The social formation of acoustic codes with "something simpler"

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

Abstract

How do humans (or other animals) acquire those cultural acoustic codes which are finite discrete repertoires of vocalizations as well as categorization systems (e.g. vowel systems in humans)? How do these acoustic codes, shared by each speakers of a given language and possibly very different from one language to the other, appeared? It has been proposed in the litterature (e.g. de Boer, 2000) that some form of non-trivial imitation was the mechanism which gave a solution to both questions. We show in this paper that a much simpler mechanism is able to account for the same phenomena. It is based on the self-organization of the coupling between perception and production both within and across agents. The assumptions on which the mechanism relies only deal with local properties of neural units as well as the ability to learn a mapping between two modalities in an unsupervised manner. No social skills or functional pressures related to communication are required. Yet, a structured discrete acoustic code shared by the society appears.

1 Introduction

Humans as well as other animals like some species of birds or whales, use acquired acoustic codes. This means that they share repertoires of sounds that they can produce and categorize. For example, humans speaking a given language produce the same set of vowels (e.g. [a], [e], [i], [o], [u]), and categorize the vowel space in the same manner. This is a cultural code because the way the vowel space is carved into categories and prototypes is arbitrary and particular to each language community (there are of course statistical regularities in human languages, but the number of existing, and so possible, vowel systems is very large). Note that there are two aspects to these acoustic codes: they are discretization of the continuous acoustic space into distinct discrete categories/modes, and this discretization is shared by all agents in the society.

How does an individual acquire an acoustic code? How do cultural discrete acoustic codes shared by populations of individuals appeared? (de Boer, 2000) proposed an answer in the case of the modelisation of the origins of vowel systems: this is the same mechanism which explains the acquisition of an acoustic code and its formation; this mechanism is imitation. 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 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". Each item in the repertoires have scores which 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 better match the sound of the speaker, or a new item is created, as close as possible to the sound of the speaker.

By the description of the game, it is clear that to perform this kind of imitation game, a lof of computational/cognitive power is needed. First of all, agents need to be able to play a game, involving successive turn-taking and assymetric changing roles. Second, they need to be able to voluntarily try to copy the sound production of others, and be able to evaluate this copy. Finally, when they are speakers, they need to recognize that they are being imitated intentionnally, and give feed-back/re-inforcement to the hearer about the success or not. The hearer then 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. As a consequence, it seems not very controversial that agents need to be able to perform some form of non-trivial imitation to play the "imitation game".

We propose in this paper that "something simpler" than imitation (Noble and Todd, 2002) might explain the social formation of sound codes. As a point of convergence with de Boer, the mechanism we present is the same for the acquisition and the formation of acoustic codes. Yet, it requires much less cognitive resources. It has similarities with what has sometimes been called "response facilitation" (Byrne and Russon, 1998): "the observer is co-

pying a motor act that is already in its repertoire, and, as a result of copying, the frequency of the particular behavioural act increases" (Miklosi, 1999). The similarity is that the observation of a behavioural act (vocalization here) increases the frequency of the production of a similar vocalization in the future, but the dissymilarity is that here agents do not copy what they hear (the increase in frequency is a distributed statistical effect). Also, all vocalizations (in a continuous space) are potentially in the repertoires of all agents initially, but this does not mean that they produce all of them.

The mechanism relies on a coupling of perception and production within and across agents. It is a generalization and abstraction of the lower-level mechanism presented in (Oudeyer, 2002a), which used two neural maps, one acoustic/percepual map and one motor map. The mechanism can be summarized by 5 assumptions that we will describe, defining local properties of neural units, and will be shown to be sufficient to generate macro-properties qualitatively different, i.e. emergent, at the society level [1].

2 The method of the artificial

The mechanism we present here is not intended to be a model of reality. Its assumptions are not based on existent knowledge in neuroscience or ethology or speech research (though they are inspired by it). The goal is to show a kind of mechanism that may lead to the kind of acoustic codes we observe in nature. It proceeds by abductive reasoning (Peirce, 1958), and its main use is to be a cautionary tale which helps to avoid uncertain intuitive reasoning which are very current in verbal theories (Steels, 2001). The claim of this paper is that complex digital acoustic codes shared by a population of agents can be formed without complex social and imitative capabilities, and without a pressure to develop a system for communication. Yet, we do not claim that the artificial system presented in this paper describes a mechanism that actually happens in nature.

3 The artificial system

This paper will remain theory neutral in terms of the variables that describe the perception and production of sounds. The perception of acoustic signals will be coded as points in abstract continuous spaces of dimension N, and the production will be coded as points in other abstract continuous spaces of dimension D. To simplify, we consider in this paper that vocalization are static configurations, and not trajectories in time. The extension to complex sounds will be describred in a further paper. For example, these points for perception could correspond to the formants, i.e. the frequencies of the peaks in the power spectrum, but also to points in the cochlea membrane activation space, in the MFCC space. For production, abstract points might be configurations in the articulator spaces, muscular spaces, proprioceptive spaces.

1. The concept of "self-organisation" and "emergence" that we use in this paper characterizes systems whose global properties are qualitatively different from its local properties; we do not include any notion of "surprise", which we believe is very subjective: the results we present are not claimed to be surprising

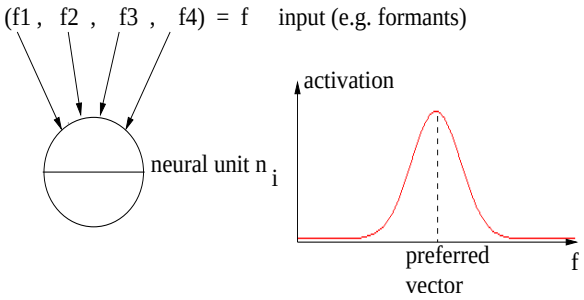

FIG. 1 *One neural unit and its tuning function, the input is an N-dimensional space (e.g. 4 first formants); for ease of representation, here we show the projection of the projection of the tuning function on one dimension.*

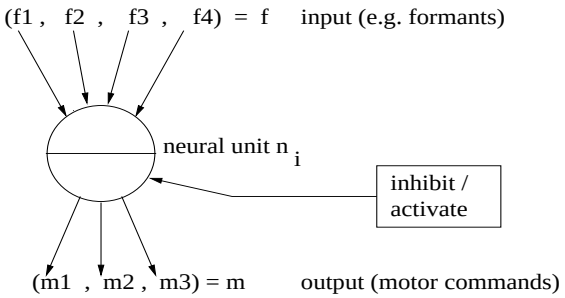

FIG. 2 *The activation of a neural unit by the "inhibit/activate" module allows to retrieve the motor commands that produce the sound to which the unit reacts maximally.*

The mechanism relies on 5 assumptions that we are now presenting.

3.1 Assumption 1: Neural units

We suppose that there are neural units N_i which have broadly tuned gaussian-like receptive fields. What we call "neural unit" could be in the brain one neuron as well as a complex neural network. The receptive field of a neural unit is the function which maps inputs to activation of the unit. Gaussian-like tuning function make that there is an input for which the unit responds maximally, which we call its preferred vector, and then when inputs get further from this preferred vector, the activation decreases along a gaussian function. When a receptive field is broadly tuned, it implies that it is not very specific, i.e. there are many inputs for which it reacts substantially (the gaussian has a large variance). Figure 1 describes one neural unit. If we note $l_{i,t}$ the tuning function of N_i at time t, f one input vector, $f_{p,i}$ the preferred vector of N_i, then:

$$l_{i,t}(f) = \frac{1}{\sqrt{2\pi}\sigma} * e^{-\frac{1}{2}\|f_{p,i}-f\|^2/\sigma^2}$$

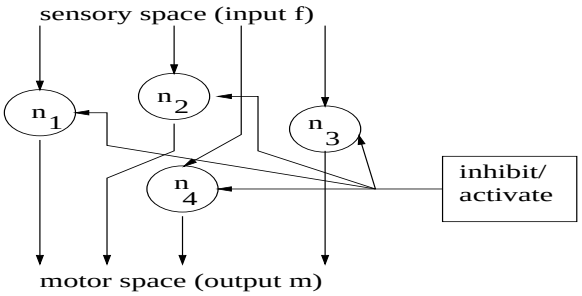

FIG. 3 *There are many neural units, no architecture is needed.*

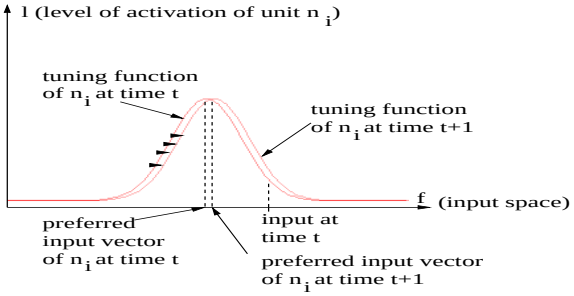

FIG. 4 *When an input activates a neural unit, the preferred vector of this unit is modified so that the unit will be more responsive to this input in the future; this modification is ponderated by the current level of activation of the unit.*

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, as used below, means that a neuron responds substantially to 10 percent of the input space).

3.2 Assumption 2: Motor commands retrieval

We assume that upon the activation of one neural unit N_i by a control system noted "inhibit/activate" on the figures, the brain is able to retrieve a set of motor commands that produce a sound corresponding to the preferred vector of N_i. There are possibly several motor commands for one sounds, we suppose that the brain can retrieve at least one. This kind of inverse-problem is known to be very difficult in general, but is possibly simpler for speech since the sounds that one agent produces are not "reversed" for other agents hearing it. "Reverse" means a symmetry like when someone moves its left hand, others in front of him see the hand at their right. Thus, several papers in the litterature have already showed that reasonable neural architectures could learn this mapping, most often with the use of mirror neurons. Note that these mirror neurons can be the result of learning, as shown in (Oudeyer, 2002a; Bailly et al., 1997). Also, this assumption does not imply that an agent is able to produce the motor commands for any sound it hears, but just for the one covered by their receptive fields centers. Figures 2 and 3 summarizes this assumption.

3.3 Assumption 3: Plasticity

The receptive fields of neural units adapt to the input. What changes is their preferred vector, their width does not evolve (they remain broad). 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 ponderated by their current activation). Basically, adaptation is an increase in sensitivity to stimuli in the environment. Figures 4 explains the process. The formula is:

$$l_{i,t+1}(f) = c_1 * e^{\|M(f_{p,i,t},l_{i,t}(f),f)\|/c_2}$$

where f is the input vector corresponding to the current segment, and M is:

$$M(v,a,s) = v + a * (s - v)$$

From a geometrical point of view, the preferred vector of each neural unit is shifted towards the input vector, and the shift is higher for unit which respond a lot than for unit which do not respond very much. [2]

3.4 Assumption 4: Production

The production of a sound is achieved through the activation of a random N_i by the control system noted "inhibit/activate" on the figures. Activating one neural unit at a given time makes that the motor variables (e.g. the articulators) take the value of the motor vector corresponding to the preferred vector of N_i. For example, if perception is coded as formant trajectories and production as articulator movements, then producing a sound amounts to choosing a formants target, which correspond to an articulatory target, and let the control system adjust the articulators. The crucial point of this assumption is that neural units N_i are both used in the perception process and in the production process. As a consequence, the distribution of targets which are used for production is the same than the distribution of receptive fields, 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. It is important to see that this is not realized through imitation, but is a side effect of an increase of sensitivity of neurons, which is a very generic local low-level neural mechanism. Agents do not imitate each other in this artificial system (but they will develop neural networks that give them the "knowledge" of how to imitate).

3.5 Assumption 5: Initial distribution

The preferred vectors of all the neural units are random along a uniform distribution in the basic form of the sys-

2. The neural network that we use is technically very similar to Self-Organizing Feature Maps. In our case, the input space is of the same dimensionality than the output space, so we do not use it to make dimensionality reduction. Feature maps are normally used to extract some regularities in high dimensional input data. Here, there is no regularity in the input data initially, which is generated also by other neural networks of the same kind. Regularities are rather created through self-organization as explained in the "dynamics" section.

tem. This means that initially they produce sounds that are randomly distributed across the space.

3.6 Non-assumptions

Among the things we do not assume is the fact that agents do not play any language game in the sense used in the literature (Hurford et al., 1998). In fact, they need not have any social skill at all. They are just in a world in which they wander around and sometimes produce sounds and adapt to the sounds they hear around them. They do not have any notion of otherness, and in particular do not imitate each other.

4 The dynamics

Now we describe what happens when a population of agents which have biological properties corresponding to the assumptions cited above. For easier visualization, the input space will be here 1-dimensional. Results extend without any change to higher dimensions. In this part, $\sigma = 0.05$ and there are 150 neural units.

Figure 5 shows the distributions of preferred vectors of 2 agents at the beginning of one simulation. We see that they are approximately uniformly distributed. As the adaptive law of neural units makes that agents tend to produce the same distribution of sounds as the one they hear around them, and as initially they all produce roughly the same uniform distribution, the initial situation is an equilibrium. Now, because there is stochasticity in the mechanism, there will be fluctuations. We now show that this equilibrium is not stable: the fluctuations drive the system in a very different state. Figure 6 shows the distribution of preferred vectors of the same two agents 2000 time steps later. We see that now they are clustered, and that these clusters are the same in the two agents. Their new distribution of preferred vectors is multi-modal. This means that the targets they use to produce sounds are now from one of several well defined modes. This corresponds to the appearance of the discretization of the space of sounds, i.e. some sort of digitalness since they produce sounds belonging to a finite discrete set (modulo the influence of noise) whereas initially they produced sounds belonging spanning a continuous space. Moreover, all the agents share this speech code. 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.

The evolution does stabilize at some point. To monitor the evolution along with time, the mean entropy of distributions was computed at each time step as well as the mean distance between agent's distributions (this was done using the KL-distance). Figures 7 and 8 shows the evolution of these two measures. On the one hand the entropy first decreases and then stabilizes, which shows the crystallization; on the other hand the distance between

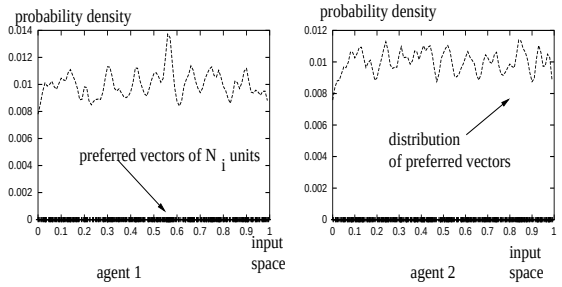

FIG. 5 *Initial distribution of preferred vectors with a one-dimensional input space; we show here the distributions of two agents, who produce sounds spread along the complete continuum of the space (there is no code).*

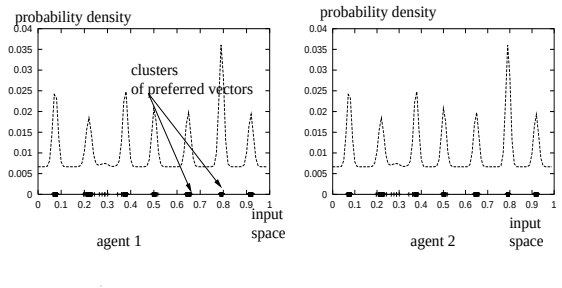

FIG. 6 *The distribution of neural units of the same two agents than in previous figure, after several thousands time steps: they are multi-modal, which means that the sounds that they produce now belong to one of several modes, and moreover these modes are the same for the two agents (the space is discretized in the same manner by all agents: this is a shared code)*

distributions does not increase (initially, they already have similar distributions since they are all uniform !), and even decreases, which shows that the peaks which appear are the same for all agents.

The reason why there is crystallization is that the natural stochasticity of the mechanism makes that initially, from times to times, some sounds get produced a little bit more by the population of agents. This can create deviations which are amplified by the adaptive mechanism through positive feedback.

Finally, if there is only one agent that adapts to its own vocalizations, then it will crystallize also on a multi-modal distribution of target. This means that there are two separable results: digitalness is explained by the coupling between production and perception through the N_i, and can be obtained with only one agent; but putting agents together makes that they synchronize their repertoires of modes. If they were to adapt to their own vocalizations without interactions, then they would end up with uncorrelated repertoires of modes.

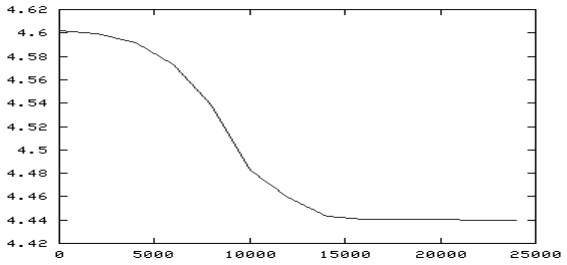

FIG. 7 *To monitor the evolution in time of the distributions of preferred vectors, their mean entropy was computed; in this simulation, we see that it decreases, which correspond to the formation of modes, and then stabilizes, which correspond to a state of convergence (with multiple modes). This curve is for a simulation with 10 agents.*

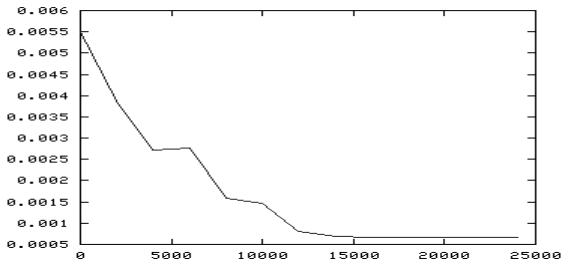

FIG. 8 *Mean distance between the distributions of preferred vectors of the agents; in this simulation, we see that they remain the same, which means that the modes are identical in each agents (this is always the case). This curve is for a simulation with 10 agents*

5 Playing with the parameters

The number of agents was changed from 2 to 50, which does not bring qualitative differences, only the convergence time is modified. If the number of neurons is too low, i.e. less than 50, then if there are too many agents, the simulation sometimes does not crystallize with shared repertoires. But if the number of neurons is increased, then nothing is modified (1000 neurons gives the same result as 150 neurons for example).

A parameter which is more crucial to the outcome of the simulation is the σ which determines the width of the gaussian which defines the tuning function. If this width is very large (¿ 0.25), then the simulation ends up with one cluster, generally in the middle of the space, for all the agents. On the contrary, if the width is too small (¡ 0.005), then the initial uniform distribution is a stable equilibrium: it stays uniform and no symmetry breaking appears. As seen in those values, the range of "interesting" width for the gaussian is large. Also, if one makes the width of the tuning functions stochastic, varying randomly around a mean value plus or minus 10 percent, the simulation also crystallizes on shared digital modes.

Another variation of the simulation deals with the initial distribution of preferred vectors. Above, they were uniformly distributed. What does happen if they have biases? We have investigated this using the constraints of the mapping from articulations to acoustics of the human vowel production system. This means that we have used the model of de Boer describing how values of lip height, lip rounding and position of the tongue map to two effective formant values (de Boer, 2000). To generate the initial distribution of preferred vectors, a uniform exploration of the articulatory space is performed, and for each articulatory configurations, the acoustic image is computed. The values of these images form the set of preferred vectors of the neural units. Hence, our initial distribution is the image of a uniform distribution in the articulatory space mapped into the acoustic space. Figure 9 shows an example of a initial distribution. Then we have studied what kind of vowel systems were generated. First of all, like in the above simulations, each run leads the society of agents to a shared digital code. Now, if we look at which vowel systems they build, we discover that they prefer 5 vowel systems, but they also generate vowel systems of different sizes (see Figure 10). This is similar to human vowel systems. If we look even closer, we discover that the frequency of each vowel system type is quite similar to the human vowel system distribution, as described in the UPSID database of 451 languages (Maddison, 1984). Figure 11 gives an example of a vowel system generated by the simulation, which correspond to the most frequent vowel system in human languages (/i/, /u/, /e/, /o/ and /a/). Figure 12 gives the distribution of all systems generated by the system, and their frequency in human languages. We see that not only do we get shared digital acoustic codes, but also with the use of the adequate bias, we are able to predict both the regularities and diversity of human vowel systems.

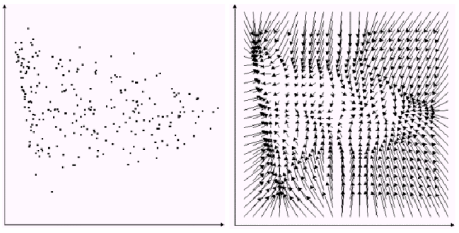

FIG. 9 *Example of initial distribution of preferred vectors when we use the model of vowel articulator. The right square shows their density (it increases in the direction of arrows)*

6 Conclusion

We showed that the formation of shared digital acoustic codes appear in societies of agents with only a very primitive form of social learning, very different from any

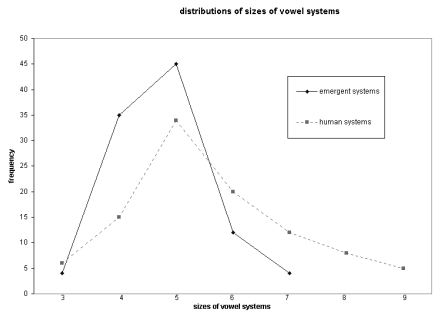

FIG. 10 *Distribution of vowel inventories sizes in emergent and UPSID human vowel systems*

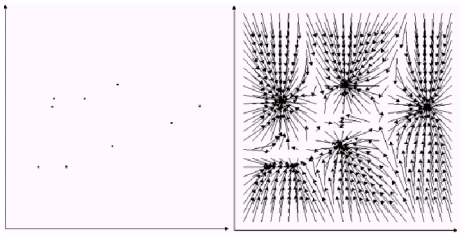

FIG. 11 *The distribution of preferred vectors of the same agent than on previous figure several thousands interactions later. Other agents have of course the same distribution in this simulation. Five vowels appear. This corresponds to the most frequent 5 vowel system in human languages.*

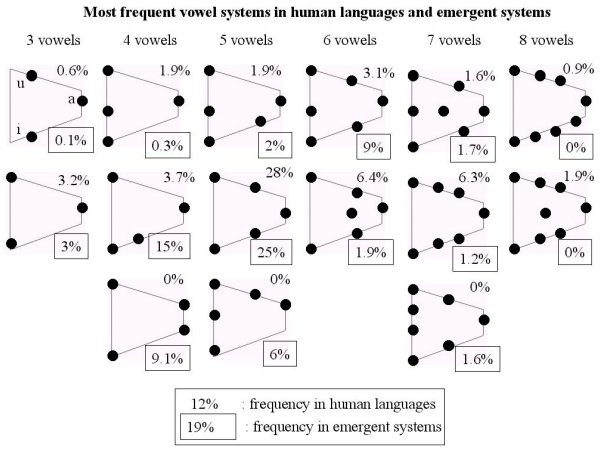

FIG. 12 *Distribution of vowel inventories structures in emergent and UPSID human vowel systems*

non-trivial form of imitation. This contrasts with previous work which pre-supposed that imitation was a pre-requisite to build similar kind of complex socially shared systems. In fact, here there are no social pre-requisites at all. Agents need not be aware of others, they are just adapting locally their neural maps in an unsupervised manner to the sounds in their environment. Through the coupling between perception and production both within and across agents, a process of self-organization with positive feedback loops takes place and global order appears. This is an example of macro-structure which appears spontaneously and independantly from any function, from the local interactions of micro-structures with very different qualitative properties. This happens often in nature, for example with the formation of snow flakes, which are macro-structures with symmetrical recursive design, from the interactions of water molecules, which are micro-structure with assymetrical non-recursive design. We believe this paper provides an original example of how some necessary building blocks to communication might bootstrap from scratch.

7 References

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

Bailly, G., Laboissière, R., and Galván, A. (1997) Learning to speak: Speech production and sensori-motor representations. In Morasso, P. and Sanguineti, V., editors, Self-Organization, Computational Maps and Motor Control, pages 593–615. Elsevier, Amsterdam.

Byrne R.W., Russon A.E. (1998) Learning by Imitation: a Hierarchical Approach, The Behavioural and Brain Sciences 21, 667-721.

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

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

Miklosi A. (1999) The Ethological Analysis of Imitation, Biological Review, 74, pp. 347-374.

Oudeyer, P-Y. (2001a), Coupled Neural Maps for the Origins of Vowel Systems. in the Proceedings of ICANN 2001, International Conference on Artificial Neural Networks, pp. 1171-1176, LNCS 2130, eds. G. Dorffner, H. Bischof, K. Hornik, Springer Verlag,

Oudeyer P-Y (2001b), The Origins Of Syllable Systems : an Operational Model. in the Proceedings of the 23rd Annual Conference of the Cognitive Science society, COGSCI'2001, pp. 744-749, eds. J. Moore, K. Stenning, Laurence Erlbaum Associates

Oudeyer, P-Y. (2002a) Phonemic coding might be a result of sensory-motor coupling dynamics, in the Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior, pp. 406-416, eds. B. Hallam, D. Floreano, J. Hallam, G. Hayes, J-A. Meyer, MIT Press.

Oudeyer P-Y. (2002b) A Unified Model for the Origins of Phonemically Coded Syllables Systems, in the Procee-

dings of the 24th Annual Conference of the Cognitive Science Society, Laurence Erlbaum Associates.

Peirce, Charles Sanders, Collected Papers. (CP). Band I-VI. (Hrsg.) Charles Hartshorne und Paul Weiß. Harvard University Press 1931-1935. Band VII, VIII. (Hrsg.) Arthur W. Burks. 1958.

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

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.

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