HAL
open science

Emergence of articulatory-acoustic systems from deictic interaction games in a "vocalize to localize" framework

Clément Moulin-Frier, Jean-Luc Schwartz, Julien Diard, Pierre Bessière

To cite this version:

Clément Moulin-Frier, Jean-Luc Schwartz, Julien Diard, Pierre Bessière. Emergence of articulatory-acoustic systems from deictic interaction games in a "vocalize to localize" framework. Anne Vilain and Jean-Luc Schwartz and Christian Abry and Jacques Vauclair. Primate communication and human language: Vocalisations, gestures, imitation and deixis in humans and non-humans, John Benjamins Pub. Co., pp.193-220, 2011, Advances in Interaction Studies. hal-00961125

HAL Id: hal-00961125

https://hal.science/hal-00961125v1

Submitted on 19 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Emergence of articulatory-acoustic systems from deictic interaction games in a "Vocalize to Localize" framework

Clément Moulin-Frier[1], Jean-Luc Schwartz[1], Julien Diard[2], Pierre Bessière[3]

1: GIPSA-Lab, ICP

FirstName.LastName@gipsa-lab.inpg.fr

2: Laboratoire de Psychologie et de NeuroCognition

julien.diard@upmf-grenoble.fr

3: Laboratoire d'Informatique de Grenoble

Pierre.Bessiere@imag.fr

CNRS – Grenoble University

Table of Contents

1 Introduction 3
2 Deriving Morphogenesis Theories from Origins Theories 3
2.1 Morphogenesis Theories 3
2.1.1 The Dispersion Theory 4
2.1.2 The Quantal Theory 4
2.1.3 The Perception for Action Control Theory (PACT) 5
2.2 Origins theories: Vocalize to Localize 6
2.3 An integrated framework 7
3 Computational models of language emergence in a society of interacting agents 8
4 Modelling 10
4.1 General principles 10
4.1.1 Deictic games 10
4.1.2 Agents knowledge 10
4.1.3 Sensory-motor systems 11
4.2 Bayesian modelling 12
4.2.1 Mathematical requirements 13
4.2.2 The inference model (declarative phase) 13
4.2.3 The interaction behaviours (procedural phase) 15
4.2.3.1 Reflex behaviour 15
4.2.3.2 Communicative behaviour 15
4.2.3.3 Hybrid behaviour 16
5 Results 16
5.1 Technical details 16
5.2 Simulations 18
5.2.1 Results for the 1-D sensory-motor system 18
5.2.1.1 Reflex behaviour 19
5.2.1.2 Communicative behaviour 19
5.2.1.3 Hybrid behaviour 20
5.2.1.4 Conclusion for the 1-D sensory-motor results 21
5.2.2 Results for the VLAM sensory-motor system 21
6 Conclusions and perspectives 23
7 References 24

1 Introduction

Since the 70's and Lindblom's proposal to “derive language from non-language” (Lindblom, 1984, p. 78), phoneticians have developed a number of “substance-based” theories. The starting point is Lindblom’s Dispersion Theory (Liljencrants & Lindblom, 1972) and Stevens’s Quantal Theory (Stevens, 1972, 1989), which open the way to a rich tradition of works attempting to determine and possibly model how phonological systems could be shaped by the perceptuo-motor substance of speech communication. These works search to derive the shapes of human languages from constraints arising from perceptual (auditory and perhaps visual) and motor (articulatory and cognitive) properties of the speech communication system: we call them “Morphogenesis Theories”.

More recently, a number of proposals were introduced in order to connect pre-linguistic primate abilities (such as vocalization, gestures, mastication or deixis) to human language. For instance, in the “Vocalize-to-Localize” framework that we adopt in the present work (Abry & al., 2004), human language is supposed to derive from a precursor deictic function, considering that language could have provided at the beginning an evolutionary development of the ability to “show with the voice”. We call this type of theories “Origins Theories”.

We propose that the principles of Morphogenesis Theories (such as dispersion principles or the quantal nature of speech) can be incorporated and to a certain extent derived from Origins Theories. While Morphogenesis Theories raise questions such as “why are vowel systems shaped the way they are?” and answer that it is to increase auditory dispersion in order to prevent confusion between them, we ask questions such as “why do humans attempt to prevent confusion between percepts?” and answer that it could be to “show with the voice”, that is, to improve the pre-linguistic deictic function. In this paper, we present a computational Bayesian model incorporating the Dispersion and Quantal Theories of speech sounds inside the Vocalize-to-Localize framework, and show how realistic simulations of vowel systems can emerge from this model.

In Section 2, we present the Morphogenesis and Origins Theories on which we shall concentrate our work, and in Section 3 we propose a survey of previous computer simulations of the emergence of some properties of language from interactions between artificial agents. Section 4 provides all methodological details about models and implementations. Section 5 describes simulation results, from simple test cases to more realistic simulations dealing with vowel systems in human languages. A discussion and perspectives towards simulations of more complex phonological sequences are proposed in Section 6.

2 Deriving Morphogenesis Theories from Origins Theories

In this part, we first expose the principles of three Morphogenesis Theories: the Dispersion Theory (DT), the Quantal Theory (QT), and the Perception-for-Action-Control Theory (PACT). Then, we expose an Origins Theory that provides our framework: “Vocalize-to-Localize”. Finally, we propose an integrating framework to incorporate DT, QT and PACT into “Vocalize-to-Localize”.

2.1 Morphogenesis Theories

Phonological systems are far from arbitrary combinations of available phonemes, as showed by the very limited number of phoneme combinations in human languages, compared with the total number of possible combinations provided by a simple combinatory rule (Boë et al., 2002). For instance, in the case of vowel systems that we shall use as a test case in the

following, there is a strong bias in favour of 5-vowel systems in terms of vowel number, and whatever this number is, most systems contain /i a u/ (Boë et al., 2002). Morphogenesis theories attempt to explain this kind of regularity. For this aim, they often propose to relate the universal tendencies to the minimization of a global score characterizing some perceptual or motor properties of a given system.

2.1.1 The Dispersion Theory

The first quantitative simulations of vowel inventories are due to Liljencrants and Linblom (1972), with their Dispersion Theory based on the maximization of auditory distances. In this framework, vowel systems tend to minimize the function:

$$G = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{1}{d_{i,j}} \right)^2$$

where n is the number of vowels and $d_{i,j}$ a perceptual distance between the vowels i and j. Various distances were considered. In their seminal paper, Liljencrants & Lindblom first considered distances in the (F1, F2) formant space, with rather good predictions of vowel systems. Particularly, this explained why /i a u/, which are at the vertices of the vocalic triangle in the (F1, F2) space, are present in most world languages. F2 was then replaced by F'2, a "perceptual formant" integrating in a nonlinear way the effects of F2 and higher formants F3 and F4. Other auditory distances directly computed on the whole spectrum were also considered (Lindblom, 1986). Schwartz et al. (1997) later argued that an additional cost related to local spectral preferences for "focal vowels" with close values of either F1 and F2, or F2 and F3, or F3 and F4, should be introduced in the predictions (Dispersion-Focalization Theory).

In 1986, Lindblom suggested to introduce an articulatory cost in the optimization function (Lindblom, 1990). Thus, this new version is not only centered on the listener's interest (by the maximization of perceptual contrasts), but also on the speaker's interest (by the minimization of articulatory effort). This lead to the "Adaptive Variability Theory" (also known as "Hyper-Hypo"), in which the function to minimize becomes:

$$G = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{A_{i,j}}{d_{i,j}} \right)^2$$

where $A_{i,j}$ represents the articulatory cost between the phonemes i and j. This allowed Lindblom to simulate some effects of the number of vowels on the distribution of sounds in the vocalic space, with more extreme configurations for systems with a larger number of vowels.

2.1.2 The Quantal Theory

In the Quantal Theory, Stevens (1972, 1989) proposes that nonlinearities in the articulatory-to-acoustic or acoustic-to-auditory transformations shape the phoneme selection. Such nonlinearities may contrast regions where articulatory variations produce small auditory variations (stability regions, I and III in Figure 1), with on the other hand instability regions where small articulatory variations lead to large auditory shifts (II). Stevens describes a number of such potential nonlinearities, and argues that phonological systems might exploit these patterns to set a contrast around instability regions, with one phoneme in the stable region I, and the other one in the stable region III, region II playing the role of a kind of natural boundary for this contrast.

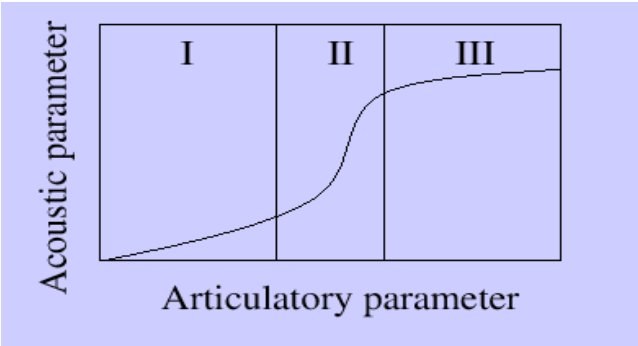


Figure 1: Non-linearity in the articulatory to acoustic transformation (Stevens, 1989).

This is for instance the case when you start from an /i/ with spread lips (lip rounding being the controlled articulatory parameter) and then progressively round the lips towards /y/. While the gesture at the beginning does almost not change the sound at all, the shift from an [i]-like to an [y]-like sound is quite abrupt, before a new stable region around the rounded [y] (e.g. Abry et al., 1989).

2.1.3 The Perception for Action Control Theory (PACT)

PACT (Schwartz et al., 2002, 2007) is a sensory-motor theory of speech communication, which attempts a synthesis inside the long history of debates between motor and auditory theories of speech perception.

On the one hand, motor theories consider that the objects of speech perception are gestures and not sounds, that is, the listener reconstructs the motor gesture from the auditory percept (e.g. Liberman & Mattingly, 1985; Fowler, 1986; Liberman & Whalen, 2000).

On the other hand, auditory theories consider that speech perception functions independently of the way the speech stimuli are produced by the articulatory system, hence there is no need to incorporate any knowledge about speech production within speech perceptual processing systems (e.g. Nearey, 1997; Massaro, 1987).

The Perception-for-Action-Control Theory claims that there are problems in both approaches.

First, motor theories fail to provide efficient predictions about regularities of phonological systems. Let us take an example in oral vowel systems. There are basically three degrees of freedom for producing oral vowels: height, front-back position, and rounding. This results in a 3-D articulatory space, illustrated in Figure 2a (with a shrinking of the space for open configurations, for which the front–back and rounding dimensions play a less important role).

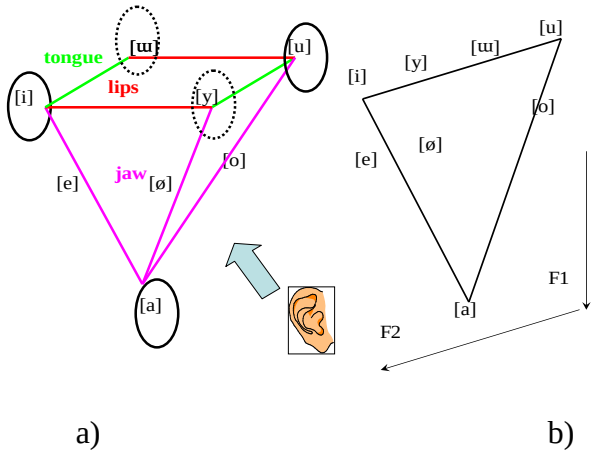


Figure 2: (a) The articulatory three-dimension space of oral vowels together with (b) its auditory projection (Schwartz & al., 2007)

What would be the best three-vowel system in this space? The system /i a u/ is a very good choice, in terms of articulatory dispersion, and it is indeed present in most world languages, as said previously. However, /y a ɯ/ provides as good a choice. It combines articulatory features differently, but the difference cannot be assessed in articulatory terms. However, this second system never appears in human languages. The reason for this is obviously an auditory one. Auditory perception is a kind of lateral projection of this 3-D space, in a 2-D (F1, F2) space (Figure 2b) in which [i u] is of course much better (in terms of dispersion) than [y ɯ]. The prevalence of /i a u/ and the absence of /y a ɯ/ clearly shows that gestures are shaped by perception.

On the other way round, auditory theories have difficulties to explain a number of phenomena where speech production leads to principled variability in speech stimuli (e.g. Fowler, 1986). Let us take the example of the vowel reduction phenomenon that is the fact that listeners are able to recover targets from coarticulated speech and particularly from reduced speech. Previous work showed that a stable articulatory target [a] can be recovered by acoustic-to-articulatory inversion, in spite of acoustic variability due to reduction in an [iai] sequence (Lœvenbruck and Perrier 1997). This suggests that listeners are able to recover the speaker's intentions, hence the need to introduce motor knowledge in speech perception (see Schwartz, 2008a).

The PACT proposes a synthesis of the motor and auditory views (Schwartz et al., 2002, 2007). In this framework, the objects of speech perception are neither purely auditory nor purely motor. They are rather multi-sensory percepts regularized by knowledge of speech production, or speech gestures shaped by perceptual processes. This sensory-motor conception also has neuroanatomical foundations through the so-called "dorsal route" of speech perception in the human cortex, linking temporal areas, considered as specialized in auditory processing and audiovisual fusion, with parietal areas, making the junction with somatosensory representations and possibly with amodal phonological representations, up to frontal areas (motor, premotor and prefrontal) connected with speech production and action understanding (Hickok & Poeppel, 2000, 2007; Skipper et al., 2007).

2.2 Origins theories: Vocalize to Localize

After a long period in the Twentieth century during which the question of language origins

was considered as taboo or scientifically unsound, the last twenty years have seen a strong emergence of proposals and debates on this topic. We shall not recall here all the elements of this debate. The present book is largely devoted to such discussions, for example about the gestural vs. orofacial precursors of human language. We shall only recall here some basic aspects of the "Vocalize-to-Localize" framework (Abry et al., 2004) that provides the selected background for the present work.

Deixis is the ability to show to a partner somebody or something in the surrounding world. Deictic abilities have been observed in monkeys and apes, involving both the orofacial and manual systems (the voice and the hand), as shown e.g. in the contributions by Zuberbuhler et al., and Hopkins et al. in the present volume. In the Vocalize-to-Localize framework, it is assumed that pointing is a precursor of language emergence, providing some bootstrap to the derivation of language from non-linguistic communicative abilities in phylogeny. It is furthermore proposed that pointing allowed a connection between the hand and the mouth, vocalizations enabling to "show with the voice" at distance, as is the case for alarm calls. Language would have thus emerged from the possibility to "localize by vocalizing".

From an ontogenetic point of view, developmental studies clearly show the importance of the coordination between manual and vocal actions in the development of language (see Volterra, this volume) and particularly the link between pointing gestures and vocalizations appearing just before the primary syntactic acquisition of the two-words sequence (Goldin-Meadow & Butcher, 2003; Volterra et al., 2005).

Another important component of the "Vocalize-to-Localize" framework is that the emergence of a vocal communication system would have required an efficient system for producing contrastive vocalizations. This is the point where the connection is done with MacNeilage and Davis's Frame-Content Theory (1998, 2000) deriving this ability from mastication, the jaw playing a crucial bootstrap role for producing efficient modulations, naturally swapping consonants and vowels (see MacNeilage, and Davis, this volume). Finally, the role of perceptual shaping of speech gestures, in the context of the previously described "Perception-for-Action-Control Theory" (PACT) is considered as essential for efficient communication.

Thus, in the Vocalize-to-Localize framework, language would build up around three basic components (see e.g. Abry & Ducey, this volume; Schwartz, 2008b):

- a hand (and a pointing and joint attention system) to show the world and produce meaning,
- a jaw (and a system for the production of orofacial actions) to achieve vocal modulations and naturally and efficiently swap consonants and vowels,
- an ear (and an eye, both connected to an audiovisual perception system) for structuring the sound flow into intelligible perceptual units.

2.3 An integrated framework

We claim that Origins Theories should encapsulate Morphogenesis Theories. While Morphogenesis Theories explore the conditions providing an efficient perceptuo-motor system for sound communication ("how to communicate?"), Origins Theories enable to embed these conditions into a rationale for communication ("why communicate?"). Instead of explaining the universals of human languages by more or less ad-hoc constraints, the aim is to derive universals directly from possible pre-linguistic functions (Figure 3). This is the purpose of the present work, in which we intend to show how a society of interacting agents, equipped with some pre-linguistics deictic abilities, could let language emerge, display some of its universal tendencies and analyze its behaviour in relation with some principles of the three Morphogenesis Theories described previously.

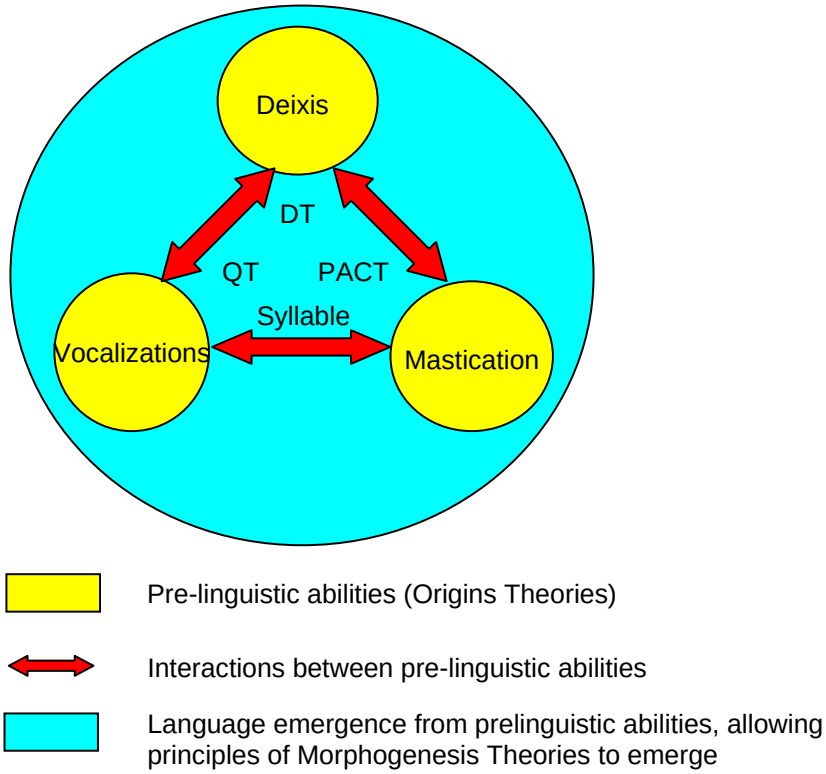


Figure 3 : Principles of Morphogenesis Theories can emerge from Origins Theories [DT : Dispersion Theory (Lindblom, 1972); QT : Quantal Theory (Stevens, 1989) ; PACT : Perception for Action Control Theory (Schwartz & al., 2007)]

3 Computational models of language emergence in a society of interacting agents

The pioneer studies by Steels in the middle 90's (e.g. Steels, 1996, 1997) opened the route to a new area of computer simulations towards "evolutionary linguistics" in which some properties of language should emerge from computational interactions between communicating artificial agents. Importantly, the interaction paradigms in these simulations intrinsically combine the "why" and the "how" questions: agents interact in some way, for some reason and through some means that the programmer must define, explain and hopefully relate to an evolutionary scenario.

According to Steels (2006), four steps are involved in setting up computer simulations:

1. Hypotheses about a link between pre-existing cognitive mechanisms and external factors and the emergence of a specific language feature.
2. Computational operationalisation of these mechanisms into "simulated agents" endowed with these processes.
3. Definition of an interaction scenario, possibly embedded in some simulation of the surrounding world, and hopefully capturing critical properties for communication.
4. Experimentation with computer simulations letting the features of interest emerge through interactions between agents.

Steels makes it clear that “this still does not prove anything about human language evolution because there may be multiple mechanisms to handle the same communicative challenges, but at least it shows a possible evolutionary pathway”.

A number of studies were published along these lines in the past ten years, with a very wide spectrum of features of interest. Most of these were focused on lexicon sharing, compositionnality, grammar emergence, or symbol grounding. Very few were concerned with the emergence of segments and phonology. For instance, works about lexicon sharing, that study how a consistent word-meaning map can emerge in a society of agents, often consider the word as an abstract object not linked to articulatory and auditory features (Kaplan, 2000, 2005; Griffiths 2005). Let us mention however three relevant precursor works dealing with the emergence of a phonetic code, generally limited to vowels.

Glotin, Berrah and colleagues (Berrah, 1999) proposed the first studies involving communicating sensori-motor agents. In the interaction paradigm they considered, agents attempt to converge towards a coherent acoustic code through an attraction-repulsion process involving vocalic items. Initially, each agent has a fixed number of items, corresponding to random points in the vocalic triangle. Then, agents interact by pairs, the speaking agent randomly selecting an item in its lexicon and producing it, and the listening agent perceiving the item and comparing it with its own set of prototypes. The closest item in this set is brought closer to the perceived sound, according to an attraction principle, while the other items are moved away, according to a repulsion force. This system, closely related to Lindblom’s Dispersion Theory, predicts the main trends of human vowel systems for a fixed number of vowels. However, it introduces a rather ad-hoc attraction-repulsion principle which is not directly interpretable in terms of pre-existing cognitive function in an evolutionary scenario.

The simulations by de Boer (2000) are more explicit in this respect. De Boer considers a population of agents able to produce and perceive vowels in a reasonably human-like plausible way. Perception is categorical: an acoustic signal is perceived as the nearest category in an agent’s repertoire. Interaction is based on so-called “imitation games”, hence imitation is the driving force in this work (see Ferrari and Fogassi, this volume). Within an imitation game, one agent selects a vowel from its repertoire, and the other agent attempts to imitate it through vowels of its own repertoire. The game may be successful or not, depending on the proximity of the speaker’s target and the listener’s imitation. Depending on this outcome, the participating agents update their repertoire, so that the expected success of subsequent imitation games is increased. Interestingly, the number of items in a given repertoire is not fixed: an agent may borrow the sound from another agent in case of a too large perceptual distance between a target and the agent’s repertoire. There is a good agreement between simulations and data on vowel systems in human languages, including the possibility to predict the preference for five-vowel systems, as in human languages.

This work was further extended by Oudeyer (2005) who attempted to reduce as much as possible the set of cognitive mechanisms necessary for vowel systems emergence. Indeed, de Boer’s work still incorporates rather ad-hoc assumptions about the ability of a pair of agents to decide whether a game is successful or not. Oudeyer proposed a number of simulations in which agents are equipped with sensory-motor maps based on Kohonen’s maps (“Self Organisation”, Kohonen 1981, 1995) and are able to adapt their own map towards the sounds they capture from their partners.

This results in very interesting sensory-motor coupling algorithms, and Oudeyer shows that these algorithms enable to converge towards systems compatible once again with the main trends of vowel systems in human languages. Furthermore, the evolutionary scenario is now rather clear: perceptual resonance drives convergence. Notice that, though Oudeyer claims

that imitation per se is not involved, this is in fact related with something like implicit imitation, in which an agent captures a sound and changes its perceptuo-motor repertoire accordingly.

However, none of these works incorporate a clear answer to the “why communicate?” question. The basis of our answer is the Vocalize-to-Localize framework, providing us with a plausible evolutionary route towards language emergence. For this aim, we propose that agents interact in what we call “deictic games”, allowing an interaction loop between two agents and objects from the environment. The next section first describes the deictic game concept as well as the agents and environment structure, and then proposes a Bayesian modelling of these principles.

4 Modelling

4.1 General principles

4.1.1 Deictic games

According to the Vocalize-to-Localize framework, we model a society of agents able to:

- produce vocalizations (as a first step, we shall consider only one articulatory parameter, then use a realistic model, VLAM (Boë & Maeda, 1997)),
- perceive vocalizations (as a first step, we shall consider only one acoustic parameter, then use a realistic model, with formants),
- focus their joint attention on objects in their environment (two agents in front of the same object identify it in the same way: hence, we posit the existence of a visual categorisation process, that is not yet implemented in the present state of the simulations).

Thus, sensory-motor agents evolve in an environment filled with objects they can identify. Over time, they randomly meet in pairs in front of an object O. They then proceed to what we call a “deictic game”, where one agent has a speaker status, and the other one has a listener status (Figure 4). In order to “show with the voice” this object, the speaking agent proposes a vocalization by achieving a motor gesture M. The gesture is transformed by acoustic and auditory processes into a sensory percept S, perceived by both agents. Deictic games occur in succession over time, each agent randomly taking either a speaker or a listener status.

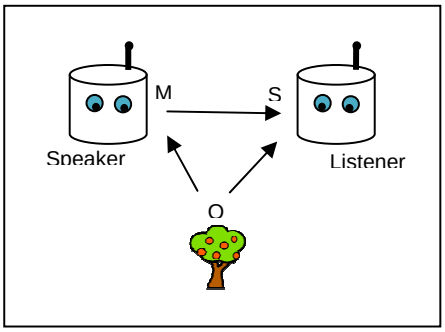


Figure 4 : A deictic games between two agents

4.1.2 Agents knowledge

During each deictic game, the agents can update their knowledge state in the following way. If

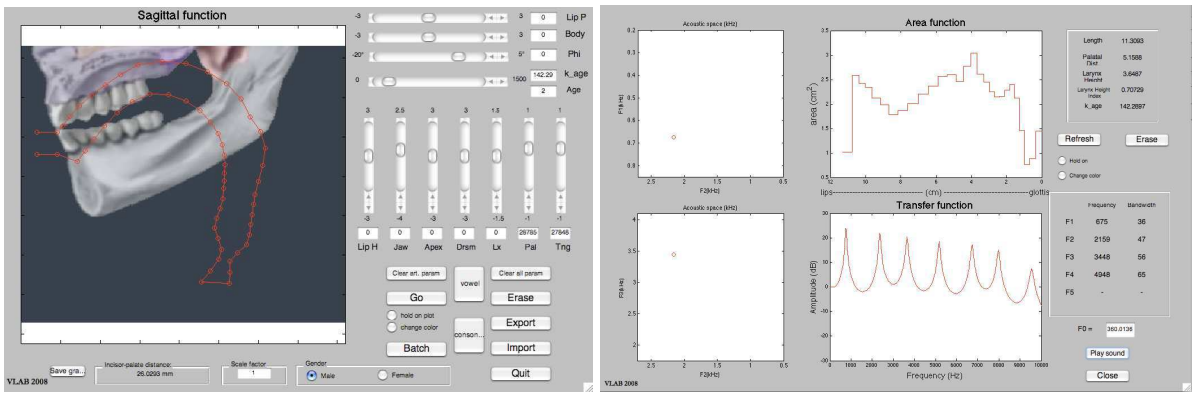
the agent is a speaker, it can update its knowledge about the relation between the considered object O and the motor gesture M associated to it. We call this (O,M) relation the Speaker Model. If it is a listener, it can update the knowledge about the relation between O and the sensory percept S associated to it. We call this (O,S) relation the Listener Model.

Concerning the relation between motor gestures M and sensory percepts S, we assume that the agents possess an internal model able to predict the sound and hence the percept that should be produced by a given motor gesture. This kind of articulatory-to-acoustic efferent copy is known to be part of the human cognitive abilities (Frith, 1992), and proposed to be consistent with the mirror neuron system found in monkeys (Iacoboni, 2005). We call this (M,S) relation the Efference Copy Model. It is supposed to be learnt from previous sensory-motor exploration of the external physical system that we shall describe now.

4.1.3 Sensory-motor systems

Sensory-motor systems establish how the vocal tract shape given by the motor configuration M physically transforms into a sensory percept S, involving acoustical and neural transformations. In the present study, we use two different systems.

As a first step, we will consider a trivial 1-D sensori-motor system with one articulatory parameter and one acoustic parameter, to establish the basic principles of the simulations (for instance, the role of a nonlinearity in the M to S transformation). Then, we shall use a realistic system modelling the vocal tract, the *Variable Linear Articulatory Model* (VLAM) which is a version of the *Speech Maps Interactive Plant* (SMIP, Boë *et al.*, 1995) that integrates a model of the vocal tract growth. The core of the SMIP is an articulatory model (Maeda, 1989) delivering sagittal contour and lips shape from seven input parameters which may be interpreted in terms of phonetic commands, and respectively correspond to the jaw (*J*), the tongue body (*TB*), dorsum (*TD*) and tip (*TT*), the lip protrusion (*LP*) and separation height (*LH*), and the larynx height (*Lx*) (Figure 5). The area function of the vocal tract is estimated from the midsagittal dimensions with a set of coefficients derived from radiographic measurements and tomographic studies. The formants and the transfer function are calculated from the area function, and a sound can be generated from formant frequencies and bandwidths.



a) b)

Figure 5: The VLAM interface. a) Articulatory part: a vocal tract shape is generated from the seven articulatory commands; b) Acoustic part: from the area function (top right), the spectrum of the vocal tract transfer function is computed (bottom right) leading to formant values positioned in the (F1, F2) and (F2, F3) spaces (left) .

In the trivial 1-D system, we consider that the gesture M produces a sensory percept S in a deterministic way. For this aim, we define a *percept* function linking the motor parameter M

and the sensory parameter S. We use this function for transforming the motor gesture produced by the speaking agent into the sensory percept heard by the listening agent during deictic games, possibly adding environmental noise.

In the VLAM system, as we have seen, an articulatory command defines an area function which delivers a sound together with its acoustic formants. These should then be transformed into some adequate perceptual representation (see Serkhane et al., 2005, for a discussion about realistic perceptual and motor representations of speech gestures in an articulatory model). For the need of reducing the complexity of the simulations, we use only three motor parameters, the tongue body (*TB*), dorsum (*TD*) and the lips separation height (LH), everything else being set to a neutral position. This allows to provide a realistic vocalic triangular space in the plan of the first and the second formants. We then consider that the motor gesture M is transformed into a sensory percept S in a probabilistic way, because of the discretization of the motor command space. Thus, for each 3-D motor command region, we compute the mono-modal 2-D distribution of the related sensory percepts in the formants plan. This provides a P(S|M) conditional distribution, that we use for drawing the sensory percept heard by the listening agent given the motor gesture produced by the speaking agent during deictic games. We also test the effect of incorporating environmental noise, by adding random ΔS values drawn from a Gaussian distribution.

For sake of simplification, we assume that the internal Efference Copy Model described in Section 4.1.2 and the external system described here are one and the same model. The hypothesis is hence that the agents are able to perfectly learn the relation between gestures and percepts and that exhaustive learning has already occurred for each agent in a previous phase, not considered here. Thus, in the case of the 1-D trivial system, we assume that agents know the *percept* function. In the case of the realistic VLAM system, we assume that the agents know the P(S|M) conditional distribution.

To summarize:

- The sensory-motor system defining how the motor configuration M physically transforms into a sensory percept S, involving acoustical and neural transformations, is
    - deterministic in the 1D case: S=percept(M),
    - probabilistic in the VLAM case due to motor space discretisation: P(S|M).
- The Efference Copy Model allowing the agent to predict the corresponding percept S of a given motor gesture M corresponds to the knowledge by the agent of:
    - the percept function in the 1D case,
    - the P(S|M) distribution in the VLAM case.
- During communication in a Deictic Game, a Gaussian noise can be added to the sensory-motor system.

4.2 Bayesian modelling

Our modelling is based on the Bayesian Robot Programming paradigm (BRP) (Lebeltel & al., 2004). This method aims at specifying the behaviour of sensory-motor agents in the framework of the Bayesian probability theory. This allows to clearly express both the hypotheses and the lack of knowledge about what is not contained inside the hypotheses set. Operations about knowledge are made by means of Bayesian inference. Moreover, this paradigm provides a clear mathematical framework, usable in order to analyze the outcomes.

4.2.1 Mathematical requirements

BRP is based on a few simple rules from the probability theory that we quickly recall here under.

The product rule [R1] (or Bayes rule) allows to express a joint distribution as a product of elementary distributions:

$$P(A\ B)=P(B).P(A|B)=P(A).P(B|A)$$

The normalization rule [R2] expresses the fact that the probabilities of all possible cases sum to 1:

$$\sum_A P(A) = 1$$

The marginalization rule [R3] is derived from [R1] and [R2] and is also frequently used:

$$\sum_A P(A, B) = P(B)$$

Given these rules and a set of variables V, we can then express all conditional distributions over the variables in V as a function of the joint distribution P(V). Typically, V is separated into three disjoint sets: the searched variables S, the known variables K, and the free variables F. The aim is then to compute the probability distribution over the search variables, knowing the known variables, that is P(S|K). For instance, this could serve a robot in order to answer the question "knowing the value of a few sensory variables (known variables K, given by sensors), what is the probability distribution over my motor variables (search variables S, corresponding to the robot commands)?". In this case, the free variables F could correspond to the unspecified sensory variables, or to internal variables unobserved.

Let us suppose that the robot is able to compute the joint distribution P(V)=P(S K F). It can then answer any question P(S|K) using the following expression (derived from [R1], [R2] and [R3]):

Equation 1:

$$P(S \mid K) = \frac{\sum_F P(S, K, F)}{\sum_{S,F} P(S, K, F)}$$

In this mathematical framework, the BRP method involves two phases. The first one is declarative and describes the model of a cognitive agent. In this phase, the programmer defines the knowledge, relevant for the domain, that the agent refines through parameters learning, in order to compute the joint probability distribution over the variables of interest (typically, motor, sensory and internal variables of the agent). The second one is procedural and describes the agent behaviour. In that phase, the agent uses its knowledge (the joint distribution) to compute any conditional distribution over its variables (for instance, what is the distribution over my motor variables, knowing my sensory variables).

4.2.2 The inference model (declarative phase)

We choose four variables of interest for each agent in the society:

- O_S represents the objects in front of which the agent can be in a speaker status,
- M represents the motor gestures that the agent can produce,
- S represents the sensory percepts that the agent can perceive,

- O_L represents the objects in front of which the agent can be in a listener status (typically the O_L domain is the same as the O_S one).

In order to compute the joint distribution $P(O_S,M, P, O_L)$, we use Bayes rule [R1] to decompose it in a product of simpler terms:

$$P(O_S, M, S, O_L)=P(O_S).P(M| O_S).P(S|M, O_S).P(O_L |S, M, O_S)$$

Then, using general principles described in Section 4.1 and making conditional independence hypotheses, we specify each of these terms. Thus:

- $P(O_S)$ is uniform (considering that objects are equiprobable in the environment)
- $P(M| O_S)$ corresponds to the Speaker Model and so can be learnt by each agent when it is a speaker during deictic games. We consider it as a Gaussian distribution family (one for each O_S value). Learnt parameters are means μ_{O_S} and variances V_{O_S}, re-estimated after each deictic game.
- $P(S|M,O_S)$ is simplified into $P(S|M)$, by considering that S is entirely defined by the knowledge of M. This distribution corresponds to the Efference Copy Model and so is supposed to be known by the agents. As a first step, in the 1-D simplified sensori-motor system (section 4.1.3), we consider it as deterministic, and hence defined by a Dirac distribution: $P(S|M)=1$ if $S=percept(M)$, 0 otherwise. Then, using the realistic VLAM sensory-motor system, we shall consider it as a fixed Gaussian distribution family, previously learnt by discretised motor space exploration.
- $P(O_L|S,M,O_S)$ simplifies in $p(O_L|S)$, considering that the listener estimates the object entirely from S, as M and O_S are not directly accessible to the listener. This distribution corresponds to the Listener Model and so can be learnt by the agent when it is a listener during deictic games. Using the Bayes rule [R1], and considering $P(O_L)$ as uniform, we have: $P(O_L \mid S) = \dfrac{P(S \mid O_L)}{\sum_{O_L} P(S \mid O_L)}$.

  $P(S|O_L)$ is considered as a Gaussian distribution family (one for each O_L value). Learnt parameters are means μ_{O_L} and variances V_{O_L}, re-estimated after each deictic game.

Thus, we obtain the following simplified expression of the joint distribution, schematized on Figure 6:

Equation 2:

$$P(O_S, M, S, O_L) \propto P(M \mid O_S).P(S \mid M).P(O_L \mid S)$$

where $P(O_L \mid S) = \dfrac{P(S \mid O_L)}{\sum_{O_L} P(S \mid O_L)}$

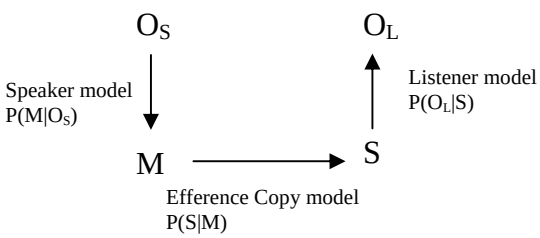


Figure 6: Joint distribution structure of an agent.

Given this joint distribution, each agent is able to compute any conditional distribution over the four involved variables, using Bayesian inference. A conditional distribution is called a "question" to the model. Note that the two terms, $P(O_S)$ and $P(S|M)$, are constant over time, and the two others, $p(M|O_S)$, $p(O_L|S)$ are learnt by the agents. Thus the joint distribution evolves during deictic games. The following subsection exposes three distinct behaviours that we elaborated for the agents, depending on how the speaker selects a motor gesture in front of an object, that is depending on which question it asks to its joint distribution.

4.2.3 The interaction behaviours (procedural phase)

Here, we expose several behaviours for the agent, which are, as we shall see later, more or less likely to lead to a common speech code between agents. A behaviour is defined as the way the speaker selects a motor gesture in front of an object during deictic games. In probabilistic terms, this corresponds to the distribution according to which it selects the variable M, that is the question it asks to the model described previously. We present three behaviours increasing in complexity: the reflex behaviour which takes into consideration only the Speaker Model, the communicative behaviour which considers only the Listener Model, and the hybrid behaviour which takes into account both the Speaker and the Listener Models.

4.2.3.1 Reflex behaviour

In this first behaviour, the speaker takes into consideration only its Speaker Model. Therefore, in front of an object o_i, it selects a motor gesture M according to the distribution $P(M|O_S=o_i)$. Thus, the agent simply selects motor gestures that it has already produced in front of the corresponding object, in a kind of "reflex" mood, without taking into account the listener's expectations. We shall see that taking into account only the speaker's interests cannot lead to the emergence of a common speech code between the agents.

4.2.3.2 Communicative behaviour

This behaviour consists, for the speaker, of attending as much as possible to the listener's expectations, by taking into consideration the Listener Model. Actually, in a deictic game, the speaker selects a motor gesture which would have allowed himself to *infer* the correct object[1]. Therefore, in front of an object oi, the speaker seeks to maximize the probability $P(O_L=o_i|M)$ over M. In fact, according to Equation 1 and Equation 2 we have:

$$
\begin{aligned}
|P(O_L = o_i \mid M) &\propto \frac{\sum_{O_S,S} P(M \mid O_S).P(S \mid M).P(O_L = o_i \mid S)}{\sum_{O_S,S,O_L} P(M \mid O_S).P(S \mid M).P(O_L \mid S)} \\
&\propto \frac{\sum_{O_S} P(M \mid O_S).\sum_S P(S \mid M).P(O_L = o_i \mid S)}{\sum_{O_S} P(M \mid O_S).\sum_S (P(S \mid M).\sum_{O_L} P(O_L \mid S))} \\
&\propto \sum_S P(S \mid M).P(O_L = o_i \mid S)
\end{aligned}
$$

Thus, the speaker selects a motor gesture producing a percept which should have the best communicative value. For example, in the 1D case where P(S|M) is a Dirac distribution, we have $\sum_S P(S \mid M).P(O_L = o_i \mid S) = P(O_L = o_i \mid S = percept(M))$

[1] It is worth noting that it is exactly what achieves the deictic function: pointing consists of producing a hand gesture which produces a visual percept (by following the finger direction) which corresponds to the pointed object.

4.2.3.3 Hybrid behaviour

This behaviour seeks to maximize both the motor and sensory qualities of the speaker's gesture by satisfying both the Speaker and the Listener Models. Thus, the speaker selects a motor gesture which it has already often selected for the object, and which in the same time would have allowed itself as a listener to easily infer it. Hence, in front of an object oi, the question asked to the model is $P(M|O_S= o_i \;\; O_L= o_i)$, which can be decomposed into:

$$\begin{aligned} P(M \mid O_S = o_i, O_L = o_i) &\propto \frac{\sum_S P(M \mid O_S).P(S \mid M).P(O_L = o_i \mid S)}{\sum_{M,S} P(M \mid O_S).P(S \mid M).P(O_L \mid S)} \\ &\propto P(M \mid O_S).\sum_S P(S \mid M).P(O_L = o_i \mid S) \end{aligned}$$

Therefore, it can be seen that the speaker selects a motor gesture according to a distribution which is the product of those of the two previous behaviours. This behaviour could thus model the relation between production and perception in speech, where a gesture is selected both for its motor and sensory qualities, as in PACT (Section 2.1.3).

Interestingly, the question asked in the Hybrid behaviour, $P(M|O_S=o_i \; O_L=o_i)$, allows to unify the three behaviours into a coherent framework, by disabling either the Speaker or the Listener Model. Disabling a model consists in setting it to a uniform distribution. Thus:

The Reflex behaviour corresponds to the question $P(M|O_S=o_i \; O_L=o_i)$ where the Listener Model is disabled, that is $P(O_L|S)$ is considered as uniform by the speaker. In this case:

$$\begin{aligned} P(M \mid O_S = o_i, O_L = o_i) &\propto \sum_S P(M \mid O_S).P(S \mid M) \\ &\propto P(M \mid O_S) \end{aligned}$$

The Communicative behaviour corresponds to the question $P(M|O_S=o_i \; O_L=o_i)$ where the Speaker Model is disabled, that is $P(M|O_S)$ is considered as uniform by the speaker. In this case :

$$P(M \mid O_S = o_i, O_L = o_i) \propto \sum_S P(S \mid M).P(O_L = o_i \mid S)$$

In the Hybrid behaviour, no model is disabled.

The next section describes the functioning of these behaviours, and their link with the Morphogenesis Theories introduced in Section 2.

5 Results

5.1 Technical details

Each simulation is run for a given number of agents N_A, a given number of objects N_O, a given behaviour B (either Reflex, Communicative or Hybrid), a given sensory-motor system SM (either 1D or VLAM, see 4.1.3) and during a given number of deictic games N_G. For each deictic game, we uniformly draw one speaker agent, one listener agent (different from the speaker) and one object. Then, the speaker agent draws a motor gesture M in a domain D_M according to the behaviour B (see 4.2.3 above for the corresponding distribution of each behaviour). This gesture is transformed into a sensory percept S in a domain D_S according to the sensory-motor system SM, that is according to a *percept* function in a deterministic manner in the 1D case, where D_M and D_S are unidimensional or according to a P(S|M) distribution in a probabilistic manner in the VLAM case, where D_M is 3-dimensional (Body, Drsm and LipH, see 4.1.3) and D_S is 2-dimensional (first and second formants). A gaussian

noise with a standard deviation SD is added to each S dimension, expressed as a percentage of D_S range.

In the 1-D case, we define the percept function as a sigmoid (considering that M_{min}= -M_{max} and S_{min}= -S_{max} for simplification):

$$percept(M) = \frac{S_{max} - S_{min}}{2.\arctan\left(NL.\frac{M_{max}-M_{min}}{2}\right)}.\arctan(NL.(M-D))$$

where NL is a non-linearity coefficient (when NL approaches 0, percept can be considered as linear; it increasingly draws away from linearity when NL increases) and D is the position of the inflection point. The aim will be to analyse the effect of a non-linearity on the common speech code, with regard to the Quantal Theory. Figure 7 proposes four percept functions for different values of NL and D.

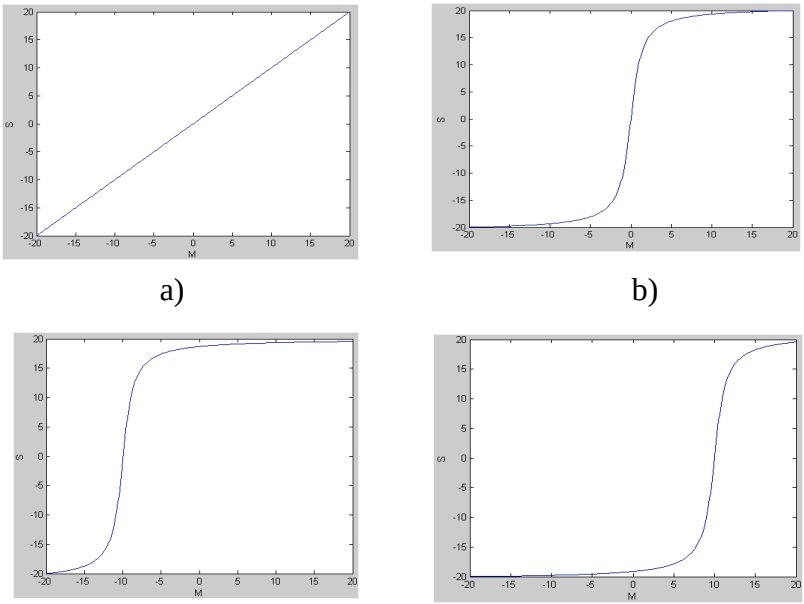


Figure 7: percept function for a) NL=10^{-5}, D=0 (linear case); b) NL=1, D=0; c) NL=1, D=-10; d) NL=1, D=10.

At the end of each deictic game, both the speaker and listener agents update their knowledge, that is the $P(M|O_S=o_i)$ gaussian distribution for the speaker and the $P(O_L=o_i)$ gaussian distribution for the listener, o_i being the object involved in the deictic game. Initially, each distribution is set with the means and variances calculated from a uniformly drawn sample of N_P points (generally N_P=1000), each with a weight set to 1. Then, during the deictic games, distributions are updated by adding a new point in the sample with a weight corresponding to a percentage F of the total weight of the sample. Thus, all the values from the beginning of the simulation are taken into account with an increasing weight for the more recent ones. F is called the forgetting coefficient (generally set to 0.1) because the higher it is, the lower the influence of the oldest values.

During the simulation, we compute what we call the understanding rate in the society. This corresponds to the percentage of successful deictic games during the N_U last deictic games (generally N_U=1000). A successful deictic game corresponds to a game in which the listener

was able to correctly infer the involved object just from the sensory percept *s* provided by the speaker, using the question $P(O_L|S=s)$. We display the understanding rate during a simulation in order to evaluate the ability of a behaviour to lead to a common speech code.

To summarize, for each simulation we shall provide a set of parameters which define it:

- N_A: the number of agents,
- N_O: the number of objects,
- N_G: the number of deictic games (thus corresponding to the duration of the simulation),
- B: the behaviour of the agents (either Reflex, Communicative or Hybrid),
- D_M, D_S: the domain of M and S, respectively.
- SM: the sensory-motor system, transforming the motor gesture M emitted by the speaker into a sensory percept S (either through a deterministic "percept" function in the 1D model, or through a probabilistic P(S|M) distribution in the VLAM model, see 4.1.3),
- NL, D: the non-linearity coefficient and the position of the inflection point of the percept function, provided only if SM=1D (in the VLAM case, P(S|M) is provided by the VLAM model as explained in 4.1.3),
- SD: the standard deviation of the gaussian environmental noise, added to each S dimension (expressed as a percentage of Ds range),
- F: the forgetting coefficient, defining the weight of new values compared to old values in the updating of the P(M|O) and P(P|O) distributions,
- N_U: the number of the last deictic games used for computing the understanding rate.

5.2 Simulations

Here we expose and analyse the results for the three behaviours described previously. The simulation window in which we observe these results is displayed on Figure x. In the upper part, there are as many windows as agents in the simulation. In each of these windows, there are as many gaussian curves as objects in the simulation. Thus, each gaussian curve corresponds either to the $P(M|O_S=o_i)$ or to the $P(S|O_L=o_i)$ distribution for a given o_i at the end of the simulation, according to what we want to observe (specified on the figures). The lower part of the simulation window corresponds to the evolution with time of the understanding rate in the society of agents, as it is defined previously.

We consider that a common speech code emerges when the $P(S|O_L)$ distributions are both different and well separated from one object to the other, and similar from one agent to another. A consequence is a high value of the understanding rate in the society, which ensures that one agent is able to correctly infer the object given a sensory percept provided by another agent.

5.2.1 Results for the 1-D sensory-motor system

These simulations are run with SM=1D, N_A=4, N_G=150 000, D_M=D_S=[-20,20], SD=0.01, F=0.1, N_U=1000. The other parameters are provided in the figures legends.

5.2.1.1 Reflex behaviour

Results about the Reflex behaviour are displayed Figure 8.

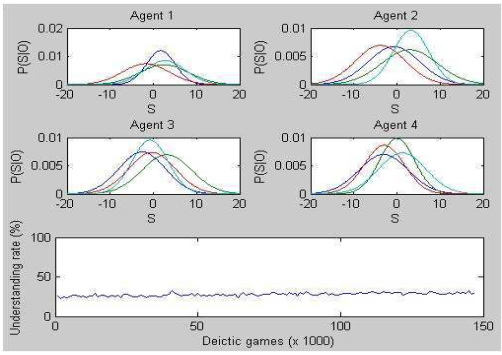


Figure 8: Simulation results for B=Reflex, No=4, NL=10^{-5} (linear *Percept* function), D=0.

We observe that the $P(S|O_L)$ distributions are neither separated between objects nor coherent between agents. Indeed, the agents draw gestures that are already often drawn for a given object (by drawing according to the distribution $P(M|O_S=o_i)$ for a given o_i, see 4.2.3.1), without taking into account the listener's expectations. Thus, distributions stay around their initial values and deictic games cannot lead to the emergence of a common speech code between the agents. In consequence, we observe that the understanding rate in the society stays around chance level, which is 25% for four objects.

5.2.1.2 Communicative behaviour

Results about the Communicative behaviour with a linear *Percept* function are displayed on Figure 9.

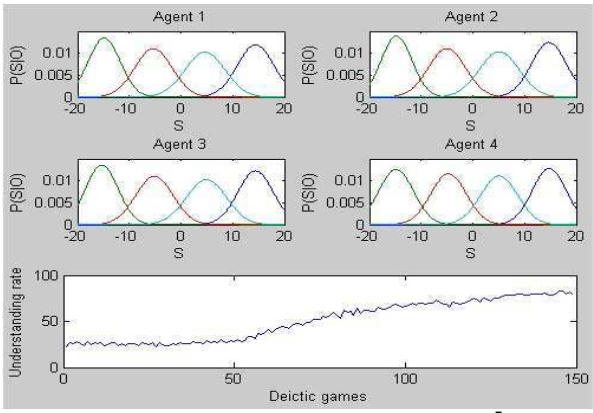


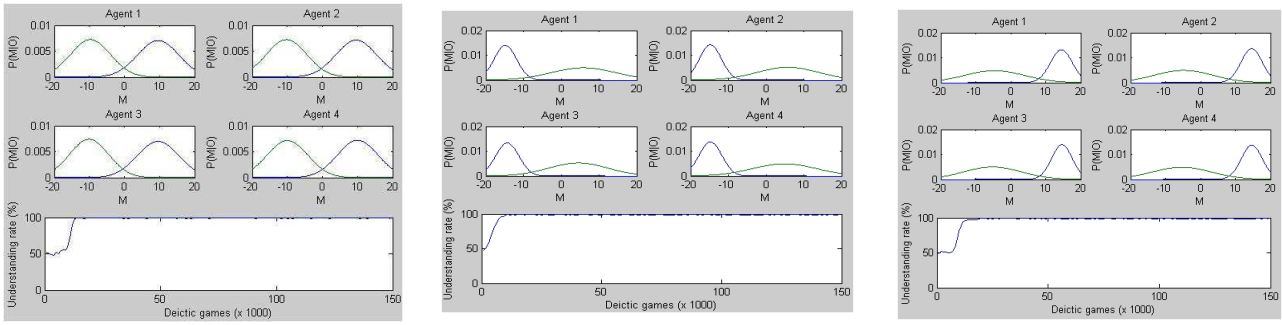
Figure 9: Simulation results for B=Communicative, No=4, NL=10^{-5} (linear *Percept* function), D=0.

We observe the emergence of a common speech code between agents. Indeed, during the simulation the agents converge towards similar $P(S|O_L)$ distributions, very different from one object to another. In consequence, the understanding rate in the society reaches around 80%. It does not reach 100% because, as we observe, there is a bit of overlapping of $P(S|O_L)$ distributions.

This fits quite well with predictions of the Dispersion Theory, observing that the means of $P(S|O_L)$ distributions seem to be scattered rather evenly with a trend of maximal dispersion between percepts. Actually, it seems possible to show that the Communicative Behaviour

should converge towards a state in which $\sum_{O_L} P(S \mid O_L)$ approximates a uniform distribution, which results in a principle of maximal dispersion. Indeed, let us consider a simple case where S=M (*Percept* is the identity function). The selected motor gestures then maximize $P(O_L \mid M) = \frac{P(M \mid O_L)}{\sum_{O_L} P(M \mid O_L)}$. After convergence, the system is stable if the drawing of M values according to this distribution does not change dramatically the $P(M|O_L)$ distributions. This results in having similar values for $P(M|O_L)$ and $P(O_L|M)$, that is when the denominator $\sum_{O_L} P(M \mid O_L)$ approximates a uniform distribution. The solution is to place the $P(M|O_L)$ gaussian distributions uniformly in the available space, as realized by the simulation in Figure 9 (though an analytical solving of this optimization problem is not trivial). This shows that the Bayesian framework nicely provides a mathematically link between hypotheses from Origins Theories and optimization problems from Morphogenesis Theories.

Considering the Quantal Theory, let us analyze the effect of a non-linearity in the *Percept* function transforming motor gestures M into sensory percepts S. Figure 10 displays how the position of the non-linearity shapes the speech code between agents (Figure 10a, Figure 10b and Figure 10c correspond to percept function of figures Figure 7b, Figure 7c and Figure 7d, respectively). Indeed, we observe that shifting the position of the non-linearity (by changing D) results in shifting accordingly the boundary between gestures associated with objects, thus producing categories driven by the nonlinearity positions, as predicted by the Quantal Theory. Moreover, the non-linearity allows to create a speech code with better quality (the understanding rate reaches 100%).



a) b) c)

Figure 10: Simulation results for B=Communicative, No=2, SD=0.1, NL=1 (nonlinear *Percept* function) and three positions of the nonlinearity, a) D=0; b) D=-10; c) D=10. Observation of the $P(M|O_L)$ for each agent.

5.2.1.3 Hybrid behaviour

Results about the Hybrid behaviour are displayed on Figure 11 for a linear *Percept* function:

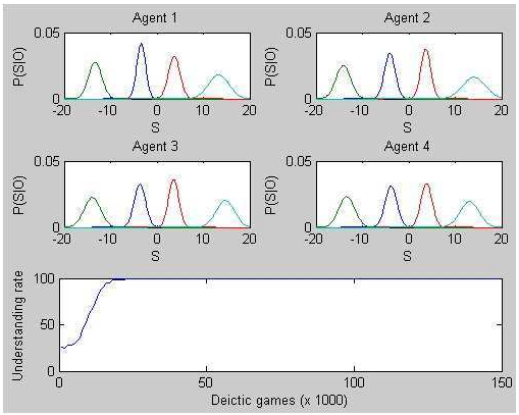


Figure 11: Simulation results for B=Hybrid, No=4, NL=10^{-5} (linear *Percept* function), D=0.

We observe the emergence of a common speech code between agents. With respect to the Communicative behaviour, adding the $P(M|O_S=o_i)$ term in the distribution setting the behaviour leads to reducing the variance of the selected sensory percepts in a conservative manner. This results in a faster convergence and a better code quality with 100% understanding rates. Moreover, this behaviour keeps the good properties of the Communicative Behaviour with respect to both the Dispersion Theory (sensory percepts for each object are dispersed, see Figure 11) and with the Quantal Theory.

5.2.1.4 Conclusion for the 1-D sensory-motor results

We suggest that the hybrid behaviour is the most attractive one in terms of both performance and theoretical basis. On one hand, it provides the fastest convergence and the highest understanding rate. On the other hand, the question to the joint distribution used for motor gesture selection, $P(M|O_S=o_i,O_L=o_i)=P(M|O_S=o_i).P(O_L=o_i|S=percept(M))$, provides a statistical implementation of a mechanism associating motor gestures conservation and sensory percepts dispersion. This is in line with the Perception for Action Control Theory developed in the last years for which gestures are selected for both their intrinsic motor and sensory properties (see Schwartz et al., 2007).

5.2.2 Results for the VLAM sensory-motor system

Starting from this conclusion, we "embodied" the hybrid behaviour into a realistic sensory-motor system: VLAM, of which the motor and sensory variables, as well as the P(S|M) distribution definition were described previously. The motor space is discretised into 1000 sections (10 for each dimension: Tongue Body, Tongue Dorsum and Lips Height). For each section, the P(S|M) distribution is provided by uniformly drawing 100 points in the section, obtaining the associate percepts thanks to the VLAM model and computing the corresponding 2-D gaussian distribution in the formant space.

We ran a simulation with the following parameters: N_A=4, B=Hybrid, SM=VLAM, D_M=[-3,3][3] ([-3,3] for each parameter, according to the VLAM convention, see Boë & Maeda, 1997)[2], D_S=[2,7]x[6,14] (corresponding to the maximal vocalic formant space for the given motor parameters, in Barks), F=0.05, N_G=20000, N_U=100.

[2] Actually, there are some configurations of the motor space which correspond to closed configurations, which are not vowels and for which formants cannot be computed in VLAM. Therefore we added a boolean variable V and a P(V|M) term in the joint probability distribution in order to represent the fact that a motor configuration must correspond to a vowel.

The noise added to each sensory dimension is drawn from a 2D Gaussian distribution with a standard deviation set to 0.4 for F1 and 1,4 for F2 (covariances are set to zero). This roughly corresponds to a 0.3 ratio between F1 and F2 noise, which is conform to the estimated weight ratio provided by Schwartz & al. 1997.

We then observe the distribution $P(S|O_L)$ for each agent, that is the distribution of sensory percepts produced by the agent for each object. We represent it by a set of dispersion ellipses (one for each object) with 1.5 standard deviation.

For an environment with three object (N_O=3), we observe that the agents select sensory percepts at the vertex of the vocalic triangle, which corresponds to the three mostly used vowels in the world languages /a, i, u/ (Figure 12)

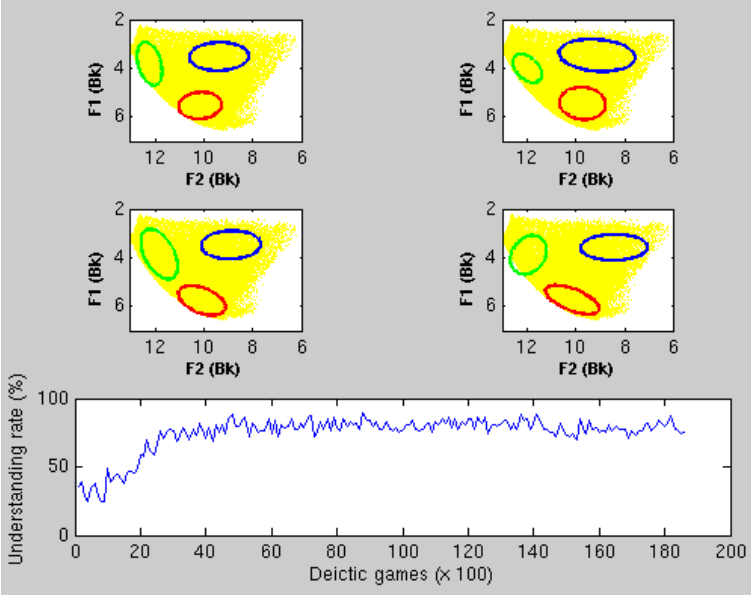


Figure 12: Results in the realistic VLAM sensory-motor system with 4 agents and 3 objects.

For an environment with five objects (N_O=5), we observe that the agents select sensory percepts which correspond to the most used vocalic system in world language /i, e, a, o, u/ (Figure 13).

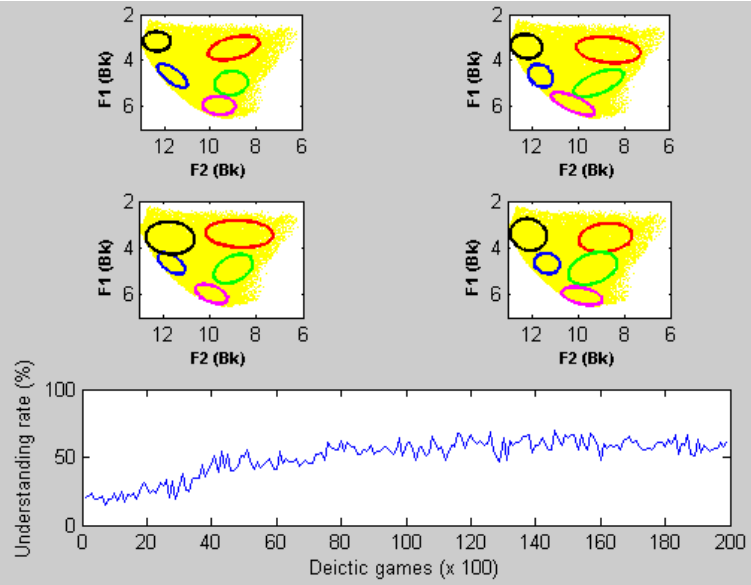


Figure 13: Results in the realistic VLAM sensory-motor system with 4 agents and 5 objects.

6 Conclusions and perspectives

In this paper, we show how principles of Morphogenesis Theories such as the dispersion between selected sensory percepts, the quantal aspect of speech, and the role of both motor and sensory knowledge in speech production can emerge from the modelling of prelinguistic functions provided by Origins Theory such as deixis. For this aim, we define and implement an integrating computational framework based on multi-agent simulations in order to link various works concerning the origins and the universals of human language.

The next step in this work will consist in going from static vocalic configurations of the vocal tract to more complex sequences. This will be achieved in connection with the Frame-then-Content Theory developed by MacNeilage and Davis (2000; MacNeilage, 1998), providing another ingredient inside Origins theories: the role of jaw cycles that would be inherited from mastication, and involved as a bootstrap for controlling modulations of vocalisations for orofacial communication. For implementing the Frame-then-Content Theory in our computational framework, we shall use the Jaw motor parameter of VLAM in order to induce a mandibular cycle in the agents vocalisations. We hope to show that acoustic/auditory nonlinearities shape the simple jaw rhythmic activity in a quantal pattern, achieving the generation of alternations of vowels and consonants in a simple way both developmentally plausible and functionally efficient. Then we predict that bilabials, dentals and velars (e.g. [b d g]) provide an optimal system in terms of auditory dispersion, provided that they are embedded in this developmental framework, pharyngeals, though auditorily salient, being eliminated by their high jaw configuration incompatible with the Frame-Content scenario (Abry, 2003; Schwartz & Boë, 2007).

In a broader perspective it must be acknowledged that the deictic function cannot be considered as more than a *bootstrap* for the emergence of a communicative system. Other ingredients could be incorporated in a further step, e.g. pantomime or other kinds of referent orofacial or bracchiomanual gestures (Arbib, 2004) possibly extending deixis towards what could be conceived as a “super-deictic” ability to evoke objects, agents and actions through gestures in various modalities.

7 References

Abry, C., Boë, L.J., & Schwartz, J.L. (1989). Plateaus, catastrophes and the structuring of vowel systems. *Journal of Phonetics*, 17, 47-54.

Abry, C. (2003). [b]-[d]-[g] as a universal triangle as acoustically optimal as [i]-[a]-[u]. *Proc. XVth International Congress of Phonetic Sciences (pp. 727–730)*. Barcelona, Spain.

Abry, C., Vilain, A., & Schwartz, J.-L (2004). Vocalize to localize? A call for better crosstalk between auditory and visual communication systems researchers. *Interaction Studies*, 5, 313-325.

Arbib, M. A. (2004). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences*, 28(2):105--124.

Berrah, A-R., Laboissière, R. (1999). SPECIES: An Evolutionary Model for the Emergence of Phonetic Structures in an Artificial Society of Speech Agents. In D. Floreano and J. Nicoud and F. Mondada, editors, *ECAL99* (pp. 674—678). Berlin: Springer-Verlag.

Boë, L. J., Gabioud, B., & Perrier, P. (1995). Speech Maps Interactive Plant « SMIP ». *Proc. XIIIth International Congress of Phonetic Sciences* (pp. 426–429). Stockholm, Sweden.

Boë, L.J., & Maeda, S. (1997). Modélisation de la croissance du conduit vocal. Espace vocalique des nouveaux-nés et des adultes. *Journées d'Etudes Linguistiques : La Voyelle dans Tous ses Etats*, 98-105.

Boë, L.-J., Vallée, N., Badin, P., Schwartz, J.-L. & Abry, C. (2002). Tendencies in phonological structures: the influence of substance on form. *Bulletin de la Communication Parlée*, 5, 35-55.

de Boer, B. (2000) Self-organization in vowel systems. *Journal of Phonetics*, 28, 441--465.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3–28.

Frith, C. (1992). *The cognitive neuropsychology of schizophrenia*. East Sussex, England: Lawrence A Erlbaum Associates.

Goldin-Meadow, S., & Butcher, C. (2003). Pointing toward two-word speech in young children. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 85-107). Lawrence Erlbaum Associates.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Science*, 4, 131–138.

Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8, 393-402.

Iacoboni, M. (2005). Understanding others: imitation, language, empathy. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From Cognitive Neuroscience to Social Science* (Vol. 1, pp. 77-99). Cambridge, MA: MIT Press.

Kaplan, F. (2000). Semiotic schemata: Selection units for linguistic cultural evolution. In Bedau, M and McCaskill, J. and Packard, N. and Rasmussen, S., editor, *Proceedings of Artificial Life VII* (pp. 372-381). Cambridge, MA. The MIT Press.

Kaplan, F. (2005). Simple models of distributed co-ordination. *Connection Science*. 17, 249-270.

Kohonen, T. (1981). Automatic formation of topological maps of patterns in a self-organizing system. In Oja, E. and Simula, O., editors, *Proceedings of 2SCIA, Scand. Conference on Image Analysis*, pages 214-220, Helsinki, Finland. Suomen Hahmontunnistustutkimuksen Seura r.y.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Berlin, Heidelberg.

Lebeltel, O., Bessière, P., Diard, J., & Mazer, E. (2004). Bayesian robot programming. *Autonomous Robots*, 16, 49-79.

Liberman, A.M, & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1–36.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Science*, 4, 187–196.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839–862.

Lindblom, B. (1984). Can the models of evolutionary biology be applied to phonetic problems? *Proc. 10th International Congress of Phonetic Sciences*, 67-81.

Lindblom, B. (1986). Phonetic universals in vowel systems. In *Experimental Phonology* (J.J. Ohala and J.J. Jaeger, eds.), pp. 13-44. New-York: Academic Press.

Lindblom, B. (1990). On the notion of « possible speech sound ». *Journal of Phonetics*. 18, 135–152.

Lœvenbruck, H., & Perrier, P. (1997). Motor control information recovering from the dynamics with the EP hypothesis. *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes – Greece, 4, 2035–2038.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-511.

MacNeilage, P. F. & Davis, B. L. (2000) On the origin of internal structure of word forms. *Science*, 288, 527--531.

Maeda, S. (1989). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and modelling* (pp. 131–149). Dordrecht: Kluwer.

Massaro, D.W. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry*. London: Laurence Erlbaum Associates.

Nearey, T.M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of Americ*a, 101, 3241–3254.

Oudeyer, P-Y. (2005) The Self-Organization of Speech Sounds. *Journal of Theoretical Biology*, 233(3), 435--449.

Serkhane, J.E., Schwartz, J.L., Boë, L.J., & Bessière, P. (2005). Building a talking baby robot: a contribution to the study of speech acquisition and evolution. *Interaction Studies, 6*, 253-286.

Schwartz, J.L., Boë,L.J., Vallée, N., & Abry, C. (1997). The Dispersion-Focalization theory of vowel systems. *Journal of Phonetics*, 25, 255–286.

Schwartz, J.L., Abry, C., Boë, L.J., & Cathiard, M.-A. (2002). Phonology in a Theory of Perception-for-Action-Control. In J. Durand and B. Laks (eds.) *Phonetics, Phonology, and Cognition* (pp. 254–280). Oxford: Oxford University Press.

Schwartz, J.L., & Boë, L.J. (2007). Grounding plosive place features in perceptuo-motor substance. Colloque "*International Conference on Features*", Paris.

Schwartz, J.L., Boë, L.J., & Abry, C. (2007). Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a Perception-for-Action-Control Theory (PACT). In M.J. Solé, P. Beddor & M. Ohala (eds.) *Experimental Approaches to Phonology* (pp. 104-124). Oxford: Oxford University Press.

Schwartz, J.L. (2008a). Eléments pour une morphogenèse des unités du langage. Colloque « *Systèmes complexes en Sciences Humaine set Sociales* », Cerisy (in press).

Schwartz, J.L. (2008b). Filling the perceptuo-motor gap. Laboratory Phonology, 10 (to appear).

Skipper, J.I., Van Wassenhove, V., Nusbaum, H.C. & Small, S.L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17, 2387-2399.

Steels, L. (1996). Emergent Adaptive Lexicons. In Maes, P. and Mataric, M. and Meyer, J.-A. and Pollack, J. and Wilson, S.W., editor, *From Animals to Animats 4: Proceedings of the*

Fourth International Conference On Simulation of Adaptive Behavior (pp. 562-567). Cambridge, MA, The MIT Press.

Steels, L. (1997) The Synthetic Modeling of Language Origins. *Evolution of Communication Journal*, 1, 1-34.

Steels, L. (2006) How to do Experiments in Artificial Language Evolution and Why. In Cangelosi, A., Smith A. and Smith K., editor, *Proceedings of the 6th International Conference on the Evolution of Language*, London. World Scientific Publishing.

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. Davis Jr. and P. B .Denes (eds.), *Human Communication: A Unified View* (pp. 51-66). New-York: Mc Graw-Hill.

Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17, 3–45.

Volterra, V., Caselli, M. C., Capirci, O., & Pizzuto, E. (2005). Gesture and the emergence and development of language. In M. Tomasello and D. Slobin (Eds.), *Beyond nature-nurture-Essays in honor of Elizabeth Bates* (pp. 3–40). Mahwah, N. J.: Lawrence Erlbaum Associates.