Self-Organization in the Evolution of Shared Systems of Speech Sounds: a Computational Study

Pierre-Yves Oudeyer

Sony Computer Science Laboratory, Paris,
http://www.csl.sony.fr/~py

Abstract

How did culturally shared systems of combinatorial speech sounds initially appear in human evolution? This paper proposes the hypothesis that their bootstrapping may have happened rather easily if one assumes an individual capacity for vocal replication, and thanks to self-organization in the neural coupling of vocal modalities and in the coupling of babbling individuals. This hypothesis is embodied in agent-based computational experiments, that allow to show that crucial phenomena, including structural regularities and diversity of sound systems, can only be accounted if speech is considered as a complex adaptive system. Thus, the second objective of this paper is to show that integrative computational approaches, even if speculative in certain respects, might be key in the understanding of speech and its evolution.

Index Terms: origins of speech, self-organization, evolution, universal tendencies, diversity, agent-based modelling, neural maps, articulatory-auditory coupling.

1. The origins of language

Unravelling the mechanisms at the origins of language has been a particularly active research area in the recent years [1, 2, 3, 4]. Three interrelated questions form the basis of this research agenda:

1) What are the biological pre-requisites that allow language learning and use?
2) How these biological capabilities evolved phylogenetically?
3) Given the biological capabilities, how new languages can form and evolve culturally in a society of individuals?

The depth of these questions has gathered specialists from many disciplines, including anthropology, linguistics, developmental psychology, cognitive neuroscience, physiologists, ethologists, evolutionary biologists, roboticists and computational modellers [5]. Understanding the origins of language involves a multitude of components interacting in complex ways in parallel on several timescales: the ontogenetic timescale, characterizing the growth of an individual person, the glossogenetic or cultural timescale which characterizes the evolution of cultures, and the phylogenetic timescale, which characterizes the evolution of species. A growing number of researchers have argued that the study of the interactions between these components and these dimensions is crucial, and that language, and its origins in particular, can only be understood as a complex adaptive system [6, 7]. The sciences of complexity have taught us that in many of the complex systems in nature, there are global phenomena that are the irreducible result of local interactions between components whose individual study would not allow us to see the global properties of the whole combined system [8, 9, 10, 11]. For example, it is not possible to understand the structure and building of termite nests by only studying individual termites (see figure 1). A termite does not possess any map of a nest, and its behaviour is mainly reactive. This is only by considering thousands of termites in interaction that ethologists managed to understand how nests are self-organized [12]. Seemingly, many properties of language might not be directly encoded in any of the components involved, but might be the self-organized outcomes of the interactions of the components.

Unfortunately, self-organized phenomena are often complicated to understand or to foresee intuitively, and to formulate in words. It is for this reason that computational models play a key role in this research area [13]. These systems are put...
to two main types of use: 1) they serve to evaluate the internal coherence of verbally expressed theories that have already been proposed, by clarifying all their hypotheses and verifying that they do indeed lead to the proposed conclusions (and quite often one discovers errors in the assumptions as well as in the conclusions, which need to be revised); 2) they serve to explore and generate new theories, which themselves often appear when one simply tries to build an artificial system reproducing the verbal behaviour of humans. A number of important results have already been obtained and have opened new ways for the resolution of previously unanswered questions: the decentralized generation of lexical and semantic conventions in populations of agents (e.g. [14, 15, 16, 17]), the formation of shared inventories of vowels or syllables in groups of agents, with features of structural regularities greatly resembling those of human languages (e.g. [31, 32, 18, 19, 13]), the formation of conventionalized syntactic and grammatical structures (e.g. [20, 21, 22]), the conditions under which combinatoriality, the property of systematic re-use, can be selected (e.g. [23, 24])2.

The following sections will present an overview of a particular example of such computational models (for a more detailed presentation and discussion, and for more sophisticated versions of this model, see [13]). It is focused on the origins of speech, and our goal is to illustrate how it can help formulate new original hypothesis and test their coherence. More precisely, it will address the issue of understanding how systems of shared speech sounds might have first bootstrapped in human evolution. The hypothesis is that thanks to self-organization, the biological transition between the ability to replicate vocalizations and the appearance of culturally shared speech systems might have been rather small.

2. How did shared combinatorial systems of speech sounds form?

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/acoustic units, which are themselves discrete elements carved up by each linguistic community in the articulatory/acoustic continuum. Moreover, the repertoires of these elementary units 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 very rarely 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 question. 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.

Cognitive innatism One possible answer, proposed by cognitive innatism [25], is that speech relies 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 make it explicit 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.

Functionalism 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 [26, 27]. 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 explain why each linguistic community has a different speech code and how one specific speech code can be “chosen” by a community.

Agent Based models Another answer, based on agent-based modelling [28, 29, 30] and focused on the question of the origins of vowels systems, was proposed by several researchers in [31, 32, 18] and does not have these limits. Let us take the example of the representative model developed by de Boer [18]. 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 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 [19], which showed how similar results could be obtained concerning the formation of shared syllable systems with the prediction of regularities in syllables structures.

Origins of language vs. origins of languages Nevertheless, if the “imitation game” is a good framework for studying the evolution of modern speech systems, 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 behavioural capabilities for agents, and assumes implicitly 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

---

2This list is in no way exhaustive, and more examples can be found in [3].
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 needs 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 explicitly 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 (“l’origine des langues” in French) than a model of the origins of the language capacity (“l’origine du langage” in French).

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 [33]). 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 whale species, vocal replication has been argued to be useful for the maintenance of social cohesion [34]. 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 distinctive vocal units serving as categories to cut a perceived trajectory into high-level segments. Moreover, in the model that I will present, there is no explicit pressure for building such a repertoire of distinctive units. Such a pressure could be introduced explicitly, like in experiments presented in [19], and would allow larger repertoires, but our aim here is to show that it is not a necessary assumption for bootstrapping shared combinatorial speech codes. Moreover, it is also possible, if we frame the experiments in an evolutionary scenario, that the assumed neural kit for vocal replication appeared under a pressure for linguistic communication. But there are other possibilities, as stated above, and the experiments remain neutral in this respect. Finally, agents do not interact in an organized manner (there is no “game” or “protocol” of interaction, which also reduces the amount of assumptions present in the model). 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 changes (even maybe no change at all) in order to generate speech codes that have the crucial properties of modern speech: in short, the evolutionary step from non-speech to speech may have been rather small.

3. Coupling perception and production in a model of vocal replication

The basic neural kit for 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 replication by oneself of a movement performed by someone else. As shown by the computational literature (e.g. [35]), the most simple system which can do this is basically a neural architecture 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 typically hebbian connections and whose purpose is to allow the transformation of the trajectory from one space to the other. Figure 2 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 instantiate this architecture in the context of vocal mimicry: the perceptual space will be auditory, and the motor space will be articulatory.

Overview of the computational system

Each agent has one artificial ear which takes measures of the vocalizations that it perceives, which are then sent to its brain. It also has a physical model of the 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 [36, 35], which are the neural devices that humans have to represent parts of the outside world (acoustic, visual, touch etc.).

Figure 2: The minimal neural kit for motor replication/mimicry.
describe the technical details of the architecture. Figure 3 gives an overview of the architecture. We will now
units from the vocalization continuum.
change. Due to stochasticity and positive feed-back loops, the
neural maps are updated and the sound distribution they model
perceive vocalizations produced by others, the neurons in each
self-organize and agents learn the correspondences. As they
babble, the neural connections between the articulatory and acoustic maps
vocalization continuum when they babble. As they babble, the
maps are random and agents produce sounds that span the all
produce a similar vocalization in the future. Initially, the neural
intermediate configurations between each articulatory objective,
neurons are activated in sequence. Each activation of a
geometrically, when
activated, the more the agent will produce articulations which
approximate the continuous sounds. Here, these points are in the
4-D space whose dimensions are the first four formants of the
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 [18].

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 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 vocalization.

Figure 3: 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. Their interactions are minimalistic: they move randomly in a simulated world, and babble with their vocal apparatus, which produces sounds heard by neighbours. The perception of a vocalization does not trigger any reactive behaviour, its only consequence is a modification of the internal neural structures which improve the probability to produce a similar vocalization in the future. Initially, the neural maps are random and agents produce sounds that span the all vocalization continuum when they babble. As they babble, the neural connections between the articulatory and acoustic maps self-organize and agents learn the correspondences. As they perceive vocalizations produced by others, the neurons in each neural maps are updated and the sound distribution they model change. Due to stochasticity and positive feed-back loops, the system will crystallize in a state in which agents begin to produce a shared set of combinatorial vocalization, carving discrete units from the vocalization continuum.

**Articulatory objective**

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 [18]. 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 [37]: lip rounding, tongue height and tongue position. The values within these dimensions are between 0 and 1, and a triplet of values \( ar = (r, h, p) \) 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:

\[ v_{j,t+1} = v_{j,t} + 0.001 G_{j,t}(s)(v_m - v_{j,t}) \]
acoustic signal. We then use a model of the cochlea, described in [38] and [18], 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 $\sigma$ 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 $s$:

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

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 [39]:

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

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 [40].

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-critical 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-critical parameter of the simulation: nothing changes except the speed of convergence.

4. 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 different targets that compose the vocalizations of agents are also randomly and uniformly distributed. Figure 5 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 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 that of Figure 6 which corresponds to Figure 5 after 2000 interactions in a population of 20 agents. Figure 6 shows that the distribution of preferred vectors is no longer uniform but...
Figure 5: 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.

Figure 6: Neural maps after 2000 interactions, corresponding to the initial state of figure 5. 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. 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 6, 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 7 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 combinatorial. 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 discretness and combinatoriality 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 combinatoriality arise thanks to the coupling between perception and production within agents, and 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 $\sigma$ which defines the width of the tuning functions. All the results presented are with a value 0.05. In [13], 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 $\sigma$ becomes too small, then the initial equilibrium of the system becomes stable and nothing changes: agents keep producing inarticulate and holistic vocalizations. If $\sigma$ is too large, then 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 systems.

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 [41] on universals in vowel systems, and is identical to the type of classification performed in [18]. The first result shows that the distribution of vowel inventory sizes is very similar to that of human vowel systems [37]: 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 [18] 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 [42]. 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 [43]. The results are shown in Figure 9. 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 8 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 [18]. 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.

5. Discussion

In this paper, we have shown a computational experiment in which a population of agents endowed with a minimal neural kit for vocal replication, self-organized a shared combinatorial speech code with structural regularities and diversity. This result is conditioned 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). Yet, the capability of vocal replication is not diminished if this value gets out of this interval within certain limits. This allows us to understand that, in an evolutionary scenario where our ancestors were capable of vocal replication but not of language, the formation of culturally shared combinatorial speech codes might have been the result of a relatively small biological change: the tuning of the width of neuronal receptive fields (for a more complete explanation, please see [13]).

Of course such a computational experiment is not a "model" in the traditional sense since some of its assumptions are speculative (yet plausible), and should not be seen as a direct proposal for evolutionary explanation. The primary goal of such a model is to develop our intuitions about the potential role of self-organization in the evolution of speech. Such experiments show that fundamental phenomena happen only when one couples simple perceptual, motor, neural and interactional mechanisms. They would never be discovered by research focused for example either on the physiology of ears or vocal tracts, or on neural pathways for speech processing, or on individual speech acquisition mechanisms. Focused and reductionist approaches are necessary, even crucial, but they are not sufficient. Their juxtaposition is also not sufficient. It might be argued that it is useless to build integrated models since we are very far from understanding components in such a way that our assumptions can be well-grounded. But this is a misleading view: while computational experiments with integrated systems will not provide testable theories of the origins of speech in the short term, they are likely to be key for the development of intellectual tools that will open new avenues for research. The understanding of bird flight has only been possible through the building of generations of (non-)flying machines that were exploring and testing speculative ideas about aerodynamics. Likewise, it is probable that the understanding of speech origins might require the building of explorative artificial systems.

6. Acknowledgements

This research has been partially supported by the ECAGENTS project founded by the Future and Emerging Technologies pro-
7. References


