same population share not only the same clusters in the spatial map, but they also share the same surviving temporal neurons. This means that rules of phoneme sequencing have appeared, which are shared by all the population. In brief, this is the self-organization 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 $p_1, p_2, ..., p_8$, then we can summarize the repertoire of allowed sequences by: $(p_6, *)$, $(p_8, *)$ et $(*, p_7)$ where $*$ means "any phoneme in $p_1, ..., p_8$". The repertoire is thus organized into patterns, in a manner similar for example to the "CV/CVC/VC" organization of syllables in Japanese. As figure 3 shows, this state is the final state of the system: there is a crystallization of the repertoire of vocalizations.

References

J.C. Ameisen. *La Sculpture du vivant. Le suicide cellulaire ou la mort cratrice.* Seuil, 2000.

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

J.P. Changeux. *L'homme neuronal*. Fayard, 1983.

B. Lindblom. Phonological units as adaptive emergents of lexical development. In Ferguson, Menn, and Stoel-Gammon, editors, *Phonological Development: Models, Research, Implications*, pages 565–604. York Press, Timonnium, MD, 1992.

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

P-Y. Oudeyer. The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3): 435–449, 2005.