**Title**
The Origins of Syllable Systems: An Operational Model

**Permalink**
https://escholarship.org/uc/item/2x99m0nb

**Journal**
Proceedings of the Annual Meeting of the Cognitive Science Society, 23(23)

**ISSN**
1069-7977

**Author**
Oudeyer, Pierre-yves

**Publication Date**
2001

Peer reviewed

# The Origins Of Syllable Systems : An Operational Model

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

## Abstract

Many models, computational or not, exist that describe the acquisition of speech: they all rely on the pre-existence of some sort of linguistic structure in the input, i.e. speech itself. Very few address the question of how this coherence and structure appeared. We try here to give a solution concerning syllable systems. We propose an operational model that shows how a society of robotic agents, endowed with a set of non-linguistically specific motor, perceptual, cognitive and social constraints (some of them are obstacles whereas others are opportunities), can collectively build a coherent and structured syllable system from scratch. As opposed to many existing abstract models of the origins of language, as few short-cuts as possible were taken in the way the constraints are implemented. The structural properties of the produced sound systems are extensively studied under the light of phonetics and phonology and more broadly language theory. The model brings more plausibility in favor of theories of language that defend the idea that there needs no innate linguistic specific abilities to explain observed regularities in world languages.

## Introduction

There are many studies about the acquisition of speech sounds, and of language in general: a lot of data is available and a lot of theories as well as operational models have been developed (Altman, 1995). Although there is great disagreement for these questions, one assumption is logically done by these models: a pre-existing language already exists, with all its associated structure and redundancies. Depending on the theoretical position, either the acquisition of a particular language consists in adjusting a number of parameters of an innate language acquisition device that already knows most of the structure of languages (Chomsky and Halle, 1968), or it relies on statistical learning techniques able to infer regularities from the data. On the contrary, very little is known about how this structure and these regularities originated, from a situation where there was no sound system at all (and no language in general) ?. In brief, how did speech emerge and why does it have the shape it has ? This ignorance is due partly because the question of the origins of language has been an actively researched question only for slightly more than a decade (Hurford et al. 1998), partly because no meaningful data of sound systems of the first speaking humans exists (by nature, speech leaves no physical trace in its environment), and partly because the questions are simply very difficult.

Because the mechanisms involved are bound to be complex and involve the interaction of many environmental, physical, neuro-cognitive and genetic entities, and because data are scarce, computational models have been increasingly used in the past 10 years in the field (Hurford and al., 1998). Indeed, their nature allows on the one hand to test the operational plausibility and feasibility of otherwise highly speculative theories, and on the other hand to gain new insights about how certain aspects can be explained by the intricate dynamics of the complex systems involved.

The research presented here concerns a computational model of the origins of syllable systems, which are thought to be a fundamental unit of the complex sound system of nowadays human languages. It aims at being a plausible implementation, and hence a proof of feasibility, of the theory that claims that sound systems originated and have properties explained by the self-organization of motor, perceptual, cognitive, social and functional constraints that are not linguistically specific, and this in a cultural manner (Steels, 1998). In brief, this theory states that speech is a complex cultural adaptive system. Among the forces at stake are articulatory ease, perceptual distinctiveness, time and memory limitations, lexicon pressure, efficiency of communication, noise in the environment and conformance to the group. The word constraint is used in its most general meaning: it can be obstacle or opportunity as we will see.

A number of computational models concerning the origins of sound systems have already been developed, mainly for phonemes, and more precisely at the vowel level. Two of them are representative: the first one was developed by Lindblom (1992), and consisted in showing that the numerical optimization of a number of motor and perceptual constraints defined analytically allowed to predict the most frequent vowel systems in the human languages (in particular the high occurrence of the 5 vowel system /e i a u o/). Whereas it gave an idea of why vowel systems have the properties observed by phoneticians, it did not give any idea of what process could have achieved this optimization. Indeed, it is not plausible that primitive humans may have willingly computed all possible vowel systems and took an optimal one (and still if this was the case, this does not tell us how they agreed on which of the many solutions that exist). This short-coming was corrected by the model of de Boer (1999)

, who places itself in a broader class of models consisting in setting up a society of agents endowed with realistic capabilities and physical constraints (whose action on the sound system is not explicit) and have them interact in order to build culturally an efficient communication system (Steels, 1998; Steels and Oudeyer 2000). More specifically, in his model no explicit optimization was performed, but rather near-optimal systems were obtained only as a side effect of adaptation to the task of building a communication system. Coherence did not come from a genetically pre-specified plan, but from the self-organization arising from positive feedback loops.

Much fewer existing models tackle the question of the origins of complex sounds, in particular syllables. Lindblom (1992) and then Redford (1998) have developed models resembling the Lindblom model for vowels: they consist in defining explicitly with analytical formulas a number of constraints, then running an optimization algorithm and showing that near-optimal systems have regularities characteristic of the most common syllable systems in the human languages. An example of regularity is the sonority hierarchy principle, which states that the sonority of syllables tends to increase until their nucleus, and then decrease. The present model aims at applying the multi-agent based modeling paradigm mentioned earlier to the question of the origins and properties of syllable systems: like de Boer's model, it should not only try to explain why syllables tend to be the way they are, but also what actual process built them. An additional requirement needed by this model is the fact that agents should be as realistic as possible, and should operate in the real world. One of the reasons for the need of realism is that previous models have shown that constraints are important to the shape of sound systems: when dealing with too abstract constraints, there is a danger to find wrong explanations. Furthermore, Redford showed that certain phenomena can be understood only by considering the interactions between constraints, so models should try to incorporate most of them.

The next section presents an overview of the model with its different modules. A more comprehensive description can be found in a companion paper (Oudeyer 2001) dedicated to the technical details of the setup. Then we present results about the behavior of the system, and discuss implications for phonetics and phonology, and more generally language epigenesis.

## The model

### The imitation game

Central to the model is the way agents interact. We use here the concept of game, recently operationalized in a number of computational models of the origins of language (Steels, 1998). A game is a sort of protocol that describes the outline of a conversation, allowing agents to coordinate by knowing who should try to say what kind of things at a particular moment. Here we use the "imitation game" developed by de Boer for his experiments on the emergence of vowel systems.

A round of a game involves two agents, one being called the speaker, and the other the hearer. Here we just note that each possess a repertoire of items/syllables/prototypes, with a score associated to each of them (this is the categorical memory described below). The speaker initiates the conversation by picking up one item in its repertoire and utters it. Then the hearer tries to imitate this sound by producing the item in its repertoire that matches best with what he heard. Then the speaker evaluates whether the imitation was good or not by checking whether the best match to this imitation in its repertoire corresponds to the item uttered initially. Then he gives a feedback signal to the hearer in a non-linguistic manner (see Steels, 1998). Finally, each agent updates its repertoire. If the imitation succeeded, the scores of involved items increase. Otherwise, the score of the item used by the speaker decreases and there are 2 possibilities for the hearer: either the score of the prototype used was below a certain threshold, and this item is modified by the agent who tries to find a better one ; or the score was above this threshold, which means that it may not be a good idea to change this item, and a new item is created, as close to the utterance of the speaker as the agent can do given its constraints and knowledge at this time of its life. Regularly the repertoire is cleant by removing the items that have a score too low. Initially, the repertoires of agents are empty. New items are added either by invention, which takes place regularly in response to the need of growing the repertoire, or by learning from others.

## The production module

**Vocal tract**    A physical model of the vocal tract is used, based on an implementation of Cook's model (Cook 1989). It consists in modeling the vocal tract together with the nasal tract as a set of tubes that act as filters, into which are sent acoustic waves produced by a model of the glottis and a noise source. There are 8 control parameters for the shape of the vocal tract, used for the production of syllables. Finally, articulators have a certain stiffness and inertia.

**Control system**    The control system is responsible for driving the vocal tract shape parameters given an articulatory program, which is the articulatory specification of the syllable. Here we consider the syllable from the point of view of the frame-content theory (MacNeilage 1998) which defines it as an oscillation of the jaw (the frame) modulated by intermediary specific articulatory configurations, which represent a segmental content (the content) corresponding to what one may call phonemes. A very important aspect of syllables is that they are not a mere sequencing of segments by juxtaposition: co-articulation takes place, which means that each segment is influenced by its neighbors. This is crucial because it determines which syllables are difficult to pronounce and imitate. We model here co-articulation in a way very similar to what is described in (Massaro 1998), where segments are targets in a number of articulatory dimen-

sions. [1] The constraint of jaw oscillation is modeled by a force pulling in the direction of the position the articulators would have if the syllable was a pure frame, which means an oscillation without intermediary targets. This can be viewed as an elastic whose rest position at each time step is the pure frame configuration at this time step. It is motivated by important neuro-scientific evidence whose synthesis can be found in (MacNeilage, 1998). Finally, and crucially, we introduce a notion of articulatory cost, which consists in measuring on the one hand the effort necessary to achieve an articulatory program and on the other hand the difficulty of this articulatory program (how well targets are reached given all the constraints). This cost is used to model the principle of least effort explained in (Lindblom 1992) : easy articulatory programs/syllables tend to be remembered more easily than others. Agents are initially given a set of pre-defined targets that can be thought to come from an imitation game on simple sounds (which means they do not involve movements of the articulators) as described in (de Boer 2000, Steels and Oudeyer 2000). Although the degrees of freedom that we can control here do not correspond exactly to the degrees that are used to define human phonemes, we chose values (see Oudeyer 2001) that allow them to be good metaphors of vowels (V), liquids (C1) and plosives (C2), which mean respectively sonorant, less sonorant, and even less sonorant phonemes (sonority is directly related to the degree of obstruction of the air flow, which means the more articulators are opened, the more they contribute to a high sonority of the phoneme).

## The perception module

The ear of agents consists of a model of the cochlea, and in particular the basilar membrane, as described in (Lyon 1997). It provides the successive excitation of this membrane over time. Each excitation trajectory is discretized both over time and frequency: 20 frequency bins are used and a sample is extracted every 10 ms. Next the trajectory is time normalized so as to be of length 25. As a measure of similarity between two perceptual trajectories, we used a technique well-known in the field of speech recognition, dynamic time warping (Sakoe and Chiba 1980). Agents use this measure to compute which item in their memory is the closest. No segmentation into "phonemes" is done in the recognition process: the recognition is done over the complete unsegmented sound. Agents discover which phonemes compose the syllable only after recognition of the syllable and by looking at the articulatory program associated to the matched perceptual trajectory in the exemplar (see below). In brief, phonemes are not relevant for perception, but only for production. This follows a view defended by a number of researchers (Seguy, Dupoux et

Mehler 1995) who showed with psychological experiments that the syllable was the primary unit of recognition, and that phoneme recognition came only after.

## The brain module

The knowledge management module of our agents consists of 2 memories of exemplars and a mechanism to shape and use them. A first memory (the inverse mapping memory) consists of a set, limited in size, of exemplars that serve in the imitation process: they represent the skills of agents for this task. Exemplars are associations between articulatory programs and corresponding perceptual trajectories. The second memory (the categorical memory) is in fact a subset of the inverse-mapping memory, to which a score is added to each exemplar. Categorical memory is used to represent the particular sounds that count as categories in the sound system being collectively built by agents (corresponding exemplars are prototypes for categories). It corresponds to the memory of prototypes classically used in the imitation game (de Boer 1999).

Initially, the inverse mapping memory is built through babbling. Agents generate random articulatory programs, execute them with the control module and perceive the produced sound. They store each trial with a probability inverse to the articulatory cost involved (prob=1-cost). The number of exemplars that can be stored in this memory is typically quite limited (in the experiments presented below, there are 100 exemplars whereas the total number of possible syllables is slightly above 12000). So initially the inverse mapping memory is composed of exemplars which tend to be more numerous in zones where the cost is low than in zones where the cost is higher. As far as the categorical memory is concerned, it is initially empty, and will grow through learning and invention.

When an agent hears a sound and wants to imitate it, he first looks up in its categorical memory (if it is not empty) and find the item whose perceptual trajectory is most similar to the one he just heard. Then he executes the associated articulatory program (noise is always added to target values). Now, when the interaction is finished, in any case (either it succeeded or failed), it will try to improve its imitation. To do that, it finds in its inverse mapping memory the item (it) whose perceptual trajectory matches best (it may not be the same as the categorical item). Then it tries through babbling a small number of articulatory variations of this item that do not belong to the memory: each articulatory trial item is a mutated version of it, i.e. one target has been changed or added or deleted. This can be thought of the agent hearing at a point "ble", and having in its memory the closest item being "fle". Then it may try "vle", "fli", or even "ble" if the chance decides so (indeed, not all possible mutations are tried, which models time constraints: here they typically try 10 mutations). The important point is that these mutation trials are not forgotten for the future (some of them may be useless now, but very useful in the future): each of them is remembered with a probabil-

---

[1] The difference is that, as described in the companion paper (Oudeyer 2001), we provide a biologically plausible implementation inspired from a number of neuroscientific findings (Bizzi and Mussa-Ivaldi 1991) and that uses techniques developed in the field of behavior-based robotics (Arkin 1999).

ity inverse to its articulatory cost. Of course, as we have memory limitation, when new items are added to the inverse mapping memory, some others have to be pruned. The strategy chosen here is the least biased: for each new item, a randomly chosen item is also deleted (only the items that belong to categorical memory can not be deleted).

The evolution of inverse mapping memory implied by this mechanism is as follows. Whereas at the beginning items are spread uniformly across "iso-cost" regions, which means skills are both general and imprecise (they have some capacity of imitation of many kinds of sounds, but not very precise), at the end examplars are clustered in certain zones corresponding to the particular sound system of the society of agents, which means skills are both specialized and precise. This is due to the fact that exemplars closest to sounds produced by other agents are differentiated and lead to an increase of exemplars in their local region at the cost of a decrease elsewhere.

## Behavior of the model

### Efficiency

The first thing one wants to know is simply whether populations of agents manage to develop a sound system of reasonable size and that allows them to communicate (imitations are successful). Figure 1 and 2 show an example of experiment involving 15 agents, with a memory limit on inverse-mapping memory of 100 exemplars, with vocalizations comprising between 2 and 4 targets included among 10 possible ones (which means that at a given moment, one agent never knows more than about 0.8 percent of the syllable space). In figure 1, each point represents the average success in the last 100 games, and on figure 2, each point represents the average size of categorical memory in the population (i.e. the mean number of syllables in agents' repertoires). We see that of course the success is very high right from the start: this is normal since at the beginning agents have basically one or two syllables in their repertoire, which implies that even if an imitation is quite bad in the absolute, it will still get well matched. The challenge is actually to remain at a high success rate while increasing the size of the repertoires. The 2 graphs shows that it is the case. To make these results convincing, the experiments was repeated 20 times (doing it more is rather infeasible since each experiment basically lasts about 2 days), and the average number of syllables and success was measured in the last 1000 games (over a total of 20000 games): 96.9 percent is the mean success and 79.1 is the mean number of categories/syllables.

The fact that the success remains high as the size of repertoire increases can be explained. At the beginning, agents have very few items in their repertoires, so even if their imitations are bad in the absolute, they will be successfully recognized since recognition is done by nearest-neighbours (for example, when 2 agents have only 1 item, no confusion is possible since their is only
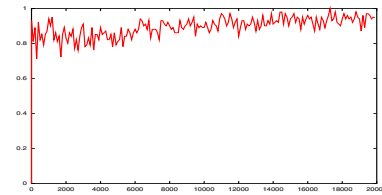


Figure 1: Example of the evolution of success in interactions for a society of agents who build a sound system from scratch
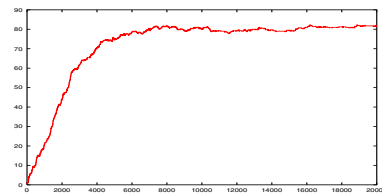


Figure 2: Corresponding evolution of mean number of items/categories in repertoires of agents along with time

1 category !). As time goes on, while their repertoires become larger, their imitation skills are also increasing : indeed, agents explore the articulatory/acoustic mapping locally in areas where they hear other utter sounds, and the new sounds they create are hence also in these areas. The consequence is a positive feed-back loop which makes that agents who knew very different parts of the mapping initially tend to synchronize their knowledge and become expert in the same (small) area (whereas at the beginning they have skills to imitate very different kinds of sounds, but are poor when it becomes to make subtle distinctions in small areas).

This result is relevant to all theories of speech (and more generally, theories of language), innatist or not. Indeed, whereas the literature is rich of reasons explaining why having complex sound systems was an advantage for the first speaking humans is, no precise account of how it could have been built was described. For instance, even Pinker and Bloom (1990), who defend the idea that nowadays humans have a lot of linguistic knowledge already encoded in the genes, acknowledge that it certainly got there through the Baldwin effect and so was initially certainly the result of a cultural process. They give cues about how acquired skills and sound systems could have been transferred into the genes, but not how they got to be acquired from a situation where there was nothing !

### Structural properties

Now that we have seen that a communication system was effectively built, one has to look whether the structural properties of the produced repertoires of syllables resemble human syllable systems. Indeed, human syllable systems are far from random: only very few combinations of types of phonemes occur in human languages compared to the high number of mathematically possible ones, and some occur significantly more often than others (Vennemann 1988) . For instance, all languages have CV sylla-

Table 1: % of syllable types in produced and random systems

| CV | CVC | CC |
|---|---|---|
| 25.3/0.2 | 20.1/1.3 | 16.1/0.5 |

| CCV | CVVC/CCVC/CVCC | other |
|---|---|---|
| 14.4/1.3 | 14.1/22.5 | 10/74.2 |

Table 2: % of syllables respecting the sonority hierarchy

| full model | jaw constraint removed | chance |
|---|---|---|
| 70.9 per | 21.5 per | 5.3 per |

bles, but CVCC is rare. The difference in frequencies exist both across and within languages. A first study about the syllable types of the produced systems was achieved. Statistics about the set of all the syllables produced by 20 runs were computed (for each run, measures were done after 20000 games). Table 1 sums up the result by giving the relative frequency in use of a number of syllable types (C means "C1 or C2"). "Relative frequency in use" means that each syllable counts as the number of times it has been used by the agents in the games it played in its life. This is a better measure than simply counting the frequency of occurrence in a syllable system, because it takes into account the fact that certain syllables tend to be adopted earlier than others, which implies that they are used more times than others, and models the relative frequency effects observed within languages. The second percentage measures the proportion of the particular type of syllable in the space of all combinatorially possible ones in the experiment. This can be viewed as a measure of syllable frequencies for randomly generated repertoires.

The first observation we can do is that there is a strong difference between the relative frequencies of syllables in actual systems and in randomly generated systems. Moreover, we find that the ordering between syllables types along their frequency is very similar to the one observed in human languages (Venemann 1988), except for the presence of CC in third position (which we think is due to too low acoustic noise, unlike in the real world). These results are rather consistent with those found by Redford, and conform to what she calls the "iterative principle of syllable structure": "simpler syllables types are expected to occur more frequently than complex ones in a systematic fashion", where the notion of "simplicity" is constructed over the most simple syllable CV: increase in complexity comes by adding C or V iteratively at the end or beginning or CV, and then after by replacing some C by V or the contrary.

A second important tendency of human languages is the "sonority hierarchy principle" [2] Whereas the measures in table 1 indicate that this seems to be the case in our experiment, they are too loose to conclude, especially because they blended C1 and C2. So we made measures over 20 runs about which proportion of syllables belong-

---

[2] in a syllable, the sonority or loudness first increases to a peak and then may decrease again. It is very rare that for instance it first decreases and then increases or that more than 1 change in sonority direction occurs in one syllable. For instance, "ble" is preferred to "lbe". Sonority/loudness is directly linked to the degree of obstruction of the air, and in particular to the degree of opening of the jaw.

ing to the repertoire of agents did obey the sonority hierarchy principle, using the fact that sonority of V is higher than sonority of C1, which is higher than sonority of C2 (due to the way they obstrue the air flow). Additionally, we made an experiment in which the oscillation of the jaw constraint was removed, in order to evaluate the hypothesis of Peter McNeilage that says that it is the main explanation for the sonority hierarchy principle. Table 1 sums up the results, with a column showing what is the proportion of syllable in the set of combinatorially possible syllables that respect this hierarchy.

We see that the sonority hierarchy is respected by most of the syllables of the emergent repertoires in the standard model. Yet, not all of them respect it, which is not that surprising since syllables like C1C1V do not imply an important deformation of the pure frame and so have a low cost, and do not respect the principle (there are 2 adjacent segments with the same sonority). Anyway, the actual percentage as compared to chance is much higher. When we remove the jaw constraint, we observe that the percentage of syllables respecting this hierarchy drops to around 20 percent, but is still substantially above chance. It indicates that the jaw constraint is crucial, but not the only responsible. In fact, when we remove the jaw constraint, we still start every syllable with the rest position corresponding to the closed jaw. So for instance syllables beginning with a vowel will still have a high articulatory cost. Of course for example C2C2 syllables will have a much lower cost in this case that in the case with jaw oscillation, but these syllables are very sensible to noise and do not have a high perceptual discriminability, which makes agents prune them quite often. As a result, a reasonable proportion of syllables that respect the hierarchy remain.

Until now, we have only looked at how the model produced syllable systems that reflect universal tendencies of human languages. We also have to look how well it matches with the diversity that exists across languages (Vennemann, 1988). Indeed, tendencies are just tendencies and there are cases of languages whose syllable systems properties significantly differ from the mean (for example, in Berber, there are many syllables with long consonant sequences, and more strikingly, there are syllables that do not contain any vowel). Additionally, two languages that have for instance the same relative ratios of syllable types may implement these in very different manners. The first kind of diversity was difficult to observe in a statistically significant manner, since the relative frequencies of syllable types most often are very close to the mean above mentioned, and since not enough experiments were conducted to study rare outliers. Nevertheless, they were observed in a number of particular cases: for example, one of the obtained population

had 55 percent of CVC/CCV syllables against only 20 percent of CV syllables. Some categorical differences were also observed: several populations did not have any CVVC or CVCC syllables for instance. The second kind of diversity was easier to observe in the system: you never get the same repertoires in 2 different runs of the experiment. In the 20 runs used for the experiments above, the mean number of common syllables was 20.2 (repertoires had sizes varying between 70 and 88), among which mainly 2-phonemes syllables due to their small number. Of course this result is not directly transposable to real languages since we always gave here the same set of phonemes in the beginning, whereas in reality these phonemes are not pre-given but should co-evolve with syllables, and so may lead to repertoire of syllables composed of very different phonemes [3]. Nonetheless, we get a good idea of how universal tendencies come from the fact that there are non-linguistically specific constraints/biases in the problem that agents are solving, whereas diversity comes from both the fact that these constraints are soft and that there exist many satisfying solutions to the problem. Operationally speaking, variety emerges because there is stochasticity locally in time and space, which makes that different societies may engage different pathways due to historical events: indeed, historicity is fundamental to the explanation of diversity. This view contrasts in different aspects with a number of innatist theories, especially optimality theory (Archangeli and Lagendoen 1997). Of course, there is a common point with optimality theory at a very general level: constraints are crucial to the explanation of language universals and diversity. Yet, a fundamental difference is the nature of constraints: in the case of optimality theory, they are linguistically specific, whereas here they are generic constraints of the motor, perceptual and cognitive apparati (we also have social constraints that are far from any concept in OT) [4]. Now, the second important difference is the way these constraints are used to explain diversity: in OT, a particular syllable system corresponds to a particular ordering of constraints (some are stronger than others, which means that a low ranked constraint may be over-ridden if one has to satisfy a higher ranked constraint), which means a different constraint satisfaction problem. Conversely, in OT, one ordering of constraint implies a fixed syllable system (in terms of syllables types). On the contrary, here we do not require a different set of constraints to obtain different kinds of systems, because there are many syllables systems that can be developed and allow efficient communication given only one set of constraints. Our model thus avoids a number of theoretical problems that OT is faced with: Where do the linguistic constraints come from ? If they are in the genes, how did they get there ? Why are there different orderings of constraints ? How one can pass from a set of constraints to another (which must happen since language evolves and syllable systems change) ?

## Conclusion

We have presented an operational model of the origins of syllable systems whose particularity is the stress on embodiment and situatedness constraints/opportunities, which implies the avoidance of many shortcuts usually taken in the literature. It illustrates in details (and brings more plausibility) the theory which states that speech originated in a cultural self-organized manner, taking as a starting point a set of generic non-linguistically specific learning, motor and perceptual capabilities. In addition to the demonstration of how an efficient communication system could be built with this parsimonious starting point, some specific properties that are known about human sound systems can be explained by our model : on the one hand, universal tendencies like the preference for CV and CVC syllable types and the sonority hierarchy principle ; on the other hand, diversity. A forthcoming paper will present other properties of human sound systems predicted by this model, among which the critical period phenomenon, the difficulty to learn a second language and the difficulty to learn artificial random sound systems as compared to "natural ones".

## References

Altman, (1995), Cognitive Models of Speech Processing, Psycholinguistics and Computational Perspectives, MIT Press.

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

Arkin, R. (1999) Behavior-based Robotics, MIT Press.

Bizzi E., Mussa-Ivaldi F., Giszter S. (1991) Computations underlying the execution of movement, Science, vol. 253, pp. 287-291.

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.

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

P. R. Cook, "Synthesis of the Singing Voice Using a Physically Parameterized Model of the Human Vocal Tract," Proc. of the International Computer Music Conference, pp. 69-72, Columbus, OH, 1989.

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

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.

Lyon, R. (1997), All pole models of auditory filtering, in Lewis et al. (eds.) Diversity in auditory mechanics, World Scientific Publishing.

Massaro, D. (1998) Perceiving talking faces, MIT Press.

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

Oudeyer P-y. (2001) The origins of syllable systems in a society of truly autonomous robots, submitted to Artificial Intelligence Journal.

Oudeyer P-y. (2001) Coupled Neural Maps for the Origins of Vowel Systems, to appear in the proceedings of ICANN'01, International Conference on Artificial Neural Networks, Springer Verlag.

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

Redford, M.A., C. Chen, and R. Miikkulainen (1998) Modeling the Emergence of Syllable Systems. In: Proceedings of the Twentieth Annual Conference of the Cognitive Science Society. Erlabum Ass. Hillsdale.

Sakoe H., Dynamic programming optimization for spoken word recognition, IEEE Transaction Acoustic., Speech, Signal Processing, vol. 26, pp. 263-266.

Segui, J., Dupoux E., Mehler J. (1995) The role of the syllable in speech segmentation, phoneme identification, and lexical access, in Altman, (ed.), Cognitive Models of Speech Processing, Psycholinguistics and Computational Perspectives, MIT Press.

Steels, (1998), Synthesizing the origins of language and meaning using co-evolution, self-organization and level formation, in Hurford, Studdert-Kennedy, Knight (eds.), Cambridge University Press, pp. 384-404.

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

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

---

[3] This is a limit of the model (that the model of Redford has also), but we think this limitation was necessary as a first step so that the resulting dynamics would not get too complicated to analyse.

[4] An example of constraint in OT is the *COMPLEX constraint which states that syllables can have at most one consonant at an edge or the NOCODA constraint which says that syllables must end with vowels.