Phonemic Coding Might Result From Sensory-Motor Coupling Dynamics

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

Résumé

Human sound systems are invariably phonemically coded. Furthermore, phoneme inventories follow very particular tendancies. To explain these phenomena, there existed so far three kinds of approaches: "Chomskyan"/cognitive innatism, morpho-perceptual innatism and the more recent approach of "language as a complex cultural system which adapts under the pressure of efficient communication". The two first approaches are clearly not satisfying, while the third, even if much more convincing, makes a lot of speculative assumptions and did not really bring answers to the question of phonemic coding. We propose here a new hypothesis based on a low-level model of sensory-motor interactions. We show that certain very simple and non language-specific neural devices allow a population of agents to build signalling systems without any functional pressure. Moreover, these systems are phonemically coded. Using a realistic vowel articulatory synthesizer, we show that the inventories of vowels have striking similarities with human vowel systems.

1. The origins of phonemic coding and other related puzzling questions

Human sound systems have very particular properties. First of all, they are phonemically coded. This means that syllables, defined as oscillations of the jaw (MacNeilage, 1998), are composed of re-usable parts. These are called phonemes. Thus, syllables of a language may look rather like la, li, na, ni, bla, bli, etc ... than like la, ze, fri, won, etc .... This might seem unavoidable for us who have a phonetic writing alphabet, but in fact our vocal tract allows to produce syllable systems in which each syllable is holistically coded and has no parts which is also used in another syllable. Yet, as opposed to writing systems for which there exists both "phonetic" coding and holistic/pictographic coding (for e.g. Chinese), all human languages are invariably phonemically coded.

Secondly, the set of re-usable parts of syllable systems, as well as the way they are combined, follows precise and surprising tendancies. For example, our vocal tract allows us to produce hundreds of different vowels. Yet, each particular vowel system uses most often only 5 or 6 vowels, and extremely rarely more than 12 (Maddieson and Ladefoged, 1996). Moreover, there are vowels that appear in these sets much more often than others. For example, most of 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 that caracterize these sets: for example, if a language contains a back rounded vowel of a certain height, for example an [o], it will usually also contain the front, unrounded vowel of the same height.

The questions are then: Why are there these regularities? How did they appear? What are the genetic, glosso-genetic/cultural, and ontogenetic components of this formation process? Several approaches have already been proposed in the litterature.

The first one, known as the "post-structuralist" Chomskian view, defends the idea that our genome contains some sort of program which is supposed to grow a language specific neural device (the so-calles Language Acquisition Device) which knows a priori all the algebric structures of language. This concerns all aspects of language, ranging from syntax (Chomsky, 1958; Archangeli and Langendoen, 1997) to phonetics (Chomsky and Halle, 1968). For example this neural device is supposed to know that syllables are composed of phonemes which are made up by the combination of a few binary features like the nasality or the roundedness. Learning a particular language only amounts to the tuning of a few parameters like the on or off state of these features. It is important to note that in this approach, the innate knowledge is completely cognitive, and no reference to morpho-perceptual properties of the human articulatory and perceptual apparatuses appears. This view is becoming more and more incompatible with neuro-biological findings (which have basically failed to find a LAD), and genetics/embryology which tend to show that the genome can not contain specific and detailed information

for the growth of so complex neural devices. Finally, even if it revealed to be true, it is not really an answer to the questions we asked earlier : it is only a displacement of the problem. How do the concerned genes get there in the course of evolution? Why were they selected? No answer has been proposed by post-structuralist linguistics.

Another approach is that of "morpho-perceptual" innatists. They argue (Stevens 1972) that the properties of human articulatory and perceptual systems explain totally the properties of sound systems. More precisely, their theory relies on the fact that the mapping between the articulatory space and the acoustic and then perceptual spaces is highly non-linear : there are a number of "plateaus" separated by sharp boundaries. Each plateau is supposed to naturally define a category. Hence in this view, phonemic coding and phoneme inventories are direct consequences of the physical properties of the body. Convincing experiments have been conducted concerning certain stop concsonants (Damper 2000) with physical models of the vocal tract and the cochlea. Yet, there are flaws to this view : first of all, it gives a poor account of the great diversity that caracterize human languages. All humans have approximately the same articulatory/perceptual mapping, and yet different language communities use different systems of categories. One could imagine that it is because some "plateaus"/natural categories are just left unused in certain languages, but perceptual experiments (Kuhl 2000) have shown that very often there are sharp perceptual nonlinearities in some part of the sound space for people speaking language L1, corresponding to boundaries in their category system, which are not perceived at all by people speaking another language L2. This means for instance that japanese speakers cannot hear the difference between the "l" in "lead" and the "r" in "read". As a consequence, it seems that there are no natural categories, and most probably the results concerning certain stop consonants are anecdotal. Moreover, the physical models of the vocal tract and of our perceptual system that have been developped in the litterature (Boersma 1998) show clearly that there are important parts of the mapping which is not at all looking like plateaus separated by sharp boundaries. Clearly, considering only physical properties of the human vocal tract and cochlea is not sufficient to explain both phonemic coding and structural regularities of sound systems.

A more recent approach proposes that the phenomena we are interested in come from self-organisation processes occuring mainly at the cultural and ontogenetic scale. The basic idea is that sound systems are good solutions to the problem of finding an efficient communicative system given articulatory, perceptual and cognitive constraints. And good solutions are caracterized by the regularities that we try to explain. This approach was initially defended by (Lindblom 1992) who showed for example that if one optimizes the energy of vowel systems as defined by a compromise between articulatory cost and perceptual distinctiveness, one finds systems which follow the structural and frequency regularities of human languages. (Schwartz et al. 1997) reproduced and extended the results to CV syllables regularities. As far as phonemic cogding is concerned, Lindblom made only simple and abstract experiments in which he showed that the optimal systems in terms of compromise between articulatory cost and acoustic distinctiveness are those in which some targets composing syllables are re-used (note that Lindblom presuposes that syllables are sequences of targets, which we will do also in this paper). Yet, these results were obtained with very low-dimensional and discrete spaces, and it remains to be seen if they remain valid when one deals with realistic spaces. Lindblom proposed another possible explanation for phonemic coding, which is the storage cost argument. It states that re-using parts requires less biological material to store the system, and thus is more advantageous. This argument seems weak for two reasons : first the additional cost of storing un-related parts is not so important, and there are many examples of cultural systems which are extremely memory unefficient (for example the pictogram based writing systems) ; secondly, it does suppose that the possibility of re-using is already there, but what "re-using" means and how it is performed by our neural systems is a fundamental question (this is similar to models of the origins of compositionality (Kirby, 1998) which in fact pre-suppose that the ability to compose basic units is already there, and in fact only show in which conditions it is used or not).

These experiments were a breakthrough as comparared to innatists theories, but provide unsatisfaying explanations : indeed, they rely on explicit optimization procedures, which never occur as such in nature. There are no little scientists in the head of humans which make calculations to find out which vowel system is cheaper. Rather, natural processes adapt and self-organise. Thus, one has to find the processes which formed these sound systems, and can be viewed only a posteriori as optimizations. It has been proposed by (de Boer 2001) that these are imitation behaviors among humans/agents. He built a computational model which consisted of a society of agents playing culturally the so-called "imitation game". Agents were given a physical model of the vocal tract, a model of the cochlea, and a simple prototype based cognitive memory. Their memory of prototypes was initially empty and grew through invention and learning from others, and scores were used to assess them and possibly prune the unefficient ones. One round of the game consisted in picking up two agents, the speaker and the hearer. The speaker utters one sound of its repertoire, and the hearer tries to imitate it. Then the speaker evaluates the imitation by checking if he ca-

tegorizes the imitation as the item he initially uttered. Finally, he gives feedback to the hearer about the result of this evaluation (good or not). de Boer showed that after a while, a society of agents forms a shared vowel system, and that the formed vowel systems follow the structural regularities of human languages. They are somewhat optimal, but this is a side effect to adaptation for efficient communication under the articulatory, perceptual and cognitive pressures and biases. These results were extended by (Oudeyer 2001b) for the case of syllable systems, where phonological rules were shown to emerge within the same process. As far as phonemic coding is concerned, (Oudeyer 2002) has made experiments which tend to indicate that the conclusions drawn from the simple experiments of Lindblom can hardly be extended to realistic settings. It seems that with realistic articulatory and perceptual spaces, non phonemically coded syllable systems that are perfectly sufficient for efficient communication emerge easily. Thus it seems that new hypothesis are needed.

This paper will present a model that follows a similar approach, yet with a crucial difference : no functional pressure will be used here. Another difference is that the cognitive architecture of the agents that we use is modeled at a lower level, which is the neural level. We will show that phonemic coding and shared vowel systems following the right regularities emerge as a consequence of basic sensory-motor coupling on the one hand, and of unsupervised interactions among agents on the other hand. In particular, we will show that phonemic coding can be explained without any reference to the articulatory/perceptual mapping, and yet how this mapping explains some of the structural regularities. The emergent vowel systems will be shown to have great efficiency if they were to be recruited for communication, and yet were not formed under any communicative pressure. This is a possible example of what has been sometimes termed "exaptation". An important aspect to keep in mind is that the neural devices of our agents are very generic and could be used to learn for example hand-eye coordination. Thus they are not at all language specific and at odds with neural devices like the LAD.

2. A low-level model of agents that interact acoustically

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. (Oudeyer 2001a) also described a first simple experiment which coupled agent and neural maps, but it involved only static sounds/articulations and abstract articulatory models. In particular, the question of phonemic coding was not studied. The present paper extends it to dynamic articulations, hence complex sounds, and will use both abstract and realistic articulatory models. We also describe in details the resulting dynamics by introducing entropy-based measures which allow to follow precisely what happens.

The model is based on topological neural maps. This type of neural network has been widely used for many models of cortical maps (Morasso et al., 1998), which are the neural devices that humans have to represent parts of the outside world (acoustic, visual, touch etc...). There are two neuroscientific findings on which our model relies, and that were initially made popular with the experiments of Georgopoulos (1988) : on the one hand, for each neuron/receptive field in the map there exist a stimulus vector to which it responds maximally (and the response decreases when stimuli get further from this vector) ; on the other hand, from the set of activities of all neurons at a given moment one can predict the perceived stimulus or the motor output, by computing what is termed the population vector (see Georgopoulos 1988) : it is the sum of all prefered vectors of the neurons ponderated by their activity (normalized like here since we are interested in both direction and amplitude of the stimulus vector). When there are many neurons and the preferred vectors are uniformly spread across the space, the population vector corresponds accurately to the stimulus that gave rise to the activities of neurons, while when the distribution is inhomogeneous, some imprecisions appear. This imprecision has been the subjects of rich research, and many people proposed more precise variants (see Abbot and Salinas, 1996) to the formula of Georgopoulos because they assumed the sensory system coded exactly stimuli (and hence that the formula of Georgopoulos must be somewhat false). On the contrary we have shown in (Oudeyer 2001a) that this imprecision allows the interpretation of "magnet effect" like psychological phenomena, i.e. sensory illusions, and so may be a fundamental characteristic of neural maps. Moreover, the neural maps are recurrent, and their relaxation consists in iterating the coding/decoding with the population vector : the imprecision coupled with positive feedback loop forming neuron clusters will provide well-define non-trivial attractors which can be interpreted as (phonemic) categories.

A neural map consists of a set of neurons n_i whose "preferred" stimulus vector is noted v_i. The activity of neuron n_i when presented stimulus v is computed with a gaussian function :

$$act(n_i) = e^{-dist(v_i,v)^2/\sigma^2} \quad (1)$$

with sigma being a parameter of the simulation (to which it is very robust). The population vector is then :

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

The normalizing term is necessary here since we are not only interested in the direction of vectors. There are arguments for this being biologically acceptable (see Reggia 1992). Stimuli are here 2 dimensional, corresponding

to the first two formants of sounds. Each neural map is fully recurrent : all the neurons n_i of the map are connected with symmetric synapses of weight

$$w_{i,j} = e^{-dist(v_i,v_j)^2/\sigma^2}$$

, which represent the correlation of activity between 2 neurons (and so could be learnt with a hebbian rule for instance, but are computed directly here for sake of efficiency of the simulation). When presented an input stimulus, two computations take place with a neural map : the population vector is calculated with the initial activation of neurons, and gives what is often interpreted as what the agent senses ; then the network is relaxed using local dynamics : the activity of each neuron is updated as :

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

This is the mechanism of competitive distribution of activation as described in (Reggia et al. 1992, Morasso et al. 1998), together with its associated dynamical properties. The fact that weights are symmetric makes that this dynamic system has point attractors. As a consequence, the iterated update of neurons activity soon arrive at a fixed point, which can be interpreted as a categorizing behavior. So the population vector allows to model how we perceive for example a particular vowel, say [e], in a particular sentence and spoken by a particular person, and the attractor in which the map falls models the behavior of categorizing this sound as being of class [e]. Moreover, it is easy to show that this local dynamics is equivalent to the global process of iteratively coding and decoding with the population vector, and each time feeding back to the input the decoded vector.

There are two neural maps : one articulatory which represents the motor space (neurons n_i), and one acoustic which represent the perceptual space (neurons l_i). Both spaces are typically isomorphic to $[0,1]^n$. The two maps are fully connected to each other : there are symmetric connections of weights $w'_{i,j}$ between every neuron of one map and the other. These weights are supposed to represent the correlation of activity between neurons, and will allow to perform the double direction acoustic/articulatory mapping. They are learnt with a hebbian learning rule (Sejnowsky 1977)

$$\delta w'_{i,j} = c_2(act_i - < act_i >)(act_j - < act_j >)$$

where act_i is the activation of neuron i and $< act_i >$ the mean activation of neuron i over a certain time interval (correlation rule). The propagation of signals in this paper always happen from the acoustic map to the articulatory map (the articulatory map can only give information to the acoustic map through the environment, as in the babbling phase described below where activity in the motor map moves articulators, which then produces sound and activate the cochlea which finnaly activate the acoustic map). Figure 1 gives an overview of the architecture.

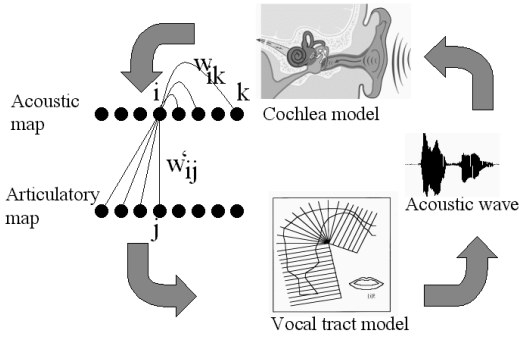

FIG. 1 *Overview of the architecture*

The network is initially made by initializing the preferred vectors of neurons (i.e. their weights in a biological implementations) to random vectors following a uniform distribution, while the $w'_{i,j}$ are set to 0. Part of the initial state can be visualized by plotting all the v_i as in one of the upper squares of figure 2 which represents the acoutic maps of two agents (the perceptual space is 2-dimensional, and points represents the preferred vectors of neurons). Also, one can visualize how agents initially perceive the sound world by plotting all the $pop(v)$ corresponding to a set of stimulus whose vectors values are the intersections of a regular grid covering the whole space. This gives graphs similar to those used in (Kuhl 2000) who discovered the perceptual magnet effect. The lower squares of figure 2 are examples of such an initial perceptual state : we see that the grid is nearly not deformed, which means as predicted by theoretical results, that the population vector is a rather accurate coding of the input space. It means that initially, agents are not subject to auditory illusions as will be the case later on. One can also visualize the initial attractors of the acoustic neural maps : figure 3 shows one example, in which each arrow has its ending point being the population coded vector after one iteration of the relaxation rule with initial activation of neurons corresponding to the population vector represented as the beginning of the arrow. In brief, if one views the relaxation rule as a function which maps one set of neuronal activities, caracterized by a unique population vector, to another one, caracterized by another population vector, then the ending point of arrows are the image of the starting points through this function. What one can notice is that initially, attractors are few, trivial and random (most often there is only one).

Before having agents interact, there is a first phase of babbling during which appropriate initial values of the $w'_{i,j}$'s are learnt with the correlation rule : random articulations are performed, which on the one hand provides activations for the neurons in the articulatory map, and on the other hand produces sounds which are percei-

ved with the cochlea which then provides activations in the acoustic map. Also, acoustic neurons who get very low activity or equivalently whose arriving $w'_{i,j}$ are very low are simply pruned. This is consistent with the well know phenomena of activity dependant growth, and in particular allows a better visualization of the neurons in the acoustic map. Once this initial phase is over, the $w'_{i,j}$'s still continue to evolve with the associated hebbian rule. This is indeed necessary since neurons in each map change their "preferred vector" during the simulation.

Then there is a learning mechanism used to update the weights/preferred vectors in the two neural maps when one agent hears a sound stimulus

$$v = (v_{t_0}, v_{t_1}, v_{t_2}, v_{t_3} ...),$$

which is represented by a temporal sequence of feature vectors here in $[0,1]^n$, typically corresponding to the formants of the sound at a moment t (formants are the frequencies for which there is a peak in the power spectrum). The delay between two feature vectors is typically a few milliseconds, corresponding to the time resolution of the cochlea. For each of these feature vectors, the activation of the neurons (just after perception) in the acoustic map is computed, and propagates to the motor map. Then, the population vector of both maps is computed, giving two vectors v_{acoust} and v_{motor} corresponding to what the networks perceived of the input. Then, each neuron of each map is updated so as to be a little bit more responsive to the perceived input next time it will occur (which means that their preferred vectors are shifted towards the perceived vectors). The actual formula is

$$\delta v_i = c_1 * e^{-\frac{dist(v_{acoustormotor}, v_i)^2}{\sigma^2}} * (v_{acoustormotor} - v_i).$$

The agents in this model produce dynamic articulations. These are generated by choosing N articulatory targets (which are configurations of the vocal tract), and then using a control mechanism which drives the articulators successively to these targets. Articulators are the parts of the vocal tract that control its shape (for example the jaw). In the experiments presented here, N=3 for sake of simplicity, but experiments have been conducted for N=2,...,5 and showed this does not change the results. The choice of the articulatory targets is made by activating successively and randomly 3 neurons of the articulatory map. Their preferred vectors code for the articulatory configuration of the target. The control mechanism that moves the articulators which we used here was very simple: it is simply a linear interpolation between the successive targets. We did not use realistic mechanisms like the propagation techniques of population codes proposed in (Morasso et al., 1998), because these would have been rather computationally unefficient for this kind of experiment, and does not alter the results. Finally, gaussian noise is introduced just before sending the target values to the control system. This noise is fixed in the present paper: the standard deviation of the gaussian is equal to 5 percent of the articulatory range (similar to experiments of de Boer).

When an articulation is performed, a model of the vocal tract is used to compute the corresponding acoustic trajectory. There are two models. The first one is abstract and serves as a test model to see which properties are due to the coupling of neural systems and which are due to the particular shape of the articulatory/acoustic mapping. This is simply a random linear mapping between he articulatory space and the acoustic space. In this paper the articulatory space is always three-dimensional, isomorphic to $[0,1]^3$, and the perceptual space is always 2-dimensional.

The second model is realistic in the sense that it reproduces the human articulatory to perceptual mapping concerning the production of vowels. We model only vowels here for sake of computational efficiency. The three major vowel articulatory parameters are used: (Ladefoged and Maddieson, 1996) tongue height, tongue position and lip rounding. To produce the acoustic output of an articulatory configuration, a simple model of the vocal tract was used, as described in (de Boer, 1999), which generates the 4 first formant values of the sound. Then, from these four values one extracts the first formant and what is called the second effective formant (de Boer, 2001), which is a highly non-linear combination of the first 4 formants. The first and second effective formant are known to represent well human perception of vowels (de Boer, 2001).

The experiment presented consists in having a population of agents (typically 20 agents) who are going to interact through the production and perception of sounds. They are endowed with the neural system and one of the articulatory synthesizers described previously. Each neural map contains 500 neurons in the simulations. Typically, they interact by pairs of two (following the evolutionary cultural scheme devised in many models of the origins of language, see Steels 1997, Steels and Oudeyer, 2000): at each round, one agent is chosen randomly and produces a dynamic articulation according to its articulatory neural map as described earlier. This produces a sound. Then another random agent is chosen, perceives the sound, and updates its neural map with the learning rule described earlier. It is crucial to note that as opposed to all simulations on the origins of language that exist in the litterature (Hurford et al., 1998) our agents do not play here a "language game", in the sense that there is no need to suppose an extra-linguitic protocol of interaction such as who should give a feedback to whom and at what particular moment and for what particular purpose. Indeed, there are no "purpose" in our agents heads. Actually, the simulation works exactly in the same way in the following setup: imagine that agents are in a world in which they have particular locations. Then, the only thing they do is to wander randomly around, pro-

duce sounds at random times, and listen to the sounds that they hear in their neighborhood. In particular, they might not make any difference between sounds produced by themselves and sounds produced by other agents. No concept of "self" is needed. They learn also on their own productions. As a consequence, the model presented here for example makes a lot less assumptions about cognitive and social pre-requisites than the model in (de Boer 2001) for the origins of vowel systems.

3. Shared crystalisation with phonemic coding : the case of abstract linear articulatory/acoustic mappings

Let us describe first what we obtain when agents use the abstract articulator. Initially, as the receptive fields 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. What is very interesting, is that this initial state situation is not stable : rapidly, agents get in the a situation like on figures 4 (for the unbiased case) or 8 which are respectively the correspondances of figures 2 and 8 after 1000 interactions in a population of 20 agents. These figures show that the distribution of receptive fields is not anymore uniform but clustered. The associated point attractors are now several, very well-defined, and non-trivial. Moreover, the receptive fields distribution and attractors are approximately the same for all agents. This means that now the targets that agents use belong to one of well-defined clusters, and moreover can be classified automatically as such by the relaxation of the network. In brief, agents produce phonemically coded sounds. The code is the same for all agents at the end of a simulation, but different across simulations due to the inherent stochasticity of the process.

Also, what we observe is that the point attractors that appear are relatively well spread across the space. The prototypes that they 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 rather large tuning function of receptive fields : indeed, if two neuron clusters just get too close, then the summation of tuning functions in the iterative process of relaxation smoothes locally their distribution and only one attractor appears.

To show this shared crystalisation phenomenon in a more systematic manner, measures were developped that track on the one hand the evolution of the clusteredness of targets for each agent, and on the other hand the similarity of target distributions between agents. The basic idea is to make an approximation of these distributions. A first possibility would have been standard binning, where the space is discretized into a number of boxes, and one counts how many receptive fields fall in each bin, and then normalize. The drawback is that the choice of the bin size is not very convenient and robust; also, as far as distribution comparison is concerned, this can lead to inadequate measures if for example there are small translations among clusters from one distribution to another. As a consequence, we did decide to make approximations of the local probability density function at a set of particular points using parzen windows (Duda et al., 2001). This can be viewed as a fuzzy binning. For a given point x, the approximation of the probability density function is calculated using a gaussian window :

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi\sigma} e^{-\frac{||x-x_i||}{\sigma^2}}$$

where the x_i are the set of targets. The width of the windows is determined by σ, and the range of satisfaying values is large and so this is easy to tune. This approximation is repeated for a set of points distributed on the crossings of a refular grid. Typically, for the 2D perceptual map/space, the grid is 10 points wide and gaussian have a variance equal to that of the noise (5 percent of range).

In order to track the evolution of clusteredness, we chose to use the concept of entropy (Duda et al. 2001). The entropy is minimal for a completely uniform distribution, and maximal for a distribution in which all points are the same (1 perfect cluster). It is defined here as :

$$entropy = -\sum_{i=1}^{l} p_n(xgrid_i) ln(p_n(xgrid_i)$$

where $xgrid_i$ are the crossings of the regular grid at which we evaluated the probability density function. As far as the comparison between two target distributions is compared, one used a symmetric measure based on the Kullback-Leibler distance defined as :

$$distance(p(x),q(x)) = \frac{1}{2}\sum_{xgrid} q(x) log(\frac{q(x)}{p(x)}) + p(x) log \frac{p(x)}{q(x)}$$

where $p(x)$ and $q(x)$ are distributions of targets.

Figure 6 shows the evolution of the mean clusteredness for 10 agents during 2000 games. We clearly see the process of crystalisation. Figure 7 shows the evolution of similarity among the distributions of targets. Each point in the curve represents the mean distance among distributions of all pairs of agents. What we expect is that the curve stays relatively stable, and does not increase. Indeed, initially, all distributions are approximately uniform, so approximately identical. What we verify here is that while each distribution becomes peaked and non-trivial, it remains close to the distributions of other agents.

Why does this phenomenon occur? To understand intuitively, one has to view the neural map that agents use, in particular the perceptual map, as modeling the distribution of sounds that are perceived, and which are produced by members of the society. The crucial point is

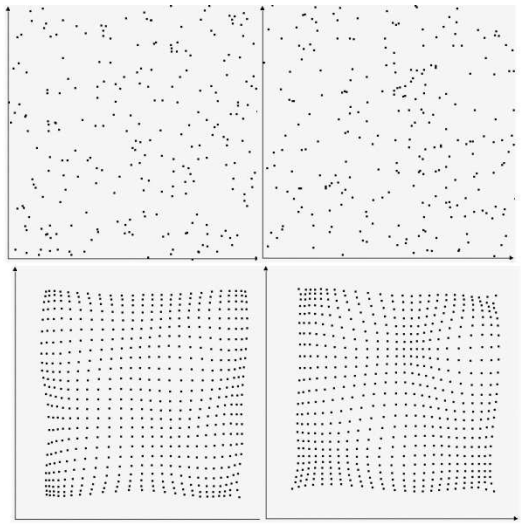

FIG. 2 *Acoustic neural maps at the beginning (top), and associated initial perceptual warping, i.e. images the points of a regular grid through the population vectir function (bottom). As with all other figures, the horizontal axis represents the first formant (F1), and the vertical axis represents the second effective formant (F2')*

that the acoustic map is coupled and evolves with the articulatory map so that the distribution of sounds which are produced is very close to the distribution of sounds which is modeled in the acoustic map. As a consequence, agents learn to produce utterances composed of sounds following the same distribution as what they hear around them. All agents initially produce, and so perceive, the same distribution. Logically, one would expect that this state remains unchanged. Yet, this is not what happens: indeed, at some point, symmetry breaks due to chance. The "uniform" state is unstable. And positive feed-back loops make that this symmetry breaking, which might happen simultaneously in several parts of the space or in several agents, gets amplified and converges to a state in which the distribution is multi-peaked, and of course still shared by agents.

These results show a real alternative to earlier described theories to explain phonemic coding as well as the formation of shared sound systems: the neural device is very generic and could have been used to learn the correspondence between other modalities (e.g. hand-eye coordination, see Morasso et al., 1998, who use similar networks), so no LAD is required (Chomskian innatists); the articulatory to perceptual mapping is linear and trivial, so there is no need for innate particularities of this mapping (morpho-perceptual innatists); agents are not playing any sort of particular language game, and there is no pressure for developing an efficient and shared signalling system (they do develop it, but this is a side effect!), so there are many fewer assumptions needed than

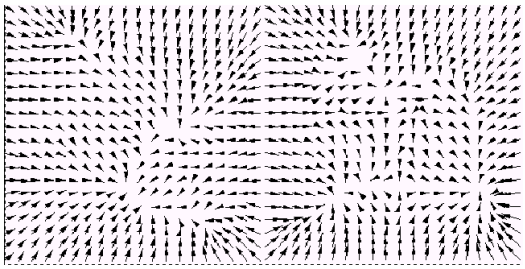

FIG. 3 *Representation of the population vector function for the initial neural maps of figure 1 : each arrow gives information about in which direction are shifted stimuli in the local area where they are drawn*

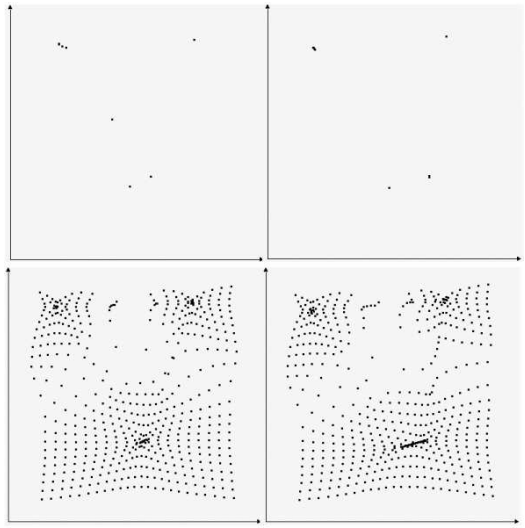

FIG. 4 *Neural maps and perceptual warping after 1000 interactions, corresponding to the intial states of figure 1)*

in Lindblom's or de Boer's approach, and as a consequence the hypothesis presented in this paper should be preferred for simplicity sake, following Occam's razor law.

4. The use of realistic articulatory/acoustic mapping

Yet, we have so far not been able to reproduce the structural regularities of for example human vowel systems as done by de Boer's model. By "structure" we mean what set of vowels (and how many of them) appear together in a vowel system. Indeed, our vocal tract theoretically allows us to have thousands of different vowel systems, but yet only very few are actually used in human languages (Ladefoged and Maddison, 1996). This is due to the fact that we used an abstract articulatory synthesizer. We are now going to use the realistic vowel articulatory synthesizer presented earlier. The mapping

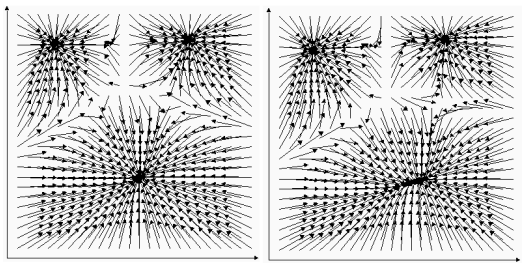

FIG. 5 *Representation of the population vector function for the final neural maps of figure 3 : each arrow gives information about in which direction are shifted stimuli in the local area where they are drawn*

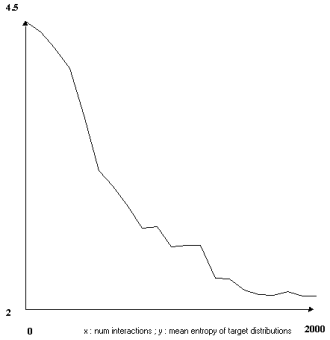

FIG. 6 *Evolution of entropy of target distributions during 2000 interactions : the emergence of clusteredness*

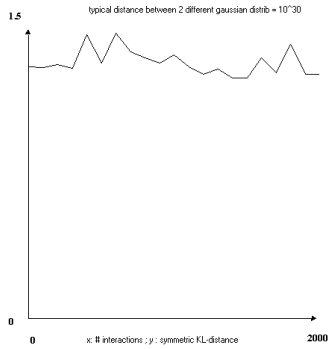

FIG. 7 *Evolution of target distributions similarity during 2000 interactions : emergent clusters are similar in different agents*

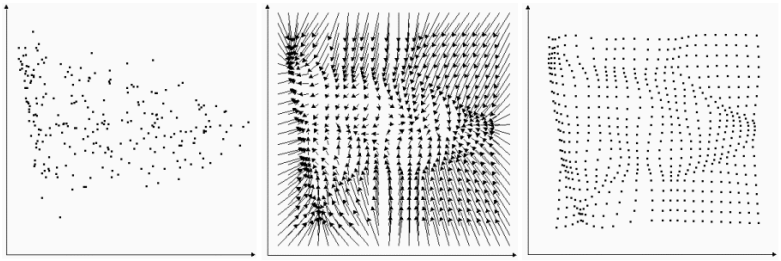

FIG. 8 *Initial neural map, population vector function representation and perceptual warping of one agent within a population of 20 agents. Here the realistic articulatiry synthesizer is used*

that it implements is not any more linear. To get an idea of it, figure 8 shows the state of the acoustic neural maps of agents just after the initial babbling phase which allows to set up initial weights for the connections with the articulatory map, and after the pruning phase which got rid of never used acoustic neurons. We see that the image of the cube $[0,1]^3$ which is uniformly explored during babbling is a triangle (the so-called vocalic triangle). A series of 500 simulations were ran with the same set of parameters, and each time the number of vowels as well as the structure of the system was checked. The first result shows that the distribution of vowel inventory size is very similar to the one of human vowel systems (Ladefoged and Maddison, 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. Also, among these 5 vowel systems, it appeared that one of them is generated much more frequently (79 percent) than others : figure 9 shows an instance of it. The remaining 5 vowel systems are either with a central vowel together with more front vowels, or with more back vowels. This agrees very well with what has been found in natural languages. (Schwartz et al. 1997) found that 89 percent of the languages had the symmetric system, while the two other types with the central vowel occur in 5 percent of the cases. For different system sizes similarly good matches between predicted systems and human vowel systems are found.

5. Conclusion

Functional and computational models of the origins of language (Hurford et al., 1998) typically make a lot of initial assumptions such as the ability to play language games or in general coordinate. The present paper presented an experiment concerning sound systems which might be a possible example of how to bootstrap these linguistic interactions. Indeed, with very simple and non-specific neural systems, without any need for explicit coordination schemes, without any deus ex-machina functional pressure, we obtained here a signalling system which could very well be recruited as a building block for a naming game for example. (Oudeyer 2002) presents a more traditional functional model of higher aspects of sound systems (phonological rules) which is based on the

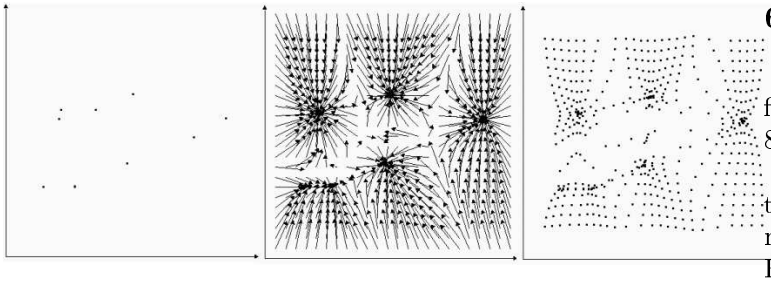

FIG. 9 *neural map, population vector function representation and perceptual warping of the agent of figure 4 after 2000 interactions with other 20 agents. The corresponding figures of other agents are nearly identical, as in figures 2 and 3. The produced vowel system corrresponds to the most frequent 5 vowel system in human languages.*

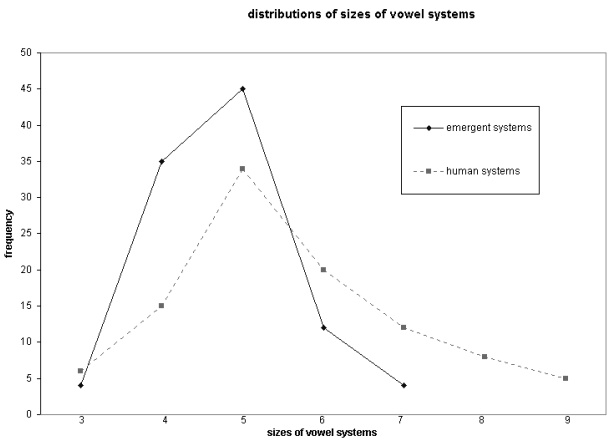

FIG. 10 *Distribution of vowel inventories sizes in emergent and exsiting human vowel systems*

bootstrapping mechanism presented here.

Moreover, it provides a very simple explanation for phonemic coding, which received only poor account in previous research. Yet, understanding truly phonemic coding might be of crucial importance to understand the origins of language: indeed, syntax is thought to be the keystone and one the hottest topics is how compositionality appeared. Interestingly, phonemic coding is a form of primitive compositionality.

Finally, the use of a realistic vowel synthesizer allowed to show that the model also predicts inventories regularities. We showed how the use of a realistic synthesizer is crucial for the prediction of these regularities, but is certainly not the explanation of phonemic coding. As far as phonemic coding and vowel inventories are concerned, the model presented in this paper is more biology-compliant than innatists models, and makes less assumptions than traditional cultural functional models.

6. References

Abbot L., Salinas E. (1994) Vector reconstruction from firing rates, Journal of computational Neuroscience, 1, 89-116.

Anderson J., Silverstein, Ritz, Jons (1977) Distinctive features, categorical perception and probability learning: some applications of a neural model, Psychologycal Review, 84, 413-451.

Archangeli D., Langendoen T. (1997) Optimality theory, an overview, Blackwell Pulishers.

de Boer, B. (1999) Investigating the Emergence of Speech Sounds. In: Dean, T. (ed.) Proceedings of IJCAI 99. Morgan Kauffman, San Francisco. pp. 364-369.

de Boer, B. (2000) The origins of vowel systems, Oxford Linguistics, Oxford University Press.

Boersma, P. (1998) Functional phonology, PhD Thesis, Amsterdam University.

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

Chomsky, N. (1958) Syntactic Structures.

Duda R., Hart P., Stork D. (2001) Pattern Classification, Wiley-interscience.

Watson G.S. (1964) Smooth regression analysis, Sankhya: The Indian Journal of Statistics. Series A, 26 359-372.

Reggia J.A., D'Autrechy C.L., Sutton G.G., Weinrich M. (1992) A competitive distribution theory of neocortical dynamics, Neural Computation, 4, 287-317.

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

R. I. Damper (2000) Ontogenetic versus phylogenetic learning in the emergence of phonetic categories. 3rd International Workshop on the Evolution of Language, Paris, France. p.55-58.

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.

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

Kirby, S. (1998), Syntax without natural selection: how compositionnality emerges from vocabulary in a population of learners, in Hurford, J., Studdert-Kennedy M., Knight C. (eds.), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Kuhl, Williams, Lacerda, Stevens, Lindblom (1992), Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 255, pp. 606-608.

Kuhl (2000) Language, mind and brain : experience alters perception, The New Cognitive Neurosciences, M. Gazzaniga (ed.), The MIT Press.

Ladefoged, P. and I. Maddison (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.

MacNeilage, P.F. (1998) The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-548.

Maddieson, I., Ladefoged P. (1996) The sounds of the world's languages, Oxford Publishers.

Morasso P., Sanguinetti V., Frisone F., Perico L., (1998) Coordinate-free sensorimotor processing: computing with population codes, Neural Networks 11, 1417-1428.

Oudeyer, P-Y. (2001a) Coupled Neural Maps for the Origins of Vowel Systems. Proceedings of ICANN 2001, International Conference on Artificial Neural Networks, Vienna, Austria, LNCS, springer verlag, Lectures Notes in Computer Science, 2001. 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. (2002) Emergent syllable systems : why functional pressure is not sufficient to explain phonemic coding, submitted.

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

Sejnowsky, T. (1977) Storing covariance woth nonlinearly interacting neurons, Journal of mathematical biology, 4:303-312, 1977.

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

Steels, L. (1997a) The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1-35.

Steels L., Oudeyer P-y. (2000) The cultural evolution of phonological constraints in phonology, in Bedau, McCaskill, Packard and Rasmussen (eds.), Proceedings of the 7th International Conference on Artificial Life, pp. 382-391, MIT Press.

Stevens, K.N. (1972) The quantal nature of speech : evidence from articulatory-acoustic data, in David, Denes (eds.), Human Communication : a unified view, pp. 51-66, New-York:McGraw-Hill.

Vennemann, T. (1988), Preference Laws for Syllable Structure, Berlin: Mouton de Gruyter.