From vocal replication to shared combinatorial speech codes: a small step for evolution, a big step for language

Pierre-Yves Oudeyer[1]

Sony CSL Paris py@csl.sony.fr

Summary. In this chapter, we show that from a minimal neural kit for vocal replication, a shared combinatorial speech code with structural regularities and diversity spontaneously self-organizes in a population of agents. This allows to understand that the evolutionary step from vocal replication systems to modern human speech systems might have been rather small.

1 Speech: a structured pre-requisite for language

Humans use spoken vocalizations, or their signed equivalent, as a physical support to carry language. This support is highly organized: vocalizations are built with the re-use of a small number of articulatory units, which are themselves discrete elements carved up by each languistic community in the articulatory continuum. Moreover, the repertoires of these elementary units (the gestures, the phonemes, the morphemes) have a number of structural regularities: for example, while our vocal tract allows physically the production of hundreds of vowels, each language uses most often 5, and never more than 20 of them. Also, certain vowels are very frequent, like /a,e,i,o,u/, and some others are very rare, like /en/. All the speakers of a given linguistic community categorize the speech sounds in the same manner, and share the same repertoire of vocalizations. Speakers of different communities may have very different ways of categorizing sounds (for example, Chinese use tones to distinguish sounds), and repertoires of vocalizations. Such an organized physical support of language is crucial for the existence of language, and thus asking how it may have appeared in the biological and/or cultural history of humans is a fundamental questions. In particular, one can wonder how much the evolution of human speech codes relied on specific evolutionary innovations, and thus how difficult (or not) it was for speech to appear.

One possible answer, proposed by cognitive innatism (Mehler et al., 2000), is that speech does rely deeply on specific biological evolutions, and thus its structure is encoded precisely in the genes. There are two limits to this approach: 1) it does not explicitates what it means to have a speech structure encoded in the genes nor how these genes could have evolved 2) it does not explain why each linguistic community has a different speech code and how one specific speech code is "chosen"

by a community. Another possible answer explains the structure of human speech as the optimal solution to efficient information transfer (in particular, perceptual distinctiveness between vocalizations) given the morpho-physiological properties of the vocal tract and the ear (Stevens, 1972; Lindblom, 1992). This approach also has a number of limits: 1) it does not explain how the optimization might be done in nature or culture; 2) like cognitive innatism, it does not explains why each linguistic community has a different speech code and how one specific speech code is "chosen" by a community.

Another answer, focused on the question of the origins of vowels systems, was proposed by (de Boer, 2001) and does not have these limits. He proposed a mechanism for explaining how a society of agents may come to agree on a vowel system. This mechanism is based on mutual imitations between agents and is called the "imitation game". He built a simulation in which agents were given a model of the vocal tract as well as a model of the ear. Agents played a game called the imitation game. Each of them had a repertoire of prototypes, which were associations between a motor program and its acoustic image. In a round of the game, one agent called the speaker, chose an item of its repertoire, and uttered it to the other agent, called the hearer. Then the hearer would search in its repertoire for the closest prototype to the speaker's sound, and produce it (he imitates). Then the speaker categorizes the utterance of the hearer and checks if the closest prototype in its repertoire is the one he used to produce its initial sound. He then tells the hearer whether it was "good" or "bad". All the items in the repertoires have scores that are used to promote items which lead to successful imitations and prune the other ones. In case of bad imitations, depending on the scores of the item used by the hearer, either this item is modified so as to match better the sound of the speaker, or a new item is created, as close as possible to the sound of the speaker.

This model is very interesting and was one of the first to demonstrate a process of cultural formation of shared vowel systems within a population of agents. This model also allowed to understand how the interaction between learning mechanisms and morpho-physiological constraints could explain both the statistical regularities that we observe in the human vowel systems and their diversity. Finally, de Boer's model was also able to deal with phenomena of sound change, showing how repertoires of vowels could evolve with time. This model was then extended in (Oudeyer, 2001), which showed how similar results could be obtained concerning the formation of shared syllable systems with the prediction of regularities in syllables structures.

Nevertheless, if the "imitation game" is a good framework for studying the evolution of modern speech system, it is less clear to see how it can allow us to understand the evolutionary origins of speech. Indeed, the imitation game implies rather complex cognitive and behavioral capabilities for agents, and assumes implicitely the pre-existence of a linguistic context which is problematic if one wants to understand the origins of language. First of all, agents need to be able to play a game which is a protocol with partly arbitrary rules, involving successive turn-taking and asymmetric changing roles. Second, they need to understand that at a point in the game, one sound produced by the speaker should be imitated, and that this imitation will undergo evaluation from the speaker, which itself needs to understand that the sound produced by the hearer is intended to be an imitation and is not related to something else happening around. Finally, the speaker need to be able to produce a feed-back signal associated with the quality of the imitation, and the hearer has to be able to understand the feedback, i.e. that from the point of view of the other,

he did or did not manage to imitate successfully. Also, in the imitation game, there is a mechanism which explicitely forces the building of a repertoire of sounds which must be distinctive from each other, and there is an explicit pressure to invent new sounds. This clearly models a need to use these sounds in order to name efficiently a growing number of "things" in the environment, i.e. the pre-existence of a linguistic context.

So, because it implies a very complex form of imitation involving both the understanding of the other's intentions and the interpretation of other's vocalization in terms of one's own repertoire of distinctive vocalizations, and because it involves the presence of a linguistic context, the "imitation game" of de Boer is more a model of the origins of particular languages than a model of the origins of the language capacity.

I will now present another model which might bring more light to this latter question. This model also involves a population of agents endowed with a vocal tract, a cochlea and an artificial brain, but what makes it special is that the neural system which is used is very basic and corresponds basically to the minimal kit necessary for analogic vocal replication. Vocal replication refers here to the capacity to reproduce precisely but in an holistic manner an acoustic/articulatory trajectory which is perceived. So, this is basically mimicry applied to the vocal domain (for a technical definition of mimicry, see (Nehaniv and Dautenhahn, 2002)). In general, the capacity for motor mimicry/copying might have appeared in evolution as a very basic form of imitation constituting the first kinds of social learning. For the vocal modality, and as present in a number of birds and whales species, vocal replication has been argued to be useful for the maintainance of social cohesion (Beecher and Brenowitz, 2005). What is interesting is that vocal replication/mimicry does not assume the understanding of intentions, and does not necessitate the existence of a repertoire of disctinctive and discrete vocal units serving as categories to cut a perceived trajectory into high-level segments. Moreover, in the model that I will present, on the one hand there is no explicit pressure for building such a repertoire of distinctive units, and on the other hand agents do not interact in an organized manner (there is no "game" or "protocol" of interaction). Yet, I will show that during the process of babbling and listening to vocalizations produced by nearby agents, a low-level and simple coupling of perception and production for vocal replication can spontaneously self-organize a shared repertoire of discrete combinatorial speech codes with structural regularities and diversity. This allows to show that the minimal neural kit for vocal replication needs very few change (even maybe no change at all) in order to generate a speech code which has the crucial properties of modern speech: in short, the evolutionary step from non-speech to speech may have been rather small.

2 Coupling perception and production in a model of vocal replication

This model is based on the building of an artificial system, composed of agents endowed with working models of the vocal tract, of the cochlea and of some parts of the brain. Before going forward to the specificities of this vocal architecture, we will describe an outline of the minimal neural kit that allows to achieve motor replication or mimicry. As stated above, motor mimicry involves the analogic and holistic

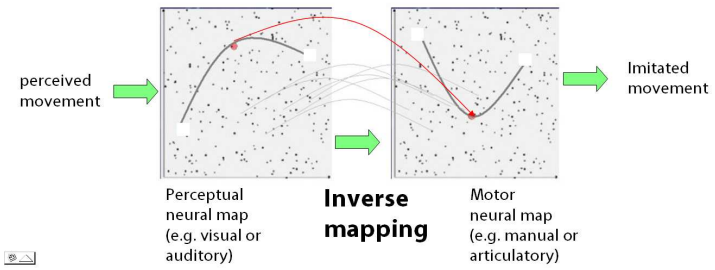

Fig. 1. The minimal neural kit for motor replication/mimicry.

replication by oneself of a movement performed by someone else. As show by the computational litterature (e.g. Morasso et al. (1998)), the most simple system which can do this is basically a neural machinery composed of three parts: one perceptual neural map encoding the movement into a perceptual trajectory, one motor neural map encoding motor trajectories and used to actually control the moving organs, and a set of connections which are typically hebbian synapses and whose purpose is to allow the transformation of the trajectory from one space to the other. Figure 1 presents a summary of this architecture. We can see that within this architecture, no mechanism of categorization is present and from a computational point of view, it amounts to map one continuous trajectory holistically from one space to the other. We will now instanciate this architecture in the context of vocal mimicry: the perceptual space will be acoustic, and the motor space will be articulatory.

Overview.

Each agent has one ear which takes measures of the vocalizations that it perceives, which are then sent to its brain. It also has a vocal tract, whose shape is controllable and is used to produce sounds. Typically, the vocal tract and the ear define three spaces: the motor space (which will be for example 3-dimensional in the vowel simulations with tongue body position, tongue height and lip rounding); the acoustic space (which will be 4-dimensional in the vowel simulation with the first four formants) and the perceptual space (which corresponds to the information the ear sends to the brain, and will be 2-dimensional in the vowel simulations with the first formant and the second effective formant).

The ear and the vocal tract are connected to the brain, which is basically a set of interconnected artificial neurons. This set of artificial neurons is organized into two neural topological maps: one perceptual map and one motor map. Topological neural maps have been widely used for many models of cortical maps (Kohonen, 1982; Morasso et al., 1998), which are the neural devices that humans have to represent parts of the outside world (acoustic, visual, touch etc.). Figure 2 gives an overview of the architecture. We will now describe the technical details of the architecture.

Motor neurons, vocal tract and production of vocalizations.

A motor neuron j is characterized by a preferred vector v_j which determines the vocal tract configuration which is to be reached when it is activated and when the

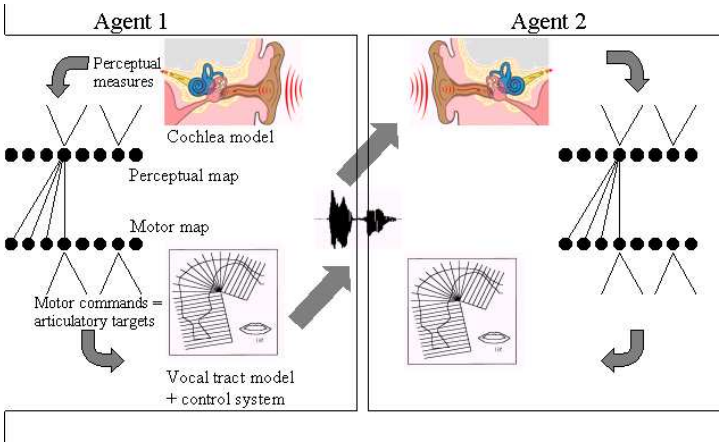

Fig. 2. Architecture of the artificial system : agents are given an artificial ear, an artificial vocal tract, and an artificial "brain" which couples these two organs. Agents are themselves coupled through their common environment : they perceive the vocalizations of their neighbours.

agent sends a GO signal to the motor neural map. This GO signal is sent at random times by the agent to the motor neural map. As a consequence, the agent produces vocalizations at random times, independently of any events.

When an agent produces a vocalization, the neurons which are activated are chosen randomly. Typically, 2, 3 or 4 neurons are chosen and activated in sequence. Each activation of a neuron specifies, through its preferred vector, a vocal tract configuration objective that a sub-system takes care of reaching by moving continuously the articulators. In this chapter, this sub-system is simply a linear interpolator, which produces 10 intermediate configurations between each articulatory objective, which is an approximation of a dynamic continuous vocalization and that we denote $ar_1, ar_2, ..., ar_N$. Figure 3 illustrates this process in the case an abstract 2-dimensional articulatory space.

An artificial vocal tract is used to compute an acoustic image of the dynamic articulations. We have re-implemented the vocal tract model of vowel production designed by (de Boer, 2001). We use vowel production only because there exists this computationally efficient and rather accurate model, but one could do simulations with a vocal tract model which models consonants if efficient ones were available. This model is based on the three major vowel articulatory parameters (Ladefoged and Maddieson, 1996): lip rounding, tongue height and tongue position. The values within these dimensions are between 0 and 1, and a triplet of values $ar_i = (r, h, p)$ defines an articulatory configuration. The acoustic image of one articulatory configuration is a point in the 4-dimensional space defined by the first four formants, which are the frequencies of the peaks in the power spectrum, and is computed with the formula defined in (de Boer, 2001).

The preferred vector of each neuron in the motor map is updated each time the motor neurons are activated (which happens both when the agent produces a vocalization and when it hears a vocalization produced by another agent, as we will

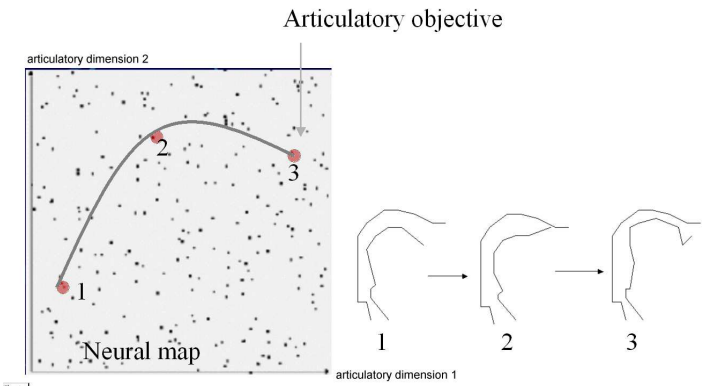

Fig. 3. When an agent produces a vocalization, several motor neurons are activated in sequence. Each of them corresponds to an articulatory configuration which has to be reached from the current configuration. A sub-control system takes care of interpolating between the different configurations.

explain below). This update is made in two steps : 1) one computes which neuron m is most activated and takes the value v_m of its preferred vector ; 2) the preferred vectors of all neurons are modified with the formula:

$$v_{j,t+1} = v_{j,t} + 0.001.G_{j,t}(s).(v_m - v_{j,t})$$

where $G_{j,t}(s)$ is the activation of neuron j at time t with the stimulus s (as we will detail later on) and $v_{j,t}$ denotes the value of v_j at time t. This law of adaptation of the preferred vectors has the consequence that the more a particular neuron is activated, the more the agent will produce articulations which are similar to the one coded by this neuron. This is because geometrically, when v_m is the preferred vector of the most active neuron, the preferred vectors of the neurons which are also highly activated are shifted a little bit towards v_m. The initial value of all the preferred vectors of the motor neurons is random and uniformly distributed. There are in this chapter 500 neurons in the motor neural map (above a certain number of neurons, which is about 150 in all the cases presented in the chapter, nothing changes if this number varies).

Ear, perception of vocalizations and perceptual neurons.

We describe here the perceptual system of the agents, which is used when they perceive a vocalization. As explained in the previous paragraphs, this perceived vocalization takes the form of an acoustic trajectory, i.e. a sequence of points which approximate the continuous sounds. Here, these points are in the 4-D space whose dimensions are the first four formants of the acoustic signal. We then use a model of the cochlea, described in (Boe et al., 1995) and (de Boer, 2001), which transforms this 4-D acoustic representation in a 2-D perceptual representation that we know is close to the way humans represent vowels.

The agent gets as input to its perceptual neural system a trajectory of perceptual points. Each of these perceptual points is then presented in sequence to its perceptual neural map (this models a discretization of the acoustic signal by the ear due to its limited time resolution).

The neurons i in the perceptual map have a gaussian tuning function which allows us to compute the activation of the neurons upon the reception of an input stimulus. If we denote by $G_{i,t}$ the tuning function of neuron i at time t, s is a stimulus vector, then the form of the function is:

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

where the notation $v_1.v_2$ denotes the scalar product between vector v_1 and vector v_2, and $v_{i,t}$ defines the center of the gaussian at time t and is called the preferred vector of the neuron. This means that when a perceptual stimulus is sent to a neuron i, then this neuron will be activated maximally if the stimulus has the same value as $v_{i,t}$. The parameter σ determines the width of the gaussian, and so if it is large the neurons are broadly tuned (a value of 0.05, which is used in all simulations here, means that a neuron responds substantially to 10 percent of the input space).

When a neuron in the perceptual map is activated because of a stimulus, then its preferred vector is changed. The mathematical formula of the new tuning function is:

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

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

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

This formula makes that the distribution of preferred vectors evolves so as to approximate the distribution of sounds which are heard.

The initial value of the preferred vectors of all perceptual neurons follows a random and uniform distribution. There are 500 neurons in the perceptual map in the simulations presented in this chapter.

Connections between the perceptual map and the motor map.

Each neuron i in the perceptual map is connected unidirectionally to all the neurons j in the motor map. The connection between the perceptual neuron i and the motor neuron j is characterized by a weight $w_{i,j}$, which is used to compute the activation of neuron j when a stimulus s has been presented to the perceptual map, with the formula :

$$G_{j,t}(s) = \frac{1}{\sqrt{2\pi}\sigma} * e^{-\sum_i w_{i,j} G_{i,t}(s)/\sigma^2}$$

The weights $w_{i,j}$ are initially set to a small random value, and evolve so as to represent the correlation of activity between neurons. This is how agents will learn the perceptual/articulatory mapping. The learning rule is hebbian (Sejnowsky, 1977):

$$\delta w_{i,j} = c_2(G_i - < G_i >)(G_j - < G_j >)$$

where G_i denotes the activation of neuron i and $< act_i >$ the mean activation of neuron i over a certain time interval (correlation rule). c_2 denotes a small constant. This learning rule applies only when the motor neural map is already activated before the activations of the perceptual map have been propagated, i.e. when an agent hears a vocalization produced by itself. This amounts to learning the perceptual/motor mapping through vocal babbling.

Note that this means that the motor neurons can be activated either through the activation of the perceptual neurons when a vocalization is perceived, or by direct activation when the agent produces a vocalization (in this case, the activation of the chosen neuron is set to 1, and the activation of the other neurons is set to 0). Because the connections are unidirectional, the propagation of activations only takes place from the perceptual to the articulatory map (this does not mean that a propagation in the other direction would change the dynamics of the system, but we did not study this variant).

This coupling between the motor map and the perceptual map has an important dynamical consequence: the agents will tend to produce more vocalizations composed of sounds that they have already heard. Said another way, when a vocalization is perceived by an agent, this increases the probability that the sounds that compose this vocalization will be re-used by the agent in its future vocalizations. It is interesting to note that this phenomenon of phonological attunement is observed in very young babies (Vihman, 1996).

Coupling of agents.

The agents are put in a world where they move randomly. At random times, a randomly chosen agent sends a GO signal and produces a vocalization. The agents which are close to it can perceive this vocalization. Here, we fix the number of agents who can hear the vocalization of another to 1 (we pick the closest one). This is a non-crucial parameter of the simulations, since basically nothing changes when we tune this parameter, except the speed of convergence of the system (and this speed is lowest when the parameter is 1). Technically, this amounts to having a list of agents, and in sequence picking up randomly two of them, have one produce a vocalization, and the other hear it. Typically, there are 20 agents in the system. This is also a non-crucial parameter of the simulation : nothing changes except the speed of convergence.

3 Dynamics

Crystallization.

The present experiment used a population of 20 agents. Initially, as the preferred vectors of neurons are randomly and uniformly distributed across the space, the

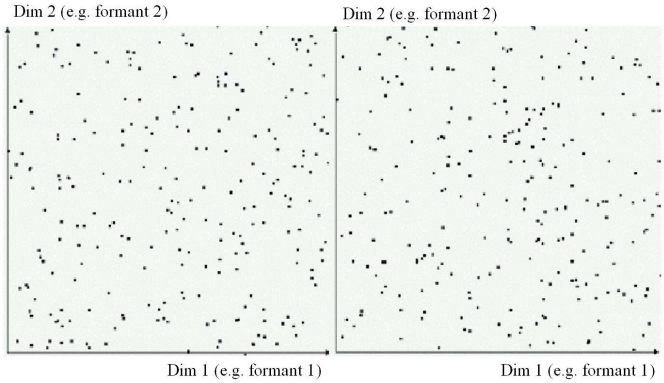

Fig. 4. Perceptual neural maps of two agents at the beginning (the two agents are chosen randomly among a set of 20 agents). Units are arbitrary. Each of both square represents the perceptual map of one agent.

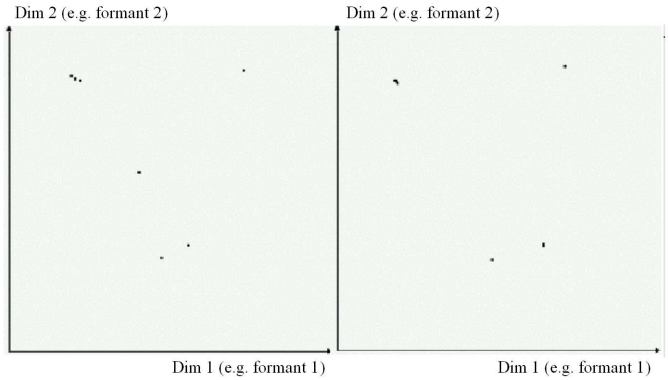

Fig. 5. Neural maps after 2000 interactions, corresponding to the initial state of figure 4 The number of points that one can see is fewer than the number of neurons, since clusters of neurons have the same preferred vectors and this is represented by only one point.

different targets that compose the vocalizations of agents are also randomly and uniformly distributed. Figure 4 shows the preferred vectors of the neurons of the perceptual map of two agents. We see that they cover the whole space uniformly. They are not organized.

The learning rule of the acoustic map is such that it evolves so as to approximate the distribution of sounds in the environment. All agents produce initially complex sounds composed of uniformly distributed targets. Hence, this situation is in equilibrium. Yet, this equilibrium is unstable, and fluctuations ensure that at some point, the symmetry of the distributions of the produced sounds breaks: from time to time, some sounds get produced a little more often than others, and these

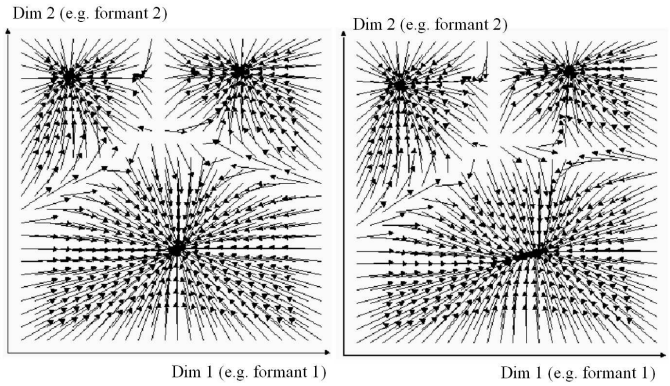

Fig. 6. Representation of the distribution of preferred vectors shown in figure 5. The arrows indicate the direction of density increase. We see that the number of clusters is fewer that the number of points in the last figure. This is because in the previous figure, some points corresponded to clusters and other to single points. This figure allows to see that a combinatorial system based on three key articulatory configurations has been built.

random fluctuations may be amplified through the positive feedback loop implied by the coupling between perception and production on the one hand, and the plasticity rules on the other hand. This leads to a multi-peaked distribution: agents get in a situation like that of Figure 5 which corresponds to Figure 4 after 2000 interactions in a population of 20 agents. Figure 5 shows that the distribution of preferred vectors is no longer uniform but clustered (the same phenomenon happens in the motor maps of the agents, so we represent here only the perceptual maps, as in the rest of the chapter). Yet, it is not so easy to visualize the clusters with the representation in Figure 5, since there are a few neurons which have preferred vectors not belonging to these clusters. They are not statistically significant, but introduce noise into the representation. Furthermore, in the clusters, basically all points have the same value so that they appear as one point. Figure 6 shows better the clusters using a representation of the distribution of preferred vectors: the arrows show the direction of increase of their density. We see that there are now three well-defined attractors or categories, and that they are the same in the two agents represented (they are also the same in the 18 other agents in the simulation). This means that the targets the agents use now belong to one of several well-defined clusters. The continuum of possible targets has been broken, sound production is now discrete. Moreover, the number of clusters that appear is low, which automatically brings it about that targets are systematically re-used to build the complex sounds that agents produce: their vocalizations are now compositional. All the agents share the same speech code in any one simulation. Yet, in each simulation, the exact set of modes at the end is different. The number of modes also varies with exactly the same set of parameters. This is due to the inherent stochasticity of the process.

It is very important to note that this result of crystallization holds for any number of agents (experimentally), and in particular with only one agent which

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

Finally, it has to be noted that a crucial parameter of the simulation is the parameter σ which defines the width of the tuning functions. All the results presented are with a value 0.05. In (Oudeyer, 2006), we present a study of what happens when we tune this parameter. This study shows that the simulation is quite robust to this parameter: indeed, there is a large zone of values in which we get a practical convergence of the system in a state where agents have a multi-peaked preferred vector distribution, as in the examples we presented. What changes is the mean number of these peaks in the distributions: for example, with $\sigma = 0.05$, we obtain between 3 and 10 clusters, and with $\sigma = 0.01$, we obtain between 6 and 15 clusters. If σ becomes too small, then the initial equilibrium of the system becomes stable and nothing changes: agents keep producing inarticulate and holistic vocalizations. If σ is too large, then the practical convergence of the system is the same as the mathematical convergence: only one cluster appears.

Structure

In the last paragraph, we showed that a system of combinatorial vocalizations self-organized, shared by the agents in the same simulation and different in agents of different simulations. We will now study the structure of these self-organized repertoires by focusing on the vowels that compose the complex dynamical vocalizations, and compare it to the structure of human vowel system.

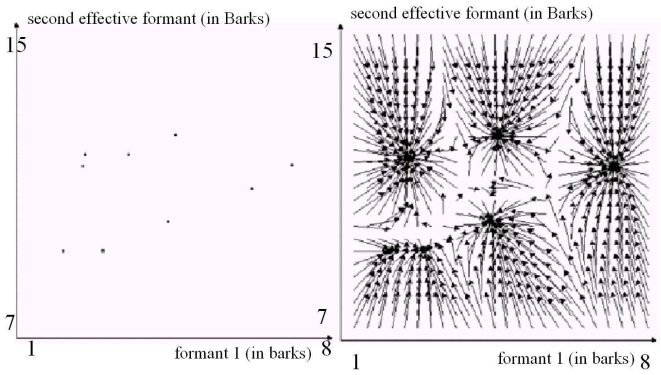

Fig. 7. Neural map and attractor field of the same agents after 2000 interactions with other 20 agents. The corresponding figures of other agents are nearly identical. The produced vowel system is here an instantiation the most frequent vowel system in human languages: /a, e, i, o, u/.

A series of 500 simulations was run with the same set of parameters, and each time the number of vowels as well as the structure of the system was checked. Each vowel system was classified according to the relative position of the vowels, as opposed to looking at the precise location of each of them. This is inspired by the work of Crothers (Crothers, 1978) on universals in vowel systems, and is identical to the type of classification performed in (de Boer, 2001). The first result shows that the distribution of vowel inventory sizes is very similar to that of human vowel systems (Ladefoged and Maddieson, 1996): experiments with the artificial system showed that there is a peak at 5 vowels, which is remarkable since 5 is neither the maximum nor the minimum number of vowels found in human languages. The prediction made by the model is even more accurate than the one provided by de Boer (de Boer, 2001) since his model predicted a peak at 4 vowels. Then the structure of the emergent vowel systems was compared to the structure of vowel systems in human languages as reported in (Schwartz et al., 1997). More precisely, the distributions of structures in the 500 emergent systems were compared to the distribution of structures in the 451 languages of the UPSID database (Maddieson, 1984). The results are shown in Figure 8. We see that the predictions are rather accurate, especially in the prediction of the most frequent system for each size of vowel system (less than 8). Figure 7 shows an instance of the most frequent system in both emergent and human vowel systems. In spite of the predictions of one 4-vowel system and one 5-vowel system which appear frequently (9.1 and 6 percent of systems) in the simulations and never appear in UPSID languages, these results compare favourably to those obtained in (de Boer, 2001). Yet, like de Boer, we are not able to predict systems with many vowels (which are admittedly rare in human languages, but do exist). This is not very surprising since this model was designed to study the mechanisms which might have allowed the bootstrapping of speech, but not how these primitive speech systems might have been recruited later on for complex linguistic communication and thus undergo severe functional pressures for larger repertoires.

4 Conclusion

In this paper, we have shown that that from a minimal neural kit for vocal replication, a shared combinatorial speech code with structural regularities and diversity could spontaneously self-organize in a population of agents. This result is conditionned by the value of the width of the tuning functions of neurons, which must be within a certain interval (but this interval is rather large). One needs to state that the capability of vocal replication is not diminished if this value gets out of this interval within certain limits. Yet, it is easy to see that from an evolutionary point of view, the transition from inarticulated speech to combinatorial and shared speech codes can be achieved just by tuning the width of these functions, which is admittedly a small modification. As a consequence, this allows to understand that, in a scenario in which our ancestors passed through a stage where motor replication, and in particular vocal replication, was present and language still absent, the evolutionary step from vocal replication systems to modern human speech systems might have been rather small.

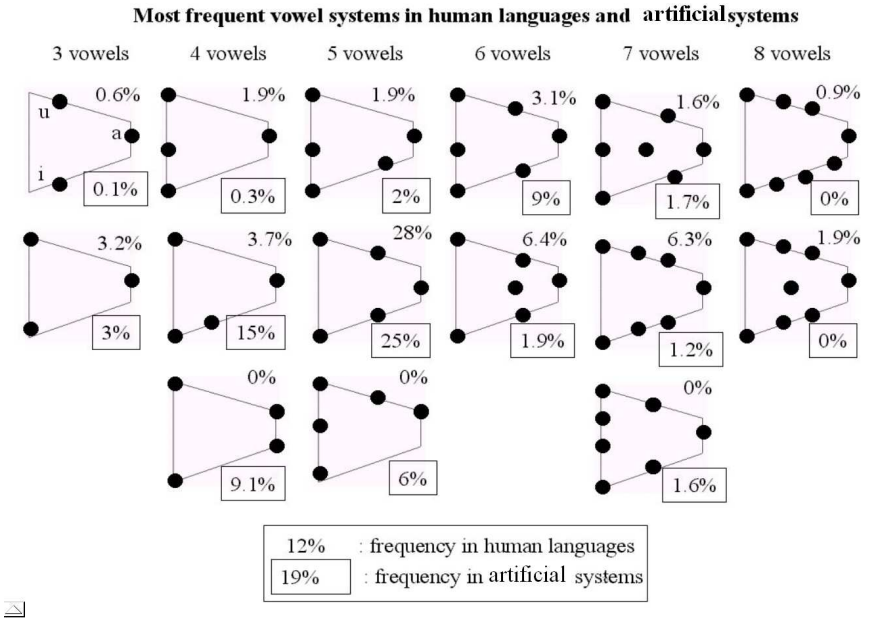

Fig. 8. Distribution of vowel inventories structures in artificial and UPSID human vowel systems. This diagram uses the same notations than the one in (Schwartz et al., 1997). Note that here, the vertical axis is also F2, but oriented downwards.

References

Beecher, M., Brenowitz, E., 2005. Functional aspects of song learning in songbirds. Trends in Ecology and Evolution 20, 143–149.

Boe, L., Schwartz, J., Valle, N., 1995. The prediction of vowel systems: perceptual contrast and stability. In: E., K. (Ed.), Fundamentals of Speech Synthesis and Recognition. Chichester:John Wiley, pp. 185–213.

Crothers, J., 1978. Typology and universals of vowels systems. Phonology 2, 93–152.

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

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43 (1), 59–69.

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

Lindblom, B., 1992. Phonological units as adaptive emergents of lexical development. In: Ferguson, Menn, Stoel-Gammon (Eds.), Phonological Development: Models, Research, Implications. York Press, Timonnium, MD, pp. 565–604.

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

Mehler, J., Christophe, A., Ramus, F., 2000. What we know about the initial state for language. In: Marantz, A., Miyashita, Y., O'Neil, W. (Eds.), Image, Language,

Brain: Papers from the first Mind-Brain Articulation Project symposium. Cambridge, MA: MIT Press, pp. 51–75.

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

Nehaniv, C. L., Dautenhahn, K., 2002. The correspondence problem. In: Nehaniv, C. L., Dautenhahn, K. (Eds.), Imitation in Animals and Artifacts. MIT Press, pp. 41–61.

Oudeyer, P.-Y., 2001. Origins and learnability of syllable systems, a cultural evolutionary model. In: Collet, P., Fonlupt, C., Hao, J., Lutton, E., Schonenauer, M. (Eds.), in Artificial Evolution. LNCS 2310. Springer Verlag, pp. 143–155.

Oudeyer, P.-Y., 2006. Self-Organization in the Evolution of Speech. Studies in the Evolution of Language. Oxford University Press.

Schwartz, J., Bo, L., Valle, N., Abry, C., 1997. Major trends in vowel systems inventories. Journal of Phonetics 25, 255–286.

Sejnowsky, T., 1977. Storing covariance with non-linearly interacting neurons. Journal of mathematical biology 4, 303–312.

Stevens, K., 1972. The quantal nature of speech: evidence from articulatory-acoustic data. New-York: Mc Graw-Hill, pp. 51–66.

Vihman, M., 1996. Phonological Development: The Origins of Language in the Child. Oxford, UK: Blackwell Publishers.