Emergence of a Language through Deictic Games within a Society of Sensory-Motor Agents in Interaction

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

1: GIPSA-Lab ICP, UMR 5216; 2: LPNC, UMR 5105 ; 3: LIG, UMR 5217
CNRS – Grenoble University
1: FirstName.LastName@gipsa-lab.inpg.fr ; 2: Julien.Diard@upmf-grenoble.fr ;
3: Pierre.Bessiere@imag.fr

Abstract

Starting from language origins theories, which connect prelinguistic primate abilities such as deixis, to modern linguistic systems, we seek to let emerge universals of human languages such as dispersion principles and the quantal aspect of speech. For this aim, we model a society of Bayesian sensory-motor agents, which interact and evolve in an environment filled with objects they can identify. We show how these agents may converge on a common phonological code for communication. This enables us to validate our assumptions and suggest probabilistic models of language origins theories.

1 Introduction

Since the 70's and Lindblom's proposal to "derive language from non-language", phoneticians have developed a panel of "substance-based" theories. The starting point is Lindblom's Dispersion Theory [1] and Stevens's Quantal Theory [2], 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 universals of world's languages from morphologic constraints arising from perceptual (auditory and perhaps visual) and motor (articulatory and cognitive) properties: 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. In the Vocalize-to-Localize framework that we adopt in the present work [3], 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 kind 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 ask 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 sounds, 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 achieve the pre-linguistic deictic function. For this aim we model a society of bayesian agents provided with a sensory-motor apparatus and evolving in an environment filled with objects they can identify. We then describe three different behaviors describing the way in which speakers seek to produce motor gestures in front of an object and study the emergence of a common speech code as well as its properties related to Morphogenesis Theories.

Section 2 exposes the model and describes the three behaviors. Then Section 3 studies these behaviors both in a simplified and in a more realistic model of the sensory-motor apparatus. Finally Section 4 concludes and provides some perspectives.

2 Modeling

According to the Vocalize-to-Localize framework [1], we model a society of agents able to produce vocalizations, perceive vocalizations and focus their joint attention on objects in their environment

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. 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 the listening agents (Figure 1). Deictic games occur in succession over time, each agent randomly taking either a speaker or a listener status.

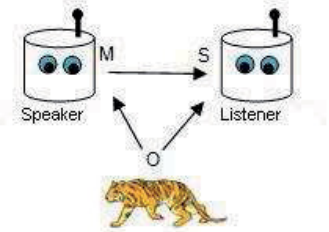

Figure 1: *A deictic game between two agents*

In the Bayesian Robots Programming framework [4], agents are defined by their knowledge of a probability distribution connecting all their variables, which are, in the present study, the object O they are in front of, together with their motor state M and their perceptual state S. We consider three sub-models enabling to describe the possible behaviors of the agents. Firsly, the relation between the considered object O and the motor gesture M associated to it is called the Speaker Model. Secondly, the relation between O and the sensory percept S associated to it is called the Listener Model. Thirdly, 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, thanks to a S=percept(M) function. This articulatory-to-acoustic efferent copy relating M and S is known to be part of the human cognitive abilities [5], and proposed to be consistent with the mirror neuron system found in monkeys [6].

If we call O_S the objects in front of which agents can be in speaker status and O_L the objects in front of which agents can be in listener status, the probability distribution $P(O_S, M, S, O_S)$ characterizing the behavior of the agent can be simplified, under adequate simplifications [7], as a product of the three sub-models:

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

This product defines the way the communication process can be implemented in each agent's knowledge state. From this state, we define three possible behaviors corresponding to different strategies of motor gesture selection in front of an object o_i.

In the *Reflex behavior*, the speaking agent takes into consideration only its Speaker Model, by drawing a motor gesture M according to the distribution $P(M| O_S = o_i)$. In this behavior, the agent processes in a kind of reflex way just taking into account the object and not the listening process that could occur in the communication partner. We shall see that this behavior cannot lead to the emergence of a common speech code between the agents.

In the *Communicative behavior*, the speaking agent takes into consideration only its Listener Model, by drawing a motor gesture M in order to maximize the probability:

$$P(O_L = o_i \mid M) = P(O_L = o_i \mid S = \text{percept}(M)).$$

In this behavior, the agent thus selects a gesture providing a percept which would have allowed himself to infer the correct object. We shall see that this behavior leads to the emergence of a common phonetic code between the agents.

In the *Hybrid behavior*, the speaking agent takes into consideration both its Speaker and its Listener Model, by drawing a motor gesture M according to the distribution:

$$P(M| O_S = o_i, O_L = o_i) = P(M| O_S = o_i)P(O_L = o_i |S=\text{percept}(M))$$

In this behavior, the agent takes into account both the object it perceives as a speaker, and the one it should perceive as a listener, which results in considering the product of the Reflex and Communicative behaviors. Interestingly, this incorporates the relation between production and perception in speech, where a gesture is selected both for its motor and sensory qualities.

3 Results

3.1 A simplified sensory-motor model

As a first step, we consider a simplified model of the sensory-motor apparatus, where M and S are simply one-dimensional bounded variables. This allows to analyze the basic properties of the three behaviors in relation with Morphogenesis Theories.

Figure 2 displays results obtained with the Reflex behavior in a society of four agents and an environment containing four objects. In a first stage, the percept function transforming motor gestures into sensory percepts is the identity function (S=M). The four upper windows represent the state of the four agents at the end of the simulation. In each window, the four Gaussian curves correspond to the distribution $P(S|\ O_L)$ for each object. The lower window provides the evolution of the society with time, that is the percentage of successful games during the 1000 last ones. A game is called successful when the listener was able to correctly infer the involved object from the stimulus *s* provided by the speaker, using the question $P(O_L\ |S=s)$. It appears that the reflex behavior provides no dispersion: all Gaussian curves stay quite close to each other, indicating a lack of differentiation of gestures for different objects. In consequence, the understanding rate stays close to random (25% for four objects).

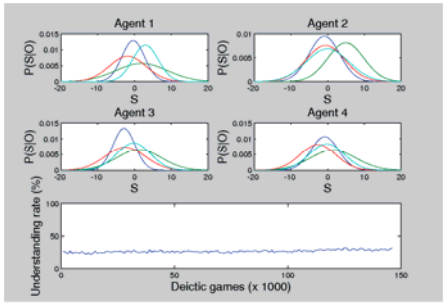

Figure 2: *Reflex behavior results in a society of four agents and an environment containing four objects*

Figure 3 displays results for the Communicative behavior in the same conditions. This behavior produces a clear differentiation between gestures associated to each object, in a coherent way from one agent to another, which results in an understanding rate reaching around 80%. This recalls the "Dispersion Theory" principle introduced by Lindblom for vowels [1]. Actually, it can be shown that the Communicative behavior leads to a kind of maximum dispersion between means of the $P(S|\ O_L)$ distributions [7].

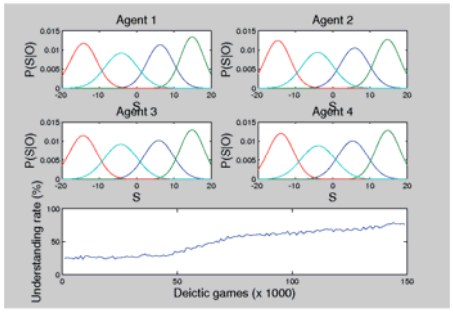

Figure 3: *Communicative behavior results*

Moreover, if one introduces nonlinearities in the percept function, we may simulate effects compatible with Stevens' Quantal Theory [2]. Indeed, we observe on Figure 4 that whether the nonlinearity is on the right or on the left of the motor space, the boundary between $P(S|\ O_L)$ distributions for the two objects is accordingly shifted towards the right or the left.

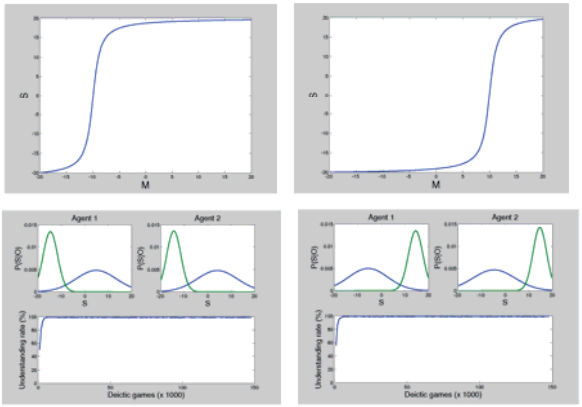

Figure 4: *Upper part : two non-linear percept functions. Lower part : corresponding results in the Communicative behavior (4 agents, 2 objects).*

Results for the Hybrid behavior are displayed on Figure 5 in a society of four agents, an environment containing four objects and a linear percept function. As with the Communicative behavior, this behavior leads to the emergence of a common speech code between the agents, with the same properties concerning dispersion and non-linearities. It seems to be the best behavior in terms of both performance

and theoretical basis. The understanding rate quickly reaches 100%. Moreover, gestures being selected both for their motor and sensory qualities, it is in line with the sensory-motor theory of speech production that we defend in our lab with the Perception for Action Control Theory (PACT, [8]).

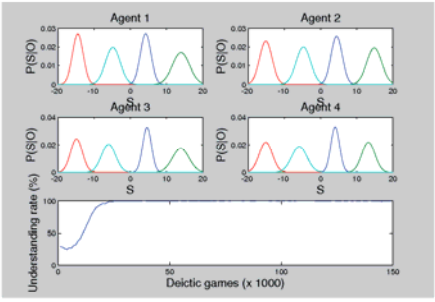

Figure 5: *Hybrid behavior results*

3.2 Realistic sensory-motor model

We now consider a more realistic model of the orofacial sensory-motor apparatus (VLAM, [9]) where motor gestures are three-dimensional (tongue body, tongue dorsum and lips height) and sensory percepts are two-dimensional (first formant and second effective one, in Bark). In this preliminary study, we consider static configurations corresponding to vowels. The Efference Copy Model P(S|M) is modeled by Gaussian distributions learnt through previous sensory-motor exploration. Speaker and Listener Models, $P(M|\ O_S)$ and $P(S|\ O_L)$ are also Gaussian distributions family, learnt during the deictic games. Figure 6 provides results for a society of two agents in a hybrid behavior and an environment containing three objects. We observe that the three vowels occupy the vertices of the vocalic triangle, that are the three vowels /a,i,u/ used in most humans languages.

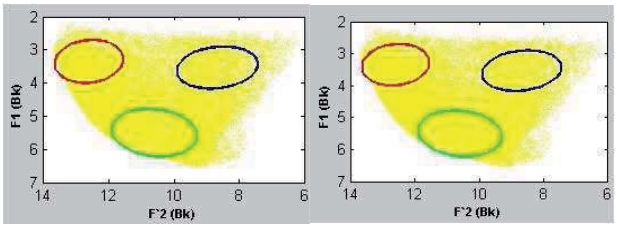

Figure 6: *Hybrid behavior in the vocalic triangle*

4 Conclusions and perspectives

This paper provides a Bayesian Robotic implementation of the "Vocalize-to-Localize" framework for communication between sensory-motor agents, in which deixis is set as the driving force. It is shown that a "Hybrid behavior" combining a Speaker and a Listener model in the agents behavior provides evolutions compatible with Lindblom's Dispersion Theory, Stevens' Quantal Theory and Schwartz et al.' Perception for Action Control Theory. The next study will consist in going towards dynamic patterns including consonants and syllables in the framework of the Frame and Content Theory by MacNeilage and Davis [10].

References

[1] J. Liljencrants and B. Lindblom. *Numerical simulation of vowel quality systems : the role of perceptual contrast*. Language, 48 :839-862, 1972.

[2] K. Stevens, On the quantal nature of speech. *Journal of Phonetics*, 17(1/2) :4-45, 1989.

[3] C. Abry, A. Vilain, and J.-L. Schwartz. Vocalize to localize. *Interaction Studies*, 5(3):313-325, 2004.

[4] O. Lebeltel, P. Bessière, J. Diard, and E. Mazer. Bayesian robot programming. *Autonomous Robots*, 16 :49-79, 2004.

[5] C. Frith. *Neuropsychologie cognitive de la schizophrenie.* Paris, PUF, 1996.

[6] M. Iacoboni, *Understanding others: imitation, language, empathy.* In S. Hurley and N. Chater (Eds.) Perspectives on Imitation: From Cognitive Neuroscience to Social Science. Cambridge, MA: MIT Press.

[7] C. Moulin-Frier, *Jeux déictiques dans une société d'agents sensori-moteurs.* Master Thesis, Grenoble Institute of Technology, June 2007.

[8] J.-L. Schwartz, L.-J. Boë, C. Abry. Linking the DFT and the 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 University Press.

[9] L.-J. Boë, S. Maeda. 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, 1997.

[10] P. F. MacNeilage. The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*,21:499-546,1998.