

COSMO ("Communicating about Objects using Sensory–Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems

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

▶ **To cite this version:**

Clément Moulin-Frier, Julien Diard, Jean-Luc Schwartz, Pierre Bessière. COSMO ("Communicating about Objects using Sensory–Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. Journal of Phonetics, 2015, 53, pp.5-41. 10.1016/j.wocn.2015.06.001 . hal-01230175

HAL Id: hal-01230175
https://hal.science/hal-01230175

Submitted on 25 Nov 2015

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.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Journal of Phonetics 53 (2015) 5–41

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

COSMO ("Communicating about Objects using Sensory–Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems

Clément Moulin-Frier [a,d,*], Julien Diard [b], Jean-Luc Schwartz [a], Pierre Bessière [c]

[a] GIPSA-Lab UMR5216, CNRS, Grenoble University, France
[b] Laboratoire de Psychologie et NeuroCognition – UMR 5105 CNRS, Grenoble University, France
[c] Laboratoire de Physiologie de la Perception et de l'Action – UMR 7152 CNRS, Collège de France, Paris, France
[d] Flowers team, Inria / ENSTA-Paristech, Bordeaux, France

ARTICLE INFO

Article history:
Received 2 July 2014
Received in revised form
26 May 2015
Accepted 8 June 2015
Available online 9 October 2015

Keywords:
Speech sound systems
Language evolution
Phonology
Cognitive modelling
Self-organization
Bayesian programming

ABSTRACT

While the origin of language remains a somewhat mysterious process, understanding how human language takes specific forms appears to be accessible by the experimental method. Languages, despite their wide variety, display obvious regularities. In this paper, we attempt to derive some properties of phonological systems (the sound systems for human languages) from speech communication principles. We introduce a model of the cognitive architecture of a communicating agent, called COSMO (for "Communicating about Objects using Sensory–Motor Operations') that allows a probabilistic expression of the main theoretical trends found in the speech production and perception literature. This enables a computational comparison of these theoretical trends, which helps us to identify the conditions that favor the emergence of linguistic codes. We present realistic simulations of phonological system emergence showing that COSMO is able to predict the main regularities in vowel, stop consonant and syllable systems in human languages.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

1.1. Universality of language forms

While the general question of language origins still seems beyond testable assumptions, the question of the origins of language forms – that is, assuming that language emerged for some reason, predicting what its units and systems should be – seems to be more amenable to an experimental approach. This question has gained much interest in recent decades and now attracts a large body of research from a wide range of disciplines. Indeed, human languages display a number of regularities, called "universals", and consequently, all human languages, while they are different from one another, are based on general principles and follow strong statistical trends. This makes them appear merely as variants of a single system. The issue that we address in this paper concerns the origins of these general principles and statistical trends, focusing on the universals of sound systems for human languages.

Universality of language forms has generated three kinds of explanation. In the Chomskyan view of a genetically specified language organ (Chomsky, 1965), a common innate language competence shared by all humans would explain the common denominator of the different languages. Another view about the universal properties of human languages may be found in the hypothesis of a common origin, by which human languages would derive from an African mother tongue (Ruhlen, 1996), imposing some common traces in spite of further cultural evolution producing their diversity (Atkinson, 2011; Gell-Mann &

* Corresponding author.
E-mail address: clement.moulinfrier@gmail.com (C. Moulin-Frier).

http://dx.doi.org/10.1016/j.wocn.2015.06.001
0095-4470 © 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Ruhlen, 2011; however, see Boë, Bessière, Ladjili, & Audibert, 2008 for strong methodological caveats against the common origin hypothesis).

A third view considers that the forms of human language are the emergent product of an optimization process, inducing some commonality in the achieved solutions because of commonality in the cognitive mechanisms at hand in the dynamic mechanisms involved, and because of common exterior constraints. This is the view first popularized by Lindblom (1984), through a proposal to "derive language from non-language". In the 1990s,a series of computational models emerged, following the seminal proposal of "language games" by Steels (Steels, 1994, 1997, 1999). In these multi-agent models, agent populations are made to interact and evolve, leading to the emergence of global properties from local interactions, with these properties being analyzed in relation to those of language.

1.2. Deriving some properties of phonological systems from general communication principles

The wide range of proposals and debates on this topic diverge in their linguistic focus (mostly syntax and to a lesser extent, lexicon or phonology); in their basic objective (e.g. explain invariance vs. illustrate and explain variability of human languages; stress phylogenetic continuities by attempting to derive human language from animal communication vs. stress discontinuities by showing what cannot be explained in human languages from pre-human cognitive abilities); in their underlying assumptions (including physical, biological, cognitive, cultural ingredients). It is out of the focus of the present paper to propose an exhaustive review of these works (see a recent review in Oudeyer, 2006, 2013).

The present paper[1] is focused on phonological universals, embracing the view that some properties of the sound systems for human languages can be derived from the communicative and perceptuo-motor abilities of the human species. This work has two major ambitions: (1) to attempt to simulate in a single integrated framework universals of various components of human sound systems, namely vowels, plosives and their associations in CV syllables; and (2) to connect within this framework various theories of speech communication that have seldom be considered together, that are theories of speech sound systems on one hand (e.g. the dispersion theory or the quantal theory, see later) and theories of speech communication on the other hand (e.g. auditory, motor or perceptuo-motor theories, see later).

For this aim, we adopt the framework of a multi-agent system, in which each agent is initially constrained by a set of prelinguistic abilities. Our aim is to examine how basic properties of phonology could emerge in such a system. A number of variants of these multi-agent models can be identified (Berrah, Glotin, Laboissiére, Bessiıere, & Boë, 1996; de Boer, 2000; Oudeyer, 2005; De Boer & Zuidema, 2010), which offer simulations that display some of the properties of sound systems, such as vowel systems with dispersion, or sound sequences with phonotactic properties comparable with those of human speech. However, none of them addresses the two major challenges mentioned previously, that are associating various categories of speech sounds in the same set of predictions, and connecting simulations with the classical theories of speech communication.

Within this framework, our work is based on an analysis of the required components of a communication system. Communication is based on the modification of the internal-knowledge state of a listener, by a speaker, through the use of communication stimuli. In essence, communication is possible and efficient provided that three subproblems are conjointly solved by the speaker and the listener. Firstly, they must select adequate communication stimuli that are reasonably easy to produce by the speaker and to process by the listener (adequacy). Secondly, a good correspondence must be ensured between the speaker's motor repertoire and the listener's perceptual repertoire so that they can then exchange roles and still communicate smoothly (parity). Thirdly, they must know the correspondence between these motor and perceptual repertoires and the objects of the external world ("reference", which is an aspect of a more general symbol-grounding problem, Harnad, 1990; Steels, 2008).

Solutions to each of these three subproblems have motivated theoretical proposals about the emergence of language that are related to both phylogenetic and ontogenetic arguments. Let us consider some of them.

Firstly, the Frame-Content Theory (MacNeilage, 1998; MacNeilage & Davis, 2000) suggests an evolutionary path to articulated speech from ingestive mandibular cyclicities and claims that this provides an elegant solution to the generation of communication stimuli that are easy to produce and to process. By supporting a prelinguistic ability to produce modulated vocalizations efficiently, such cyclicities could have produced a bootstrap for a solution to the adequacy problem. Secondly, recent proposals about the existence of mirror neurons (located in the premotor and parietal cortex of the macaque, active both when the macaque performs a transitive action and when it observes another individual performing the same action, Rizzolatti & Arbib, 1998) and the mirror system in humans (active when a human both performs and observes a particular transitive action, Rizzolatti, Fadiga, Gallese, & Fogassi, 1996; Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995), are claimed to provide a solution for parity (Arbib, 2005a; Rizzolatti & Arbib, 1998). Starting from the very definition of parity by Liberman and Mattingly (1989), summarized by Liberman and Whalen (2000) by the well-known formula "*what counts for the speaker must count for the listener*", Arbib (2005b) is explicit on this point:

> "*The parity requirement for language in humans – that what counts for the speaker must count approximately the same for the hearer – is met because Broca's area evolved atop the mirror system for grasping with its capacity to generate and recognize a set of associations.*"

[1] The present paper compiles a series of experiments performed during the PhD thesis of Clément Moulin-Frier, supervised by Jean-Luc Schwartz, Julien Diard and Pierre Bessière (Moulin-Frier, 2011). It is based on our previous works linking standard speech communication theories to the study and simulation of phonological system emergence in interacting sensory–motor agents (Moulin-Frier, Schwartz, Diard, & Bessière, 2008, 2010, 2011). Here, we present these results in a larger framework, including unpublished original research on the simulation of stop and syllable emergence and emphasizing our view on the cognitive nature of speech sound systems in line with the focus of the present special issue.

Finally, the reference problem has generated strong debates between proponents, on the one hand, of a gestural origin for language, which involves a strong iconicity of gestures (Corballis, 2002; Gentilucci & Corballis, 2006) and the use of complex gestural imitation procedures (Arbib, 2005a), and proponents, on the other hand, of a vocal route, possibly using vocal deixis as a bootstrap (Abry, Vilain, & Schwartz, 2004). In both cases, shared attention would be a crucial aspect of these processes (Leavens & Bard, 2011).

We do not claim that solving these three subproblems explains human phonology completely. It could be argued that specific mechanisms are necessary, such as those that deal with the question of compositionality in phonology (Hauser, Chomsky, & Fitch, 2002), although we will provide some perspective on this point in our final discussion. Moreover, interactions among the suggested phylogenetic precursors during language evolution are certainly a crucial point. It has been suggested that mirror neurons, for example, can provide a solution for the reference problem through action recognition (Arbib, 2005b) and that synchrony of mandibular cyclicities (adequacy) and deictic pointing gestures (reference) could provide a developmental rendezvous toward first-words production (Abry, Ducey Kaufmann, Vilain, & Lalevée, 2008). Our main claim in this paper is that, if cognitive processes capable of solving the three subproblems of adequacy, parity and reference exist in a society of interacting agents, then a number of properties could be obtained by multi-agent simulation, which would appear similar to those of human phonological systems.

1.3. Assumptions, predictions and summary

The present work therefore aims (i) to define interacting agents equipped with a basic ability to produce and perceive sounds and with a cognitive system able to solve the “adequacy”, “parity” and “reference” subproblems, (ii) to implement simulations within which these agents orally interact, (iii) to study oral communication systems emerging from the interactions between these agents and (iv) to assess whether these systems display a number of properties of human phonological systems.

Therefore, we do not attempt in this work to derive language from nothing. On the contrary, we capitalize on a number of assumptions about orofacial production and auditory perception, and about cognitive abilities described previously. From these, we propose a computational framework for expressing communication between agents and for defining how a society of agents might evolve through inter-agent communication. We then analyze the results of these simulations and make comparisons with some known properties of human sound systems.

To address these aims, the next section presents the universals and statistical properties of phonological systems that we are interested in, and the way that they have been described and sometimes simulated in previous theoretical and modeling studies.

We then offer our three main contributions:

Firstly, in the Section 3, we present the core of our own approach; namely, the analysis of the communication paradigm and the basic constraints that this imposes on the form of an adequate communication system. This leads to the cognitive architecture at the center of our modeling approach; namely, “Communicating about Objects using Sensory–Motor Operations” (COSMO). We show that COSMO integrates various theoretical proposals about speech communication into a single framework, and we introduce a Bayesian implementation that leads to a computational version of COSMO. In Section 4, we define “deictic games” as the key tool for our simulations, in which computational agents interact in the presence of objects that they attempt to designate by the voice.

Our second contribution, in Section 5, is an experimental study involving deictic games in a simplified sensory–motor space, of our main theoretical proposals about speech communication. It enables us to extract some cognitive conditions necessary for the emergence of a speech code in COSMO.

Our third and final contribution, in Section 6, shows how the model is able to predict coherent vowel, consonant and syllabic phonological systems by using a realistic vocal-tract model and a phylogenetically and ontogenetically plausible control framework. We analyze our simulation results with respect to some of the regularities in the sound systems for human languages described previously.

In the final section, Section 7, we discuss the validity, limitations and perspectives offered by our work in the field of theories about the emergence of human language forms.

2. Regularities in phonological systems

2.1. Universals and major trends in sound systems for human languages

Languages are characterized by a “double articulation” in meaning and sound units. That is, there is a combination of elementary meaning units, typically words, which are themselves a combination of elementary sound units, typically phonemes. Considering sound, which will be the main focus of this paper, it appears that phonological systems display more or less regular alternations of vowels and consonants in syllables, usually corresponding to open and close configurations of the vocal tract, respectively. Within the phonological systems themselves, strong regularities can be observed.

2.1.1. Vowels

Vowels can be characterized by their formant values. The debate between local parametric representations based on formants and global spectral representations, based on e.g. critical band spectra or cepstral parameters, is classical (see a discussion in Schwartz, Boë, Vallée, & Abry, 1997a). While global parameters are the only adequate way to deal with automatic processing of acoustic stimuli, formants remain the reference in simulations about phonological inventories, therefore they will be used in the present work.

Vowel system inventories available in the phonological database UPSID, which samples about 10% of human languages (Maddieson, 1984; Maddieson & Precoda, 1989), display clear regularities (Maddieson & Precoda, 1989; Vallée, 1994; Schwartz et al., 1997b). In particular, almost every language has /a,i,u/, with /a/ as in "car" (from a low jaw and a low and central tongue), /i/ as in "bee" (from a high jaw and a high and front tongue) and /u/ as in "food" (from a high jaw and a high and back tongue). The most common system is the five-vowel system /i,e,a,o,u/, with two additional vowels /e/ and /o/ (from jaw–tongue configurations intermediate between those for /a,i,u/). These vowels are distributed rather evenly in the space of the two first formants (*F*1, *F*2) (see Fig. 1).

2.1.2. Consonants

Consonants are more difficult to characterize as a class than vowels, both acoustically and from an articulatory point of view. Among the consonants, the stop consonants (or "stops"), which are characterized by a complete "closure" of the vocal tract, are relatively well described (see Schwartz, Boë, Badin, & Sawallis, 2012b). We therefore use them as the focus of the present study. For stops, either the tongue or the lips are put in such a configuration that there is a place inside the vocal tract of complete closure where air cannot pass. It is only when the closure is "released" by an appropriate articulatory maneuver that the characteristic sound of the stop is uttered. Stops are best characterized by the place of the closure (called the "place of articulation") and the formant values (*F*1, *F*2, *F*3) at the time of release (Schwartz,et al., 2012b) (see Fig. 2).

The "best" stops in phonological systems (clearly demonstrated in UPSID) are bilabial, coronal and velar plosives. Indeed, if a language has three stops (as in 3% of UPSID languages), it has plain voiceless /p t k/, rather than other combinations of place or of place and secondary articulations. If a language has six stops, which is the most frequent number (in 24% of the languages of UPSID), it has mostly /p t k b d g/. With 9 stops, the basic /p t k b d g/ series combines with secondary feature sets such as aspiration, prenasalization, palatalization or laryngealization. Altogether, in the UPSID extension to 556 languages (Maddieson, 2001), 45% of the systems include the 6 stops /p t k b d g/. In the following, we will not deal any more on voicing, which is out of the field of our simulations, and concentrate on plosive place of articulation. Voiceless stops are more frequent than voiced in human languages (Boë, Vallée, Badin, Schwartz, & Abry, 2000), but the formant transitions are more fully visible in voiced than in voiceless stops, and hence easier to model in the articulatory models that will be described later. Therefore, for expository convenience, we will use and cite voiced rather than voiceless stops for the discussion and simulation of plosive place throughout this paper.

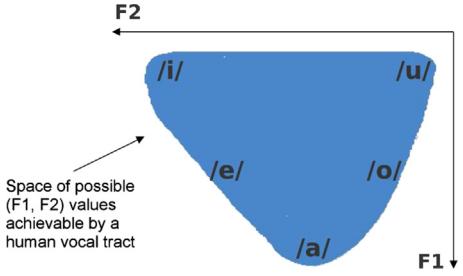


Fig. 1. Vowel characterization. The vowel (*F*1, *F*2) space and the most common system /i,e,a,o,u/.

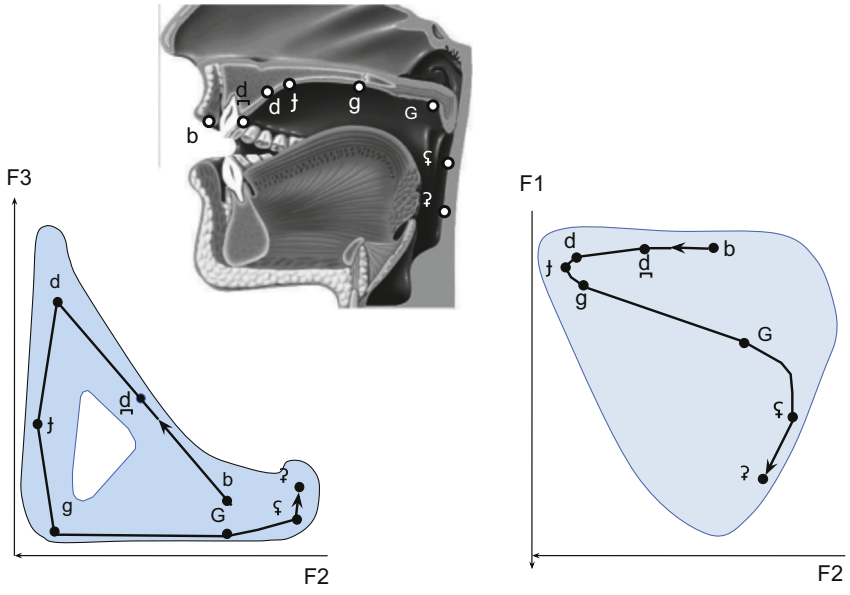


Fig. 2. Plosive consonant characterization. Top: the closure places for stops. Bottom: acoustic spaces for stops. Adapted from Schwartz et al. (2012)b – vocal tract infography by Sophie Jacopin.

Stops are all produced with a rather closed jaw, enabling a vocal tract closure at the lips (for the bilabial /b/), toward the front of the palate with the tip of the tongue (for the alveolar /d/) and toward the middle of the palate with the tongue dorsum (for the velar /g/) (see Fig. 2, top). They correspond to small *F*1 values and are rather well differentiated in the (*F*2, *F*3) plane (see Fig. 2, bottom).

In contrast, the "back" stops from /ɢ/ to /ʡ/ are performed with the dorsum of the tongue closing the vocal tract toward the back of the mouth (see Fig. 2, top), with the jaw being lower and even largely open in order to push the back of the tongue toward the pharyngeal part of the palate in /ʡ/. They are less favored in phonological inventories, despite their rather distinctive high *F*1 values (see Fig. 2, bottom).

2.1.3. Co-occurrences in syllables

Finally, it has been suggested that some associations in human languages between stops and vowels are more frequent when the tongue is in the same region of the vocal tract for both the consonant and the vowel. Examples include the tongue being toward the front (with a front /i/ and an alveodental /d/ in /di/), toward the middle (with a central /a/ and a labial /b/ with no active tongue movement in /ba/) and toward the back (with a back /u/ and a velar /g/ in /gu/) (MacNeilage & Davis, 2000; Vallée et al., 2009).

2.2. Predicting form from substance

At the beginning of the 1970s, evidence for the existence of such regularities in sound inventories led researchers to develop what Lindblom called "substance-based theories", in which form was supposed to emerge from properties of the articulatory-acoustic substance of speech communication. Two such theories were introduced in the 1970s and developed over the subsequent 20 years.

According to Lindblom's "Dispersion Theory" (Liljencrants & Lindblom, 1972; Lindblom, 1984, 1986, 1990; Schwartz et al., 1997a), human languages have sound systems for which auditory distances between pairs are maximal in order to enhance distinguishability. In this context, auditory distances are typically defined in the formant (*F*1, *F*2) space. This leads to clear-cut predictions about vowel systems, for which sounds would be most distant from one another in the vowel triangle, generating such systems as /i,a,u/ for three-vowel systems and /i,e,a,o,u/ for five-vowel systems (see Fig. 1 for a visualization of these vowels in the (*F*1, *F*2) space).

Stevens' "Quantal Theory" (Stevens, 1972, 1989; Stevens & Keyser, 2010) claims that nonlinearities in the articulatory-to-acoustic transformation may lead to the existence of so-called "quantal regions", where plateaus, in which percepts are almost invariant to articulation, are separated by sharp natural boundaries, where sound changes quickly for very small articulatory displacements. These "substance" natural boundaries provide, according to Stevens, optimal niches for phonological contrasts in sound systems.

However, these theories appear as "phenomenological laws" associating a phonological contrast (form) to a phonetic behavior (substance): auditory distinctiveness in one case, articulatory-to-acoustic nonlinearity in the other. Emergence simulations are a way to ground these phenomenological laws in principles of communication per se. The next section proposes a set of such principles.

3. The COSMO model

This section presents the cognitive architecture at the center of our modeling approach; namely, "Communicating about Objects using Sensory–Motor Operations" (COSMO). We first expose the conceptual principles of COSMO, where an agent fully internalizes the structure of a communication situation. Then we propose a computational version based on a Bayesian implementation, to finally show that COSMO integrates various theoretical proposals about speech communication.

3.1. COSMO principles

3.1.1. A speech-communication prototypical paradigm

The scenario that we propose for the emergence of speech communication assumes that the three evolutionary requirements suggested in the introduction – namely, those related to adequacy, parity and reference – are solved in one way or another. In its general form, our scenario is agnostic in relation to choices about adequate phylogenetic precursors of these requirements.

However, in the implementations of the model that we describe in the following sections, we make several choices that guide the simulations. For present purposes, it suffices to say that the speech-communication situation that we consider is illustrated in Fig. 3. Here, a speaker has the desire to communicate about an object O_S to a listener.

Before describing Fig. 3 in more detail, we emphasize that the term "object" in COSMO may take a variety of meanings. It can be a physical object related to a word, or an event or a concept related to a more complex syntactic structure. It can also refer directly to a

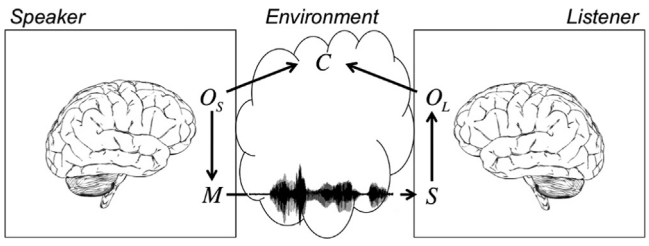


Fig. 3. Schema of the speech communication situation.

"first articulation unit"; i.e., a phoneme or a syllable. Of course, this assigns to the "object" category a very general sense, which will have to be specified and developed in future work. In the present work and for the sake of simplicity, we consider that a physical object (e.g. an apple) is associated with a single phonological unit (e.g. a vowel or a syllable). In the simulations that we will describe, the number of objects (and therefore the number of phonological units) will be fixed in advance, and the challenge will be to learn a mapping between these units and articulatory–auditory configurations, allowing the agents to communicate efficiently about the physical objects. "Objects" will then refer to the phonological units used to name physical objects, providing the communication units for the systems studied.

In Fig. 3, the speaker is provided with a brain that includes a set of cognitive control processes acting on a vocal tract through different articulators. We denote this set of articulators globally as M (for *Motor*). The vocal tract then produces a sound wave, from which the listener has to infer the communication object by means of an ear allowing perception of the sound and a brain that includes a set of auditory processing capabilities. We denote this set globally as S (for *Sensory*), with O_L being the object inferred by the listener. Finally, communication success is defined by the condition $O_S = O_L$ and denoted by $C = 1$.

To summarize, in this speech-communication situation

- the path from O_S to M represents the speaker's production task,
- the path from M to S is the articulatory-to-acoustic transformation resulting from the physics of sound creation and transmission,
- the path from S to O_L represents the listener's perception task, and
- C is the communication success condition, simply defined as $O_L = O_S$.

3.1.2. COSMO and the "*internalization*" *hypothesis*

We now present the COSMO model for speech communication and language emergence. The COSMO acronym also represents the five variables around which the basic structure of the model is built (see Fig. 3). In COSMO, communication (C) is a success when an object O_S in the speaker's mind is transferred, via sensory and motor mechanisms S and M, to the listener's mind, where it is correctly recovered as O_L.

The central hypothesis of the COSMO model is that a communicating agent, which is both a speaker and a listener, is able to internalize fully the communication situation described previously (see Fig. 3) inside an internal model (see Fig. 4).

Firstly, the agent can take both roles, listener and speaker, and therefore contains both motor and sensory subsystems. Secondly, the agent has acquired by learning some knowledge about the articulatory-to-acoustic transformation performed by the environment. When it is internalized, it takes the form of an internal forward model, allowing the agent to predict the sensory consequences of motor gestures. Finally, the communication condition is also internalized, with the agent having two internal representations of an object, linked by a system that verifies whether they refer to the same object. This "internalization" hypothesis therefore results in an architecture combining

- a motor system able to associate communication objects O_S with motor gestures M,
- an auditory system able to associate communication objects O_L with auditory stimuli S,
- a sensory–motor link able to associate motor gestures M with auditory stimuli S (providing an internal model of the articulatory-to-acoustic transformation), and
- a fusion system able to associate the communication objects in both the motor (O_S) and auditory (O_L) branches through C.

The C variable is interpreted somewhat differently in Figs. 3 and 4. Although it corresponds to the success of the communication between two agents in the first case, it is treated as an internal variable of the cognitive architecture of an agent in the second case. As explained below, the latter form is used for the *coherence variable* (Bessiére, Mazer, Ahuactzin, & Mekhnacha, 2013; Gilet, Diard, & Bessiére, 2011; Pradalier, Colas, & Bessiere, 2003) that enables an agent to perform a fusion of the information available in both its motor and auditory branches.

This internalization hypothesis could be discussed within the framework of general cognitive theories of social communication and human evolution (Baron-Cohen, 1997; Tomasello, Carpenter, Call, Behne, & Moll, 2005) (see also Moore, 2007 for similar views about internalization, expressed in a control theory framework).

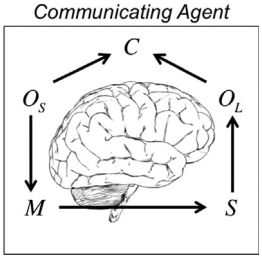


Fig. 4. Schema of the communicating agent model.

An agent with this cognitive architecture possesses a model of the entire communication situation and is therefore able to perform both production and perception tasks within an integrated perceptuo-motor architecture. The internalization hypothesis provides, as we show below (Section 3.3), a basis for the unification of various theories of speech communication.

3.2. A Bayesian implementation of COSMO

3.2.1. From a conceptual model to a computational model

The first step in our simulation program is to propose a computational implementation of COSMO sensory–motor agents. This is done in the Bayesian programming framework. Applying Bayesian modeling, particularly when dealing with sensory–motor behaviors, finds its justification in the fact that any model of a real phenomenon is by nature incomplete (Lebeltel, Bessiere, Diard, & Mazer, 2004): there are always hidden variables that influence the phenomenon and are however not taken into account in the model. Thus, the use of classical logic quickly finds its limits and Jaynes (2003) proposed the use of probability theory as an alternative to logic for reasoning about incomplete and uncertain knowledge. In this framework, the subjective interpretation of probability theory allows the programmer to specify in a rigorous and precise way incomplete and uncertain knowledge of the sensory–motor agent, to propose learning and interaction processes and to simulate them by automated processes of probabilistic inference.

Bayesian modeling provides a mathematical framework that allows comparative assessment of cognitive theories, by embedding them within a common mathematical framework based on probabilities. Such comparisons are more and more widespread in cognitive science; see for instance the recent works on causal inference and probability matching strategies in multimodal perception (Kording et al., 2007), or on theoretical comparison of memory models (Myung & Pitt, 2009). This is why Bayesian programming is at the heart of the present work for comparing various kinds of theories and simulations – though no assumption is done in this paper about the view that Bayesian computations could be indeed at work in the human brain.

In Bayesian programming (Bessiére et al., 2013; Lebeltel et al., 2004), one specifies knowledge of a sensory–motor agent as a joint probability distribution over variables of interest (typically motor, sensory and internal variables). This joint probability distribution is generally broken down into a product of simpler distributions, using Bayes' rule and conditional independence hypotheses. Once it is defined, this knowledge can then be used, with sensory–motor behavior being defined as a conditional probability distribution computed from the joint distribution. For example, given the values of some sensory variables, what is the probability distribution over the motor variables? The term to be computed is called a question, one that is to be answered using Bayesian inference. We describe the mathematical bases of this framework in Appendices A.1–A.3.

The Bayesian model of COSMO is a direct translation of the conceptual cognitive architecture of Fig. 4, which can be interpreted as the structure of a Bayesian Network model. We define five variables M, S, O_S, O_L and C, where

- M represents the motor gestures that an agent is able to produce,
- S represents the auditory stimuli that an agent is able to perceive,
- O_S and O_L represent the communication objects (when the agent takes the speaker and listener point of view, respectively), and
- C represents the internalization of the communication success condition and will allow the fusion of motor and auditory branches (it is a Boolean variable that is *true* (denoted by the value 1) when $O_L = O_S$ and *false* otherwise).

We then define four subsystems:

- The motor subsystem is defined as a conditional probability distribution $P(M \mid O_S)$: given an object to communicate, this is the probability distribution over the speaker's motor gestures.
- The sensory–motor subsystem is defined as a conditional probability distribution $P(S \mid M)$: given a motor gesture, this is the probability distribution over the corresponding auditory stimuli.
- The auditory subsystem is defined as a conditional probability distribution $P(O_L \mid S)$: given a sensory stimulus, this is the probability distribution over the objects that can be inferred by the listener.
- The fusion subsystem of motor and auditory branches is defined as a conditional probability distribution $P(C \mid O_S\ O_L)$: given objects O_S and O_L, this is the probability distribution over the successful communications.

The COSMO Bayesian model is the joint probability distribution $P(O_S\ M\ S\ O_L\ C)$, defined by

$$P(O_S\ M\ S\ O_L\ C) = P(O_S)P(M \mid O_S)P(S \mid M)P(O_L \mid S)P(C \mid O_S\ O_L). \tag{1}$$

The mathematical assumptions behind the above equation are described in Appendix A.2

In the remainder of this paper, two of the five terms of this decomposition will have a fixed definition:

- $P(O_S)$, the prior distribution over objects O_S, is assumed to be a uniform probability distribution (which is just a helpful assumption that a speaker talks about each talkable object with equal probability).
- $P(C \mid O_S\ O_L)$, the fusion system, is defined by a Dirac probability distribution: $P([C = 1] \mid O_S\ O_L)$ is 1 if and only if $O_S = O_L$, and is 0 otherwise.

The motor, sensory–motor and auditory subsystems will be defined differently for the various simulation experiments that we describe below. For the moment, assume that they are associated with mathematical forms, so that the joint probability distribution of Eq. (1) is fully defined: it contains all the agent's cognitive knowledge about the communication situation.

3.2.2. Using the computational model for solving communication tasks

After the joint probability distribution of Eq. (1) is defined, production and perception tasks can be defined as questions about this joint probability distribution (see Appendix A.3 for further details). In both tasks, the agent will activate the fusion system by setting $C=1$; i.e., ensuring that motor and auditory branches both relate to the same communication object.

In production tasks, we compute probability distributions over motor gestures, given an object to communicate about. This is an inference path from O_S to M in the Bayesian network of Fig. 4. In probabilistic terms, it involves computing a distribution over motor gestures given an object $[O_S=o_i]$ and the activation of the fusion system $[C=1]$; i.e., the probabilistic question $P(M\mid[O_S=o_i][C=1])$. Bayesian inference, from the joint probability distribution of Eq. (1), leads to

$$P(M\mid[O_S=o_i][C=1])\propto P(M\mid[O_S=o_i])\sum_S P(S\mid M)P([O_L=o_i]\mid S). \tag{2}$$

The symbol "$\propto$" denotes proportionality: it just means that we omit a normalization factor in the equation, see Appendix A.3 for further explanation.

In perception tasks, we compute probability distributions over objects, given an input sensory signal. This is an inference path from S to O_L in the Bayesian network of Fig. 4. In probabilistic terms, it involves computing a distribution over objects given an auditory stimulus $[S=s]$ and the activation of the fusion system $[C=1]$; i.e., the probabilistic question $P(O_L\mid[S=s][C=1])$. Bayesian inference, from the joint probability distribution of Eq. (1), leads to

$$P(O_L\mid[S=s][C=1])\propto P(O_L\mid[S=s])\sum_M P(M\mid[O_S=O_L])P([S=s]\mid M). \tag{3}$$

The next section interprets these two complex probabilistic formulations and shows that they allow a computational expression of the main theoretical trends in the speech production and perception literature.

3.3. Probabilistic unification of motor, auditory and sensory–motor theories of speech communication in COSMO

3.3.1. Nature of the content of communication in speech communication theories

A key question in speech science concerns the nature of the reference frame of communication, with three major frameworks being motor, auditory, and sensory–motor theories of speech communication.

Motor theories consider the reference frames for speech communication as gestures. This results in speech-production models conceived as combinations of articulatory gestures able to express the context-dependent variability of speech, without explicitly taking into account the auditory consequences of a motor event (the Articulatory Phonology framework, Browman & Goldstein, 1986, 1989, 1992). On the speech-perception side, the Motor Theory of Speech Perception (Liberman & Mattingly, 1985) suggests that perceiving speech amounts to perceiving gestures, supposed to be more invariant than sounds (notice that the "other" motor theory provided by the Direct Realist Theory of speech perception does not introduce this assumption, considering that perceiving speech is perceiving gestures for general reasons associated with direct realism as a general process, Fowler, 1986). This necessitates an inverse model that allows retrieval of the speaker's intended motor gesture from the received acoustic stimulus.

Auditory theories consider that the reference frame for speech is auditory. In the case of speech perception, proponents of auditory theories assume that speech perception involves auditory or multisensory representations and processing, with no explicit call on knowledge coming from speech production (Diehl, Lotto, & Holt, 2004). In the case of speech production, the target would be a region in the auditory space (Guenther, Hampson, & Johnson, 1998). Auditory theories therefore place the inverse problem on the production side and limit perception to a signal-processing task.

Sensory–motor theories have emerged recently for both speech perception and production. They usually consider auditory frames as the core of communication, but they include a sensory–motor link inside the global architecture, to deal with feedback adjustment processes in speech production (Guenther, 2006) or speech perception in complex or adverse conditions (Skipper, Van Wassenhove, Nusbaum, & Small, 2007; Schwartz, Basirat, Ménard, & Sato, 2012a).

We consider the COSMO model as a general, sensory–motor computational model of speech communication; i.e., a proposed computational implementation of sensory–motor theories (Moulin-Frier, Laurent, Bessière, Schwartz, & Diard, 2012). We now show how, by disabling portions of the COSMO model, we can recognize implementations of either motor or auditory theories.

3.3.2. COSMO implementation of motor theories

Firstly, in motor theories, an agent only has knowledge about its motor representations[2] and the articulatory-to-acoustic transformation (through the $P(M\mid O_S)$ and $P(S\mid M)$ distributions, respectively). We disable the direct connection between S and O_L

[2] Henceforth, in the framework of Bayesian programming of sensory–motor agents (Lebeltel et al., 2004; Bessière et al., 2008) the term "representation" will refer to an instantiation of external variables or a probabilistic relation between variables.

by setting the distribution $P(O_L \mid S)$ as a uniform probability distribution, which encodes a complete lack of knowledge. Answers to the production and perception questions are therefore simplified in the following way.

Consider first a speech production task $P(M \mid [O_S = o_i][C = 1])$. Setting $P(O_L \mid S)$ as uniform in (2) yields a simple expression for speech motor control; namely, $P(M \mid [O_S = o_i])$: what is the speaker's usual action for a given object o_i? (see, for example, Browman & Goldstein, 1986.)

Consider next a speech perception task $P(O_L \mid [S = s][C = 1])$. Setting $P(O_L \mid S)$ as uniform in (3) yields a more complex expression; namely, $\sum_M P(M \mid [O_S = O_L])P([S = s] \mid M)$. This question can be interpreted as a probabilistic "motor theory of speech perception". First, it is important to notice that motor terms are involved in the equation, whereas the auditory subsystem $P(O_L \mid S)$ is not, which provides a "speech specific" way to process sensory inputs. Interpreting the probabilistic equation in algorithmic terms provides intuitive understanding of the relationship to the classical Motor Theory of Speech Perception. Indeed, summing over M the $P([S = s] \mid M)$ term implements in a Bayesian framework the search for motor values able to lead to the perceived sensory input s (this provides a probabilistic equivalent to the "inversion" or "analysis by synthesis" processes). This "search" is weighted by a $P(M \mid [O_S = o_i])$ factor, which can be conceived of as an "articulatory decoder", that assumes that invariance for a given object o_i lies in motor rather than auditory cues (see, for example, Liberman & Mattingly, 1985). This should make clear that "deterministic" inversion in which ad hoc terms are introduced to guide inversion towards a specific articulatory configuration (see a recent review in Demange & Ouni, 2013) is replaced here by a Bayesian implementation in which no specific configuration is selected, but all configurations are considered for articulatory decoding, and weighted by their probability of providing the corresponding auditory input.

3.3.3. COSMO implementation of auditory theories

Secondly, in auditory theories, an agent only has knowledge about its sensory representations and the articulatory-to-acoustic transformation (through the $P(O_L \mid S)$ and $P(S \mid M)$ distributions, respectively). We disable the direct connection between O_S and M by setting the distribution $P(M \mid O_S)$ as a uniform probability distribution. Again, this simplifies the answers to the production and perception questions.

This process leads to auditory theories of speech perception, for which Eq. (3) simplifies into $P(O_L \mid [S = s])$, which involves direct inference without any motor knowledge, typical of these theories (see, for example, Diehl et al., 2004).

For auditory theories of speech production, Eq. (2) simplifies into $\sum_S P(S \mid M)P([O_L = o_i] \mid S)$. This estimates, through auditory inference, the gestures that the speaker should produce to make the listener perceive the appropriate object. This corresponds to associating auditory targets, defined by the term $P([O_L = o_i] \mid S)$, with the inversion of the forward model, defined by the term $P(S \mid M)$, for estimating appropriate motor commands (see e.g. Guenther et al., 1998).

3.3.4. COSMO implementation of sensory–motor theories

Finally, sensory–motor theories take into account the information provided by both motor and sensory subsystems. In other words, neither subsystem is disabled. Bayesian inference therefore yields Eqs. (2) and (3); namely, $P(M \mid [O_S = o_i])\sum_S P(S \mid M)P([O_L = o_i] \mid S)$ for speech production tasks and $P(O_L \mid [S = s])\sum_M P(M \mid [O_S = O_L])P([S = s] \mid M)$ for speech perception tasks.

These can be seen as combinations of the two previous inferences; i.e., fusions between purely perceptual and purely motor inferences. This corresponds to a sensory–motor theory that takes the computational form of products of both motor and auditory processes (Guenther, 2006; Schwartz, et al., 2012a; Skipper et al., 2007).

In the case of production tasks, given an object o_i to communicate, the chosen motor gesture is then a trade-off between a usual gesture for the speaker (the $P(M \mid [O_S = o_i])$ term) and a gesture that has a good communicative value for the listener (the $\sum_S P(S \mid M)P([O_L = o_i] \mid S)$ term).

In the case of perception, the object is inferred using information both from the purely auditory decoder (the $P(O_L \mid [S = s])$ term) and from the motor decoder by inversion (the $\sum_M P(M \mid [O_S = O_L])P([S = s] \mid M)$ term).

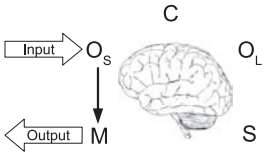
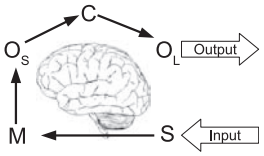
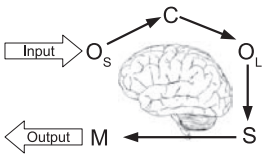
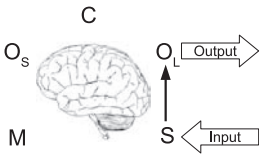
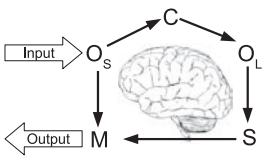
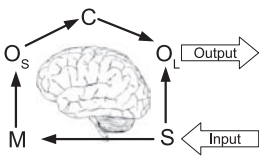
3.3.5. COSMO as a unifying framework where the sensory–motor link is key

These results are summarized in Table 1. Each cell in the figure contains a computational expression, which can be seen as a probabilistic implementation of major theories of speech communication. In other words, COSMO is a unifying Bayesian model: it implements sensory–motor theories of speech production and perception when it is fully instantiated, and it can implement auditory or motor theories of speech production and perception when some components are deactivated. This allows a rigorous association between computational expressions and theories of speech communication, which constitutes the first contribution of our work.

It is important to understand that all the variants of COSMO (i.e. the "auditory", "motor" or "sensory–motor" one) have a number of properties in common:

- they are all able to operate as either a speaker or a listener, depending on the question asked to the model, that is a "production" task based on computing $P(M \mid [O_S = o_i][C = 1])$, or a "perception" task based on computing $P(O_L \mid [S = s][C = 1])$;
- they all contain and exploit knowledge about the relationship between gestures M and sounds S, either for perception, or production, or both; this knowledge is based on the link $P(S \mid M)$;
- in addition, to this direct link, there is a secondary link through the variable C, which enables us to fuse perceptual and motor information inside both the production and perception tasks, provided that such information is available, which depends on the type of model implemented in COSMO;

Table 1
Synthesis of the three main theoretical frameworks in speech production and perception. Light arrows represent the corresponding information flow in the cognitive agent model (not to be confused with the direction of probabilistic dependencies in the model). Probabilistic inferences are derived from the unified Bayesian model. Typical references to the litterature are provided at the top of each cell.

Type of theory	Production	Perception
Motor	Browman and Goldstein (1986)	Liberman and Mattingly (1985)
		
	$P(M\|[O_S=o_i][C=1]) \propto P(M\|[O_S=o_i])$	$P(O_L\|[S=s][C=1]) \propto \sum_M \left(\begin{matrix} P(M\|O_S=O_L) \\ P([S=s]\|M) \end{matrix} \right)$
Auditory	Guenther et al. (1998)	Diehl et al. (2004)
		
	$P(M\|[O_S=o_i][C=1]) \propto \sum_S \left(\begin{matrix} P(S\|M) \\ P([O_L=o_i]\|S) \end{matrix} \right)$	$P(O_L\|[S=s][C=1]) \propto P(O_L\|[S=s])$
Sensory–motor	Guenther (2006)	Schwartz et al. (2012a), Skipper et al. (2007)
		
	$P(M\|[O_S=o_i][C=1]) \propto P(M\|[O_S=o_i]) \sum_S \left(\begin{matrix} P(S\|M) \\ P([O_L=o_i]\|S) \end{matrix} \right)$	$P(O_L\|[S=s][C=1]) \propto P(O_L\|[S=s]) \sum_M \left(\begin{matrix} P(M\|O_S=O_L) \\ P([S=s]\|M) \end{matrix} \right)$

- this is where the difference arises between the various instantiations of COSMO. Indeed, in our formalism, no direct knowledge is available on the link between objects and gestures in auditory theories: it is useless in perception, and indirect, mediated by sounds, in production; and no direct knowledge is available on the link between objects and sounds in motor theories: it is useless in production, and indirect, mediated by gestures, in perception.

4. Reference: deictic games

4.1. Deixis as the background for reference

For the experimental simulations that we present in later sections, we consider deixis as a possible bootstrap for the reference requirement described in the introduction. Deixis is the ability to show an element of the environment to someone else, pointing to it by the hand, the face, the eye or any posture of the body, and then to manage the shared attention. This behavior could have provided a bootstrap to a prelinguistic reference, which would allow, combined with vocalizations, evolution toward a meaningful language ("Vocalize-to-Localize", Abry et al., 2004). This refers to the ability of a number of prehuman primates to "show with the voice" by adapting their calls to different kinds of danger in the environment (Cheney & Seyfarth, 1982; Manser & Fletcher, 2004). Deixis also seems to be a necessary developmental step toward language, with deictic gestures usually appearing just before a child's syntax acquisition and vocabulary expansion (Goldin-Meadow & Butcher, 2003; Volterra, Caselli, Capirci, & Pizzuto, 2005).

4.2. Deictic games

Following the language-game paradigm introduced by Steels (1997) and applied to phonological systems (Berrah, 1998; Berrah et al., 1996; de Boer, 2000; Oudeyer, 2005), we simulate societies of evolving communicating agents by exploiting language games, based on deixis, that we call "deictic games" and that may be considered as a variant of Steels' "naming games" (Steels, 1994, 1997). These kinds of interaction games leading to evolution and convergence of a society of agents towards a stable communication set whose properties are compared with those of human languages have been the topic of a very large set of studies, including in a number of cases Bayesian paradigms as in the present work (e.g. Griffiths & Kalish, 2007; Hurford, 1989; Oliphant, 1996; Zuidema & Westermann, 2003); see a review of this kind of algorithms applied in the field of speech sound systems in Oudeyer (2006, 2013).

In a deictic game, two agents communicate in front of a given "object" on which their attention is jointly focused via a deictic process of some kind, such as pointing. We drastically simplify the simulations by assuming that deixis is perfect, in that agents perfectly identify objects that they are facing with their communicating partner, and there is no ambiguity about objects. This would obviously not be the case in realistic settings, where complex notions about categories ("a fruit" vs. "an apple") or specifiers ("a red apple") refer to highly complex discovery mechanisms (see, for example, Dominey, 2007; Kemp & Tenenbaum, 2008; Roy, 2005; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). However, this simplification is in line with the topic of our work – namely, phonology – and it allows us to focus our simulations on the "adequacy" and "parity" requirements. Fig. 5 illustrates a deictic game.

Let us define N as the number of agents in the society, Q as the number of objects on which the agents can share attention and TransMS : $M \rightarrow S$ as the articulatory-to-auditory transformation. That is, each time an agent produces an articulatory configuration $m \in M$, both agents involved in the deictic game will perceive the corresponding auditory stimulus $s = \text{TransMS}(m) \in S$. The COSMO Bayesian model of Eq. (1) is embedded in each agent, with no prior knowledge in the motor and auditory subsystems. That is, the $P(M \mid O_S)$ and $P(O_L \mid S)$ distributions are largely uncertain, being, for example, uniform or Gaussian with a large variance. These distributions will be refined by learning from the data collected during their interactions. In our simulations, all agents in the society will embed the same implementation of COSMO; namely, just one of the motor, auditory or sensory–motor theories (defined in Section 3.3 and summarized in Table 1).

The simulation loop for a deictic game has these steps, which are repeated until there is convergence of the speech code:

- random selection of a first agent from among the N agents of the society, which receives the "speaker" status,
- random selection of another agent from among the $N-1$ remaining agents of the society, which receives the "listener" status,
- random selection of an object o_i from among the Q objects, received by both agents as an input, via the assumed shared-attention mechanism,
- the speaker selects a motor gesture $m \in M$ in order to name o_i, according to its production behavior (defined in Table 1), and emits it,
- the simulated environment propagates this motor gesture, by the articulatory-to-acoustic transformation, resulting in an acoustic signal $s = \text{TransMS}(m)$, possibly distorted by simulated acoustic noise, and
- both agents update their knowledge according to their respective experience, with the speaker updating its motor prototypes based on its knowledge of the object o_i and the motor gesture m, and the listener updating its auditory prototypes based on its knowledge of the object o_i and the acoustic stimulus s.

Mathematical details about the last step of this simulation loop, allowing the update of the joint distribution terms in (1), are described in Appendix A.4.

Computationally, the shared-attention mechanism will act as a learning supervision signal, indicating to agents that $C = 1$ (i.e., $O_S = O_L$ in Fig. 3) before each interaction throughout the simulation process. At the end of the simulation, to assess the phonological system emergence, we will test whether the learning process has converged to a state where agents can ensure the success of communication ($C = 1$) using only vocal production and perception tasks (i.e., without requiring further use of the shared-attention mechanism). This will be the basis of the evaluation of our simulation results in the next section. Let us recall that setting the internal variable C in Fig. 4 to 1 is not an assumption about communication but a way to fuse the auditory and motor branches inside each agent.

The end of a simulation is defined empirically, by when the agent productions for each object do not evolve further. In all the following simulations, 200,000 deictic games were sufficient to ensure convergence.

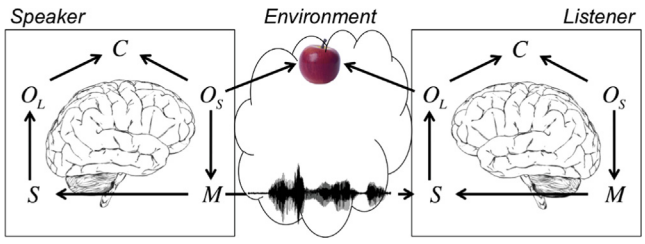


Fig. 5. Schema of two communicating agents involved in a deictic game.

5. Parity: conditions of speech code emergence

We now turn to the experimental simulations with COSMO. We argued in the introduction that there were three requirements for communication; namely, reference, parity and adequacy. The current section focuses on parity and the conditions for ensuring speech-code emergence. This constitutes the second contribution of our work.

5.1. Framework and assumptions

5.1.1. Objectives

Our aim is to compare three versions of the communicating agent model; namely, the motor, auditory and sensory–motor versions, as summarized in Table 1.

The first objective in this comparative evaluation of the three kinds of models is to test the emergence of efficient codes; i.e., codes that are similar for all agents of the society and well understood by all, enabling communication to take place in good conditions.

The second objective is to assess whether the predictions of "substance-based" theories – namely, the Dispersion and Quantal Theories (see Section 2.2) – might be derived from simulations as properties of the emerging code. That is, we compare the dispersion and quantal properties of the emergent codes with the predictions of the corresponding theories.

5.1.2. A simplified unidimensional articulatory-to-acoustic model

We consider a simple unidimensional instantiation of the model, which allows the extraction of the general properties of motor, auditory and sensory–motor behaviors. The motor and sensory variables, M and S, are defined as integer values in the range $[-10,\ldots,10]$. The articulatory-to-acoustic transformation TransMS : $M \rightarrow S$ is a sigmoid function parameterized in a way that allows changes from a purely linear function to a step function (see Fig. 6). It is defined as

$$\mathrm{TransMS}(m) = S_{max}\left(\frac{\arctan(NL\ (m-D))}{\arctan(NL\ M_{max})}\right), \tag{4}$$

where $S_{max} = M_{max} = 10$ according to the M and S range specifications.

5.2. Methodology

5.2.1. Agent specification

In each simulation, we consider a society of N agents interacting in an environment containing Q objects. Q is a fixed parameter, specified explicitly for each simulation.

Each agent implements a COSMO Bayesian model as defined by Eq. (1), in which three terms remain to be clarified; namely, the motor subsystem $P(M \mid O_S)$, the sensory–motor subsystem, $P(S \mid M)$ and the auditory subsystem $P(O_L \mid S)$. They are defined within regularly discretized motor and acoustic spaces.

$P(M \mid O_S)$ is a family of motor prototypes, associating with each object o_i a Gaussian probability distribution parameterized by its mean μ_i and standard deviation σ_i (the distributions are univariate because of the simplified sensory–motor space that we consider in this section).

$P(S \mid M)$ is a family of sensory–motor relationships, associating with each motor value m a Gaussian probability distribution parameterized by its mean TransMS(m) and standard deviation σ_{Ag}. We therefore consider that the agents already know the articulatory-to-acoustic transformation (assuming, for example, a previous sensory–motor exploration) with an uncertainty encoded by σ_{Ag}.

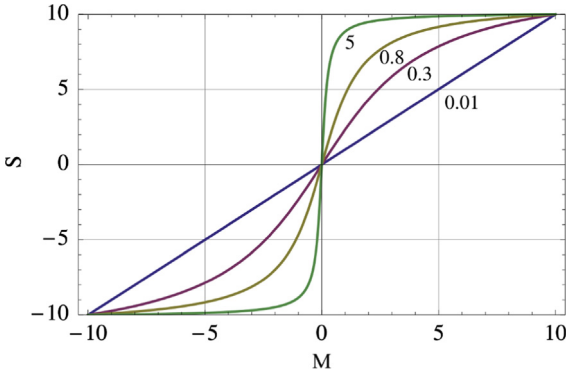


Fig. 6. Articulatory-to-acoustic transformation from Eq. (4). This sigmoid function, which maps motor gestures M to the resulting sensory stimuli S, is parameterized by NL, which allows changes from a purely linear function ($NL = 0.01$) to a step function ($NL = 5$). Parameter D sets the position of the inflection point ($D = 0$ for all curves here).

$P(O_L \mid S)$ is defined as a classifier using Gaussian auditory prototypes $P(S \mid O_L)$. Under the assumption of a uniform prior for the objects O_L, Bayesian inference yields

$$P([O_L = o_i] \mid S) = \frac{P(S \mid [O_L = o_i])}{\sum_{O_L} P(S \mid O_L)}. \tag{5}$$

Auditory subsystems are therefore parameterized by the means and standard deviations of the Gaussian auditory prototypes $P(S \mid [O_L = o_i])$.

In a given simulation, all agents implement the same version of the COSMO model: they are either motor, auditory or sensory–motor, as defined in Table 1.

Initially, the Gaussian prototypes $P(M \mid O_S)$ and $P(S \mid O_L)$ encode a complete lack of knowledge, with centered means and large standard deviations[3]. A simulation then comprises a series of deictic games. During a deictic game concerning object o_i, the simulated environment propagates the proposed motor gesture m according to TransMS, adding a simulated Gaussian noise with standard deviation σ_{Env}, resulting in an acoustic stimulus s. At the end of a deictic game, the speaker and listener agents, respectively, update the parameters of their motor and auditory Gaussian prototypes, according to the observed $\langle o_i, m \rangle$ pair for the speaker, and the $\langle o_i, s \rangle$ pair for the listener.

5.2.2. Evaluation

All along the course of simulations, we proceed to an evaluation to assess how the system evolves and possibly converges towards efficient communication. Evaluation is an independent process, different from the deictic games involved in training. During deictic games in training, communication always occurs as intended, with $O_L = O_S$. Therefore, since the reference requirement is supposed to be perfectly solved through deixis, the listener agent knows which object is designated by the speaker and does not try to recognize the object on the basis of the auditory stimulus S. However, in the evaluation process, we assess how communication occurs when it is no more ensured by deixis. This involves a perception stage for evaluation purposes. This allows computation of the evolution of the recognition rate in the society during a simulation. In this perception step, the listener agent infers the object designated by the speaker using its perception behavior, ignoring the "true" object identity provided by the nonverbal deixis behavior. The recognition rate in the society is obtained by the mean value of this perceptual inference process.

We also evaluate the ability of the simulation to allow the emergence of properties of the Dispersion and Quantal Theories described in Section 2.2.

To evaluate the Dispersion Theory, we compute the final dispersion of the auditory prototypes, using the Dispersion Theory formula (Liljencrants & Lindblom, 1972):

$$G = \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \left(\frac{1}{d_{i,j}} \right)^2, \tag{6}$$

where Q is the number of elements in the phonological system considered (here, the number of objects), and $d_{i,j}$ is the auditory distance between two elements i and j, taken as the distance between the means of auditory stimuli produced by the whole society in the last deictic game of a given simulation. In this formula, a lower G indicates a higher global dispersion for the emergent system. Note that this equation is used here only as a measure of the ability of COSMO simulations to enable dispersion to emerge from deictic games and is never part of the agents' knowledge.

To evaluate the Quantal Theory, we study the effects of the nonlinearity in the articulatory-to-acoustic transformation TransMS, with respect to both robustness toward environmental noise and the structure of the emerging code. We expect to observe an increase in robustness with sharp nonlinearities (which favor the existence of phonetic contrast) and a dependence of sound categories on the position of the nonlinearity, with one category on each side of the natural boundary provided by the nonlinearity.

5.3. Results

5.3.1. Conditions for emergence of a speech code

Fig. 7 shows the state of agent societies at the end of three typical simulations, one for each version of the COSMO model. In this first set of simulations, the articulatory-to-acoustic transformation TransMS is linear and the environmental noise σ_{Env} is low.

Fig. 7 (top) shows the results at the end of a simulation of "motor" agents; namely, a simulation whereby agents produce gestures according to the motor production behavior of Table 1. We observe that the motor prototypes of the agents have evolved relatively randomly during the learning process and that the recognition rate stays at the chance level (around 25%). Because the agents do not use auditory knowledge, and therefore have no way to link their productions in front of each object with the auditory inputs they receive from the others, there is no chance that such a society would enable the emergence of a structured and efficient speech code to designate objects.

Fig. 7 (middle) shows the results at the end of a simulation of "auditory" agents; namely, a simulation whereby agents produce gestures according to the auditory production behavior of Table 1. We observe that the motor prototypes for each object are well

[3] Infinitely large standard deviations would result from applying a uniform distribution, as used in the theoretical reasoning in Section 3.3. Here, however, we want these distributions to evolve progressively from a lack of knowledge toward acquired knowledge, thereby peaking these distributions. Therefore, the initial standard deviations are large but finite, and they typically decrease during simulations.

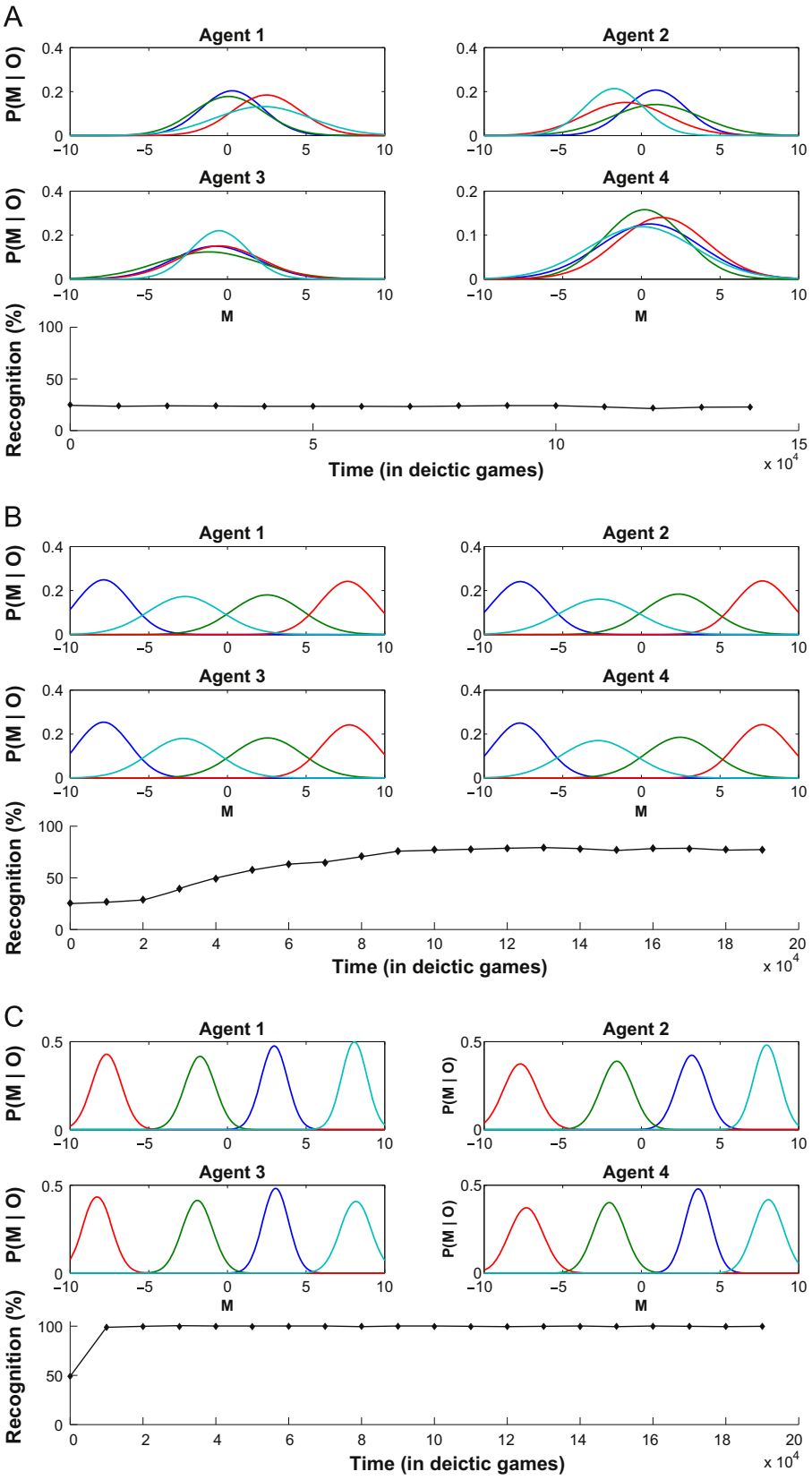


Fig. 7. Typical simulations of "motor" agents (A), "auditory" agents (B) and "sensory–motor" agents (C). Each simulation involves four agents, whose final motor prototypes are shown in the four panels. Each panel shows $P(M|[O_S = o_i])$ for each of the four possible objects o_i. The lower plots represent the evolution of successful communication rates as a function of simulated time.

contrasted and that an agreement between agents has emerged. Moreover, the recognition rate has increased during the simulation. Using auditory production behavior, agents prefer to produce gestures that result in auditory stimuli favored by their auditory subsystem; namely, those that have a high probability in the presence of o_i and a low probability in the presence of other objects (5).

Fig. 7 (bottom) shows the results at the end of a simulation of "sensory–motor" agents; namely, a simulation whereby agents produce gestures according to the sensory–motor production behavior of Table 1. We still observe a dispersion of motor prototypes for each object with an agreement between agents. In this case, however, prototypes are optimally distinguishable, leading rapidly to a 100% recognition rate. Compared with the auditory production behavior, adding the motor subsystem enables a reduction in the variability of the chosen motor gestures. This is caused by the $P(M \mid [O_S = o_i])$ factor in the sensory–motor speech-production behavior (see Table 1), which results in firmer anchoring of speech production behavior around selected prototypes.

Therefore, both auditory and sensory–motor agents can provide the conditions for the emergence of a speech code, with a higher level of communication efficiency for the latter case. However, motor agents do not provide the conditions for the emergence of a speech code that enables communication.

5.3.2. Communication and dispersion

In the following, we analyze only those behaviors that lead to the emergence of speech codes; namely, the auditory and the sensory–motor behaviors. The articulatory-to-acoustic transformation remains set for the linear case. The question here is to assess possible differences between auditory and sensory–motor agents in terms of convergence, communication and dispersion.

Fig. 8 shows how environmental noise σ_{Env} affects the recognition rate for both behaviors. We observe that the sensory–motor behavior is more robust, providing better communication scores all along the noise range. This is in line with the reduction of the variance of motor prototypes observed in Fig. 7.

Fig. 9 shows how environmental noise σ_{Env} affects prototype dispersion. We observe that the sensory–motor behavior increases dispersion (i.e. Lindblom's measure decreases) as the noise increases, until $\sigma_{Env} = 2$, which suggests that there could exist an optimal dispersion for this noise value. Above this level, the dispersion begins to decrease (i.e., Lindblom's measure increases) for both models. Globally, once again, the sensory–motor model, producing more dispersion than the auditory one, appears to lead to more efficient communication systems.

We next analyze the stability condition for our simulations; i.e., the time horizon when the agents' learning processes cease modifying their motor prototypes. Consider, for simplicity, the case of a linear articulatory-to-acoustic transformation, without noise, where $s = \text{TransMS}(m) = m$. Starting from Eq. (2), with $P(M \mid O_S)$ being uniform and $P(S \mid M)$ being a Dirac distribution with $S = M$, the

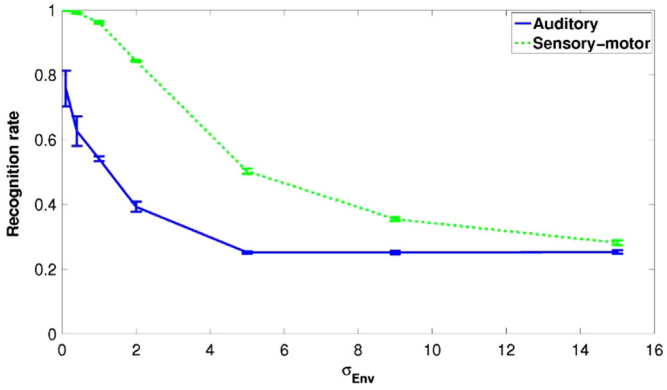


Fig. 8. Recognition rate as a function of the environmental noise σ_{Env} at the end of simulations of auditory and sensory–motor agents. Ten independent simulations were run for each condition, for which means and standard deviations were computed. Each involved four agents and four objects. For all simulations, σ_{Ag} was set to a low value (0.1).

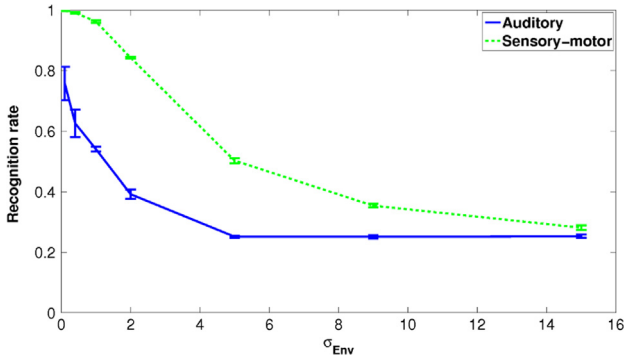


Fig. 9. Logarithm of Lindblom's dispersion as a function of the environmental noise σ_{Env}. Low values correspond to high dispersions, as defined by Eq. (6). Conditions are the same as for Fig. 8.

auditory production behavior can then be rewritten as

$$P(M \mid [O_S = o_i][C = 1]) \propto \frac{P(S \mid [O_L = o_i])}{\sum_{o_j \in O_L} P(S \mid [O_L = o_j])}.$$

Stability occurs when the successive motor gestures drawn according to this distribution do not change the auditory prototypes $P(S \mid O_L)$, which are the only updatable terms in this behavior. This happens when

$$\sum_{o_j \in O_L} P(S \mid [O_L = o_j]) \propto \mathbf{U}(S), \tag{7}$$

where $\mathbf{U}(S)$ is the uniform probability distribution over S.

Intuitively, a solution would be to distribute the prototypes to cover the whole auditory space and therefore to disperse them as observed in Fig. 7 (middle).

A similar kind of derivation as that leading to Eq. (7) shows that the sensory–motor production behavior tends to select motor gestures in regions that produce sensory stimuli leading to the clear identification of objects; i.e., regions where the auditory classifier of Eq. (5) gives distinguishable results.

In summary, simulations with both auditory and sensory–motor agents naturally result in the optimization of auditory dispersion in the sense defined by Lindblom (Liljencrants & Lindblom, 1972). However, sensory–motor populations result in higher communication efficiency and higher resistance to noise compared with auditory agents. This seems to be due to a reduction in the prototype variance (that is, motor prototypes are more precisely defined since they are better anchored by the motor component).

5.3.3. Quantal aspects

To assess the effects of the nonlinearity in the articulatory-to-acoustic transformation TransMS of Eq. (4), we now consider simulations with two objects. Fig. 10 shows how the nonlinearity strength affects the recognition rate for high environmental noise. We observe that it considerably improves the auditory behavior performance, because the nonlinearity reduces the variance of the auditory prototypes in stable regions, thereby making them more distinguishable. With respect to sensory–motor agents, there is no such improvement because the variances are already low in this case.

Fig. 11 shows how a nonlinearity shapes the emerging code, with respect to Stevens' Quantal Theory. To address our aim, we consider a strong nonlinearity ($NL = 5$) and vary the position D of the inflection point in the motor space (see (4)). That is, we vary the position of the unstable region.

We observe that motor gestures chosen by agents are strongly correlated with the position of the inflection point D, as predicted by the Quantal Theory, in that they occupy both sides of the nonlinearity boundary. Such a structure will indeed make the code efficient, and both the auditory and sensory–motor behaviors are able to use their auditory subsystems to let such an efficient structure emerge.

Moreover, whereas variances of motor prototypes are almost null for the auditory agents, they are much larger when considering sensory–motor agents. The reason is that the sensory–motor behavior tends to produce sensory stimuli which are dispersed enough to allow the communication success. All the motor gestures in the stable regions are therefore good candidates, hence the variability across various simulations, especially observed in large stable regions. It is not the case when considering the auditory behavior, which rather tends to produce stimuli optimally spanning the whole sensory space (see (7) and the corresponding analysis).

In summary, these simulations are in line with the major predictions of the Quantal Theory: nonlinearity improves communication and shapes the selected sound system, producing a natural sensory boundary exploited by the system with one prototype on each side of the boundary.

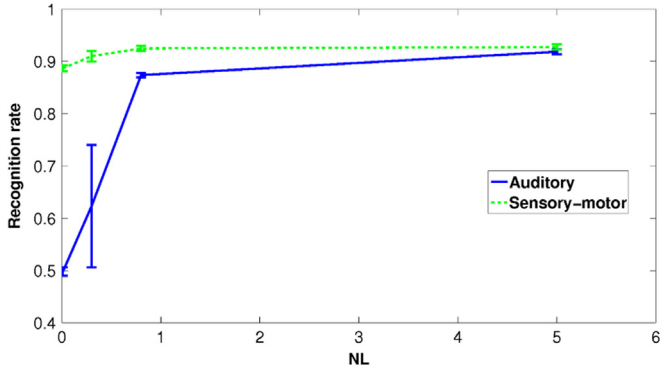


Fig. 10. Recognition rate as a function of the nonlinearity strength NL, at the end of simulations of auditory and sensory–motor agents. Ten independent simulations were run for each condition, from which means and standard deviations were computed. Each involved four agents and two objects. In all simulations, σ_{Env} was set to a rather high value (5), and σ_{Ag} was set to a low value (0.1).

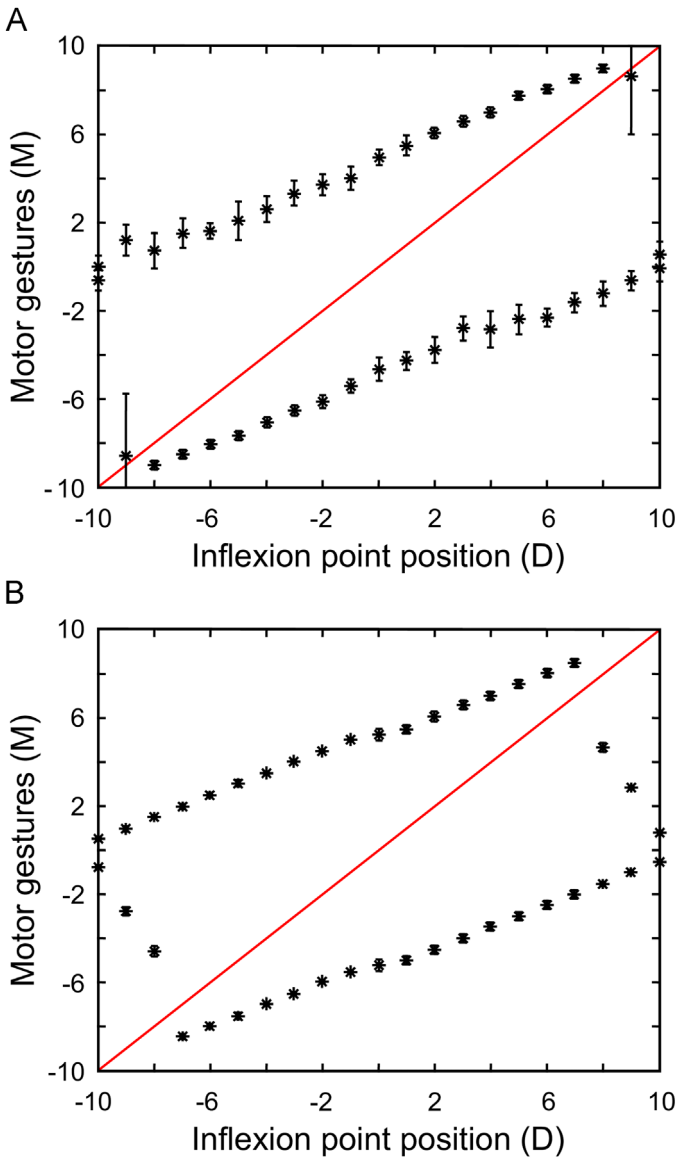


Fig. 11. Motor gestures at the end of simulations with four agents and two objects, as a function of the position *D* of the inflection point of TransMS. (A) Simulations of auditory agents. (B) Simulations of sensory–motor agents. For each value of *D*, 10 independent simulations are performed, each one providing a mean motor gesture for each object, computed over the whole society at the end of the simulation; then the mean and standard deviation of these 10 values is displayed on the figure. The identity function shows the inflection point position (*x*-axis) in the motor space (*y*-axis).

5.4. Conclusion

The following results emerge from these simulations.

5.4.1. Sensory–motor parity is required for the emergence of a speech code

The first and major lesson of these simulations is the impossibility of a speech code emerging from a society of motor agents. The only difference between motor and sensory–motor agents is that for the former, the auditory value of gestures is never evaluated in terms of the efficiency of allowing the communicating partner to understand the message (see Table 1). This shows that with respect to our interaction paradigm, a sensory–motor link is needed for the emergence of efficient speech codes. This validates the parity requirement. Without a capacity to associate what they hear from others with their own motor gestures and to select adequate gestures in relationship with the sensory information they provide to their interlocutor, agents are unable to converge to a conventional code. This can be related to a remark by Schwartz, Boë, and Abry (2007), who pointed out that motor theories do not provide any prediction about phonological systems in human languages, which is a serious limitation on these theories. The crucial point there is that the characterization of the perceptual nature of speech gestures is completely lacking in the motor theory, and this produces a weakness which has been often discussed and debated, but which becomes insuperable when dealing with the prediction of sound systems in human languages. One can attempt to add a perceptuo-motor component in motor theories – see e.g.

Studdert-Kennedy and Goldstein (2003). But this implicitly modifies the very concept of gesture, which becomes a perceptuo-motor rather than a pure motor concept, and therefore deeply changes the Motor Theory itself.

A second important, and less expected, result is that the sensory–motor behavior provides agents with increased robustness with respect to distinguishability and environmental noise. We found that all simulations with sensory–motor agents outperformed those with auditory agents in terms of global communication efficiency.

5.4.2. Dispersion and quantal theories emerge from deictic games in COSMO

A third result concerns the ability of these simulations to allow the emergence from the deictic game paradigm of regularities associated with the Dispersion and Quantal Theories.

With respect to dispersion, we showed that it might appear as a condition for convergence in the case of auditory agents. We also observed that dispersion may be reduced if communication does not require it to be high, such as with low noise or with a strong reduction of variance using sensory–motor agents. This is in line with further refinement of the Dispersion Theory, in which Lindblom proposed that dispersion might be "sufficient" instead of "maximal", and even "adaptive" to the conditions of communication (Lindblom, 1986, 1990).

With respect to quantal behaviors, they are obtained as a consequence of nonlinearities in the articulatory-to-acoustic transformation, for both auditory and sensory–motor agents, and they result in vocal codes that naturally exploit optimal sensory–motor "niches" provided by these nonlinearities, with the consequence that communication is globally favored relative to linear situations.

These two theories can therefore be derived from communication in COSMO, provided that parity is ensured by a sensory–motor link in auditory or sensory–motor agents. In this way, a basic level of adequacy is obtained in these simulations. The next section describes how more realistic articulatory–auditory environments constrain adequacy in a richer way, leading to accurate predictions of sound systems for human languages.

6. Adequacy: realistic predictions of phonological systems

We now consider a final series of COSMO simulations, in which we further specify the model and compare its predictions with actual data about sound systems for human languages, as presented in Section 2. This will constitute our third contribution.

For this aim, we capitalize on the results of the previous section that showed that motor agents do not let an efficient communication system emerge from deictic games. Furthermore, while both auditory and sensory–motor agents lead to coherent systems, sensory–motor agents are more efficient in terms of dispersion, convergence speed and robustness to noise. We will therefore continue by considering only the sensory–motor version of COSMO, for which the parity requirement is solved most efficiently.

In the simulations of the previous section with sensory–motor agents, parity was achieved via the internal link between motor M and auditory S variables inside the agent's architecture. Reference was ensured by deixis. However, no proposal about adequacy has yet been introduced. The objective in the current section is to make precise proposals about perceptuo-motor adequacy, to derive quantitative simulations from these proposals and to assess these simulations in the light of quantitative data about phonological systems. The evaluation of these simulations will therefore be based on an analysis of the structures of emergent phonological systems and systematically compared with the properties of human sound systems presented in Section 2.

We still consider that "objects" are directly phonological, in that we do not consider here the higher levels of language (words and morphemes) and their combination into syntactic and semantic sequences. In short, objects are phonological units. In this section, we consider three major types of units; namely, vowels, stops and syllables.

6.1. Framework and assumptions

6.1.1. Adequacy in the frame-content theory of speech emergence

Our basis for the generation of adequate sensory–motor stimuli lies in the "Frame-Content Theory" (FCT) (MacNeilage, 1998), which suggests an evolutionary path from ingestive mandibular cyclicities to articulated speech. According to FCT, speech would have exploited the ability to produce orofacial movements arising out of the cyclical use of the jaw during ingestion. A crucial point is that cyclical jaw movements naturally produce alternations between open and closed positions of the vocal tract, which result in highly contrasted acoustic configurations with consonants linked to the closed periods with high jaw, and vowels to the open periods with low jaw. This ensures the auditory adequacy of the articulatory modulation. This first stage, which would provide the basis for precursors of syllables, is called the "frame". At this stage, the open and closed portions are not independent, being characterized by similar tongue and lip settings while only the jaw varies, thereby producing the frame. The "content" comes next, in which other articulators (such as tongue and lips) become controlled to alternate between independent vowels and consonants. Interestingly, this "frame-then-content sequence" is also present during language acquisition, where babbling (a jaw cycle coupled with simple vocalizations) is a necessary step during child development, with the later development of more complex control involving the other articulators (MacNeilage & Davis, 2000, 2001).

In computational terms, FCT constrains the way that speakers control their vocal tract (i.e., constrains the variations in *M*) and the way that consonants and vowels are coordinated, beginning with a strong dependence (frame) and evolving progressively toward independence (content).

6.1.2. Realistic motor variables: the Variable Linear Articulatory Model (VLAM)

We use a realistic computer model of the vocal tract called VLAM (Boë, 1999), derived from the Maeda (1989) articulatory model. This latter model was conceived from a statistical analysis of 519 vocal-tract sagittal contours, obtained from radiographic measurements and tomographic studies of sentences pronounced by a French speaker. These contours were segmented into 28 sections, from the glottis to the lips, from which the corresponding vocal-tract areas were calculated. This analysis provided seven parameters that explained 88% of the data variance and that may be interpreted in terms of phonetic commands corresponding to the jaw (*J*), tongue body (*TB*), tongue dorsum (*TD*), tongue tip, lip protrusion, lip separation height (*LH*) and larynx height. These parameters can be linearly combined to reconstruct the sagittal contour and the area function, from which formants and a transfer function can be computed. Sound is then generated from the formant frequencies and bandwidths.

Each configuration can also be characterized by the so-called constrictive information from the area function, comprising the minimal area section (constriction position X_c), the minimal area size (constriction area A_c) and the area of the lip section (lip area A_l).

To simplify the simulations, we use only four of the motor parameters – namely, *J*, *TB*, *TD* and *LH* – with the other parameters set to a neutral position. This approximation is sufficient to generate most of the possible variability in vowels and stops.

6.1.3. Realistic sensory variables: formants and weighted Euclidean distances

As mentioned previously, the nature of auditory representations of speech sounds was debated widely in the 1970s and 1980s. Despite the difficulty of extracting formants automatically from spectra, several perceptual experiments have pointed out the specific role of formants in the determination of vowel quality (for example, see Carlson, Granström, & Klatt, 1979; Klatt, 1982), and most simulations of phonological systems have considered formants as an adequate characterization of the auditory input. Formants are then expressed on a perceptual scale, typically linear at low frequencies and logarithmic at high frequencies. We use the Bark scale defined according to the formula proposed by Schroeder, Atal, and Hall (1979) and used in many studies on auditory perception.

Then, because we describe probability distributions in terms of Gaussian laws, we need to define auditory distances. Classically, the Euclidean distances between the formants *F*1, *F*2 and possibly *F*3 of two sounds are considered as providing an adequate representation of the perceptual distance between the two sounds. However, unweighted distances, which give the same importance to all formants, do not take into account spectral masking phenomena, according to which low-frequency components decrease the perceptual role of higher-frequency components. This led to the proposal (Schwartz et al., 1997a) that *F*1 should have typically three times the weight of an “effective second formant” grouping of the roles of *F*2 and *F*3, and that in the computation of this effective second formant *F*2 would be twice more important than *F*3. That is, *F*1 would have three times the weight of *F*2 and six times the weight of *F*3. These weights were shown to produce the best agreement between simulations of vowel systems in human languages and the major trends in vowel system inventories.

In the present study, instead of systematically varying the weights in search of the best simulations, which could be very costly considering the dimensionality of our models, we will systematically compare two kinds of distances; namely, unweighted distances for *F*1, *F*2 and *F*3 in Barks and weighted distances (with relative weights (1), (1/3) and (1/6) for *F*1, *F*2 and *F*3 in Barks, respectively).

Note that while FCT is proposed as an explicit landmark in human evolution toward language, we do not assume that there is any specific discontinuity in motor and auditory processes between nonhuman and human primates. For auditory variables, this is widely accepted (Pickles, 2012). For vocal tract anatomy and control, it is more widely debated (for example, see Lieberman, 1984, 2012 vs. Boë et al., 2007, 2013), but this is beyond the scope of the present paper.

Taken together, VLAM and FCT, in addition to the corresponding choices of auditory parameters and distances, enable specification of the sensory–motor COSMO agents from which all simulations of vowel, stop and syllable systems are derived, as described in the following sections.

6.2. Vowel systems

6.2.1. Objectives

In this first series of simulations, we consider vowel systems. Changing the unidimensional articulatory-to-acoustic model of the previous section to the multidimensional motor and sensory spaces of VLAM produces a large increase in simulation complexity. Therefore, a preliminary objective is to check that simulations of COSMO agents in the VLAM framework are tractable.

We study two types of vowel systems, involving either three or five vowels. Data about human languages presented in Section 2 show that three-vowel systems are most often of the type /i,a,u/, and five-vowel systems are most often of the type /i,e,a,o,u/ (statistics for UPSID, Maddieson, 1984; Vallée, 1994; Schwartz et al., 1997b). Therefore, one objective is to check that these systems emerge from COSMO simulations in VLAM.

For this objective, the relative dimensions of the sensory space must be specified adequately. As mentioned above, we shall compare two versions; namely, with equal weights for *F*1, *F*2 and *F*3 and with relative weights (1), (1/3) and (1/6), respectively, for *F*1, *F*2 and *F*3.

A final objective is to test how dispersion depends on communication noise, to extend the results obtained in the previous section with the unidimensional model.

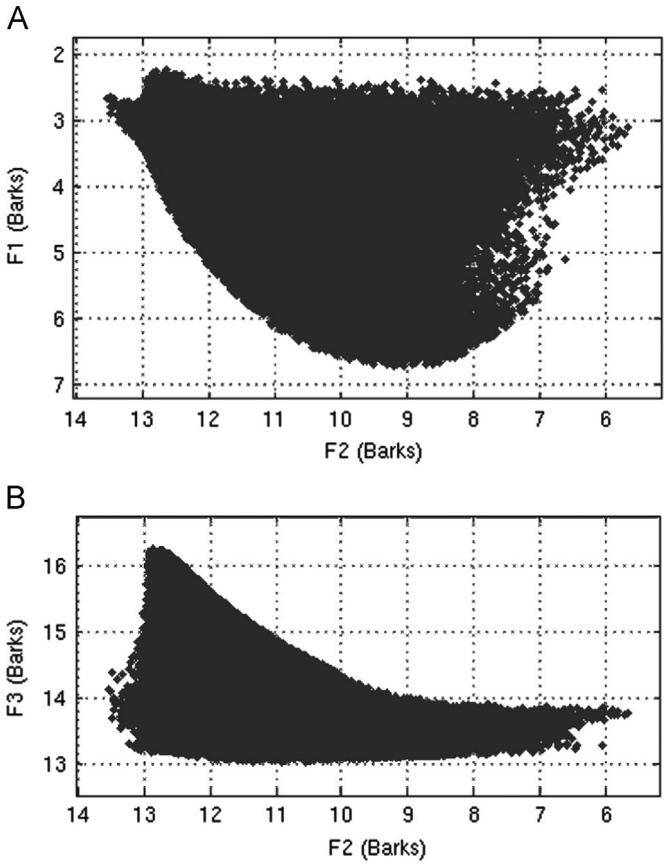


Fig. 12. A dictionary of motor configurations with the jaw in a neutral position yields an acoustic space. (A) Shown as a projection onto the (*F*1, *F*2) plane. (B) Shown as a projection onto the (*F*2, *F*3) plane. The point clouds are 100,000 motor configurations randomly drawn to correspond to open configurations, which we consider as vowels.

6.2.2. Methodology

Agent specification: The motor variables M are discretizations of the (J, TB, TD, LH) parameters, and the sensory variables S are discretizations of the ($F1$, $F2$, $F3$) parameters (in Barks). Because all variables are discrete, we can reduce the information loss by a probabilistic definition of the articulatory-to-acoustic transformation. Therefore, we randomly draw a large number of values in the motor space and compute, using VLAM, the corresponding values in the acoustic space, thereby compiling a dictionary. Then, for each motor command region, we compute the (supposedly) unimodal distribution of the related acoustic stimuli in the space of the first three formants, using the dictionary. This provides a $P(S \mid M)$ conditional distribution, modeling the articulatory-to-acoustic transformation performed by the environment, that we use to infer the sensory stimulus heard by the listening agent, given the motor gesture produced by the speaking agent during a deictic game.

To study the emergence of vowel systems, we consider only open configurations of the vocal tract. To do this, we compute probability distributions from a dictionary for which the jaw parameter J is fixed to a neutral position and where both A_c and A_l are greater than $0.15\,\mathrm{cm}^2$. We therefore use three of the articulatory parameters in this section; namely, TB, TD and LH (the jaw parameter J will be used for the consonant and syllable simulations in subsequent sections). The acoustic space covered by motor configurations in the vowel dictionary is shown in Fig. 12. Notice that the limitation induced by fixing the jaw parameter J to zero does almost not change at all the available acoustic range, and particularly does not decrease the available $F1$-range: TB and TD are sufficient to enable tongue configurations with all $F1$ values for vowels.

Probability distributions: Let us recall that in COSMO, three subsystems have yet to be defined; namely, the motor, sensory–motor and auditory subsystems.

The motor subsystem is expressed as

$$\begin{aligned} &P(M \mid O_S) \\ &\quad = P(TB\ \ TD\ \ LH \mid O_S) \\ &\quad = P(TB \mid O_S)P(TD \mid O_S)P(LH \mid O_S). \end{aligned}$$

This last equality expresses a conditional independence hypothesis; namely, that TB, TD and LH are considered independent when the object identity O_S is given. Each term of this product is defined as a univariate Gaussian distribution, initialized with a centered mean and large variance and then learned via simulation data. Learning of these distributions occurs when an agent is a speaker during a deictic game, as detailed in Section 4.2.

The sensory–motor subsystem $P(S \mid M)$ corresponds here to $P(F1\ F2\ F3 \mid TB\ TD\ LH)$ and is learned from the dictionary before the simulation starts (thereby considering, as for the unidimensional simulation, a previous sensory–motor exploration).

Finally, the auditory model $P(O_L \mid S)$ is expressed as a function of the auditory prototypes $P(S \mid O_L)$ (see (5)), which correspond here to $P(F1\ F2\ F3 \mid O_L)$ probability distributions. As for the motor model, they are products of learned univariate Gaussian distributions $P(F_i \mid O_L)$, where data are collected by the agent from its experience as a listener in deictic games, as opposed to the speaker's experience for the motor subsystem.

Deictic games: In each deictic game, the speaker agent emits a single vowel, for which, as mentioned above, we make no distinction between semantic and phonological categories. We study the emergence of three- and five-vowel systems by running simulations with either three or five objects. To shorten the convergence time, we use two agents running 150,000 deictic games in each simulation.

The simulation parameters also include the simulated environmental noise in each formant dimension, such that $\sigma_{Env} = (\sigma_{F1}, \sigma_{F2}, \sigma_{F3})$. It is important to note the interpretation of these parameters: the more significant the noise in a given auditory dimension, the less useful will this dimension be for generating dispersion and ensuring efficient communication. We therefore use the noise parameters to control the relative significance of each auditory dimension in the simulation. In a first set of simulations (called "1–1–1"), we use equal values for σ_{F1}, σ_{F2} and σ_{F3}, which means that all formants (in Barks) have the same significance in the simulations. However, in line with the proposal that $F1$ should have a greater weight and $F3$ a smaller one (Schwartz et al., 1997a), we also consider a second set of simulations (called "1–3–6"), in which $\sigma_{F2} = 3\sigma_{F1}$ and $\sigma_{F3} = 6\sigma_{F1}$, which means that $F1$ has three times more weight than $F2$ and six times more weight than $F3$. For both of these sets, the value of σ_{F1} is used to control globally the

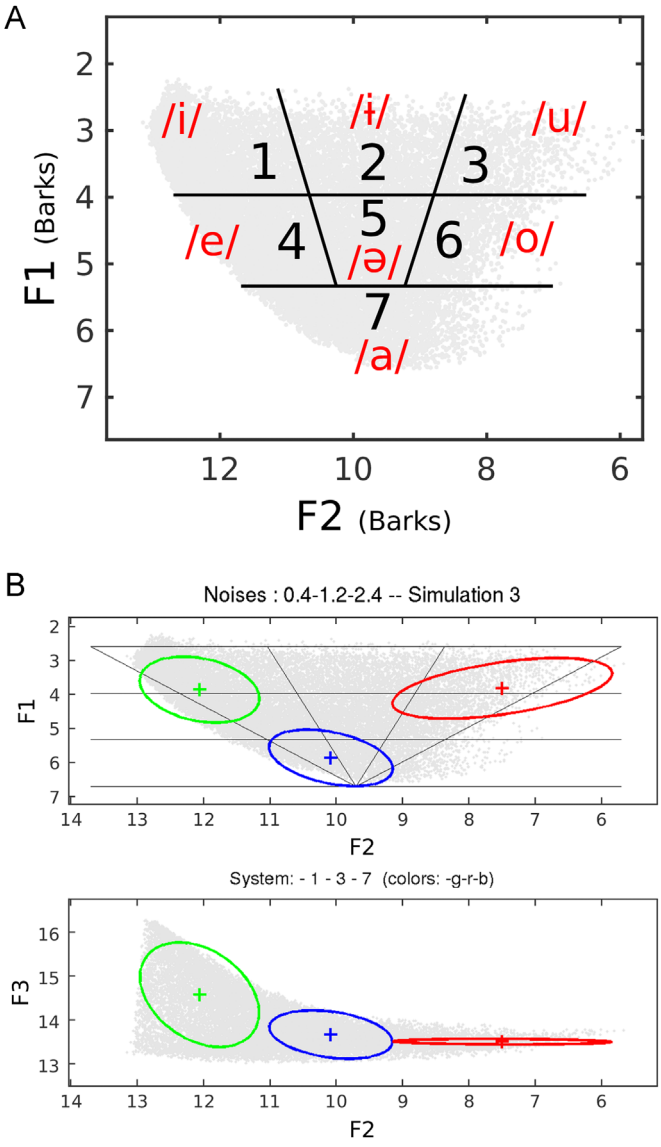


Fig. 13. Vowel classification. (A) The classification system used to process simulation results automatically in the ($F1$, $F2$) formant space. Each of the seven grid cells displays the corresponding phonetic symbol. (B) An example of a three-vowel simulation that converged to a [1 3 7] system, corresponding to an /i,a,u/ vowel system. The caption at the top of the panel indicates the noise values, in Barks, for each formant dimension. The caption in the middle of the panel maps the system elements (here, the vowels) to the ellipse colors. In this particular example, it means that the vowels /i/, /u/ and /a/ (classification code 1–3–7), respectively, correspond to the green, red and blue ellipses (color: g–r–b). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

communication noise level, to test its role in relation to dispersion.

Evaluation: Evaluating phonological systems in emergence simulations is a complicated task, usually done manually (Berrah et al., 1996; de Boer, 2000; Oudeyer, 2005). Here, however, we classify emergent vowel systems automatically, using a grid in the ($F1$, $F2$) formant plane, as shown in Fig. 13 (top). This grid defines seven categories. For a given simulation, the vowel associated with an object is classified according to this grid, using the mean in ($F1$, $F2$), of the auditory stimuli produced by the whole society for that object. This is performed during the last of the deictic games, before addition of the environmental noise. Fig. 13 (bottom) shows an example of a system with three objects, after 150,000 deictic games, in the "1–3–6" set with a baseline noise $\sigma_{F1} = 0.4$. Here, there was convergence to a [1 3 7] system in our classification scheme. That is, to designate objects, agents in this simulation would use vowels that approximate an /i,a,u/ vowel system.

Although using only seven categories might appear simplistic, it is sufficient for our goal of evaluating the global coherence of our results in terms of world-language data and is compatible with the manual notation used in comparable studies (Berrah et al., 1996; de Boer, 2000; Oudeyer, 2005). Note that with seven categories, there are 84 (or 462) possible systems for three-object (or five-

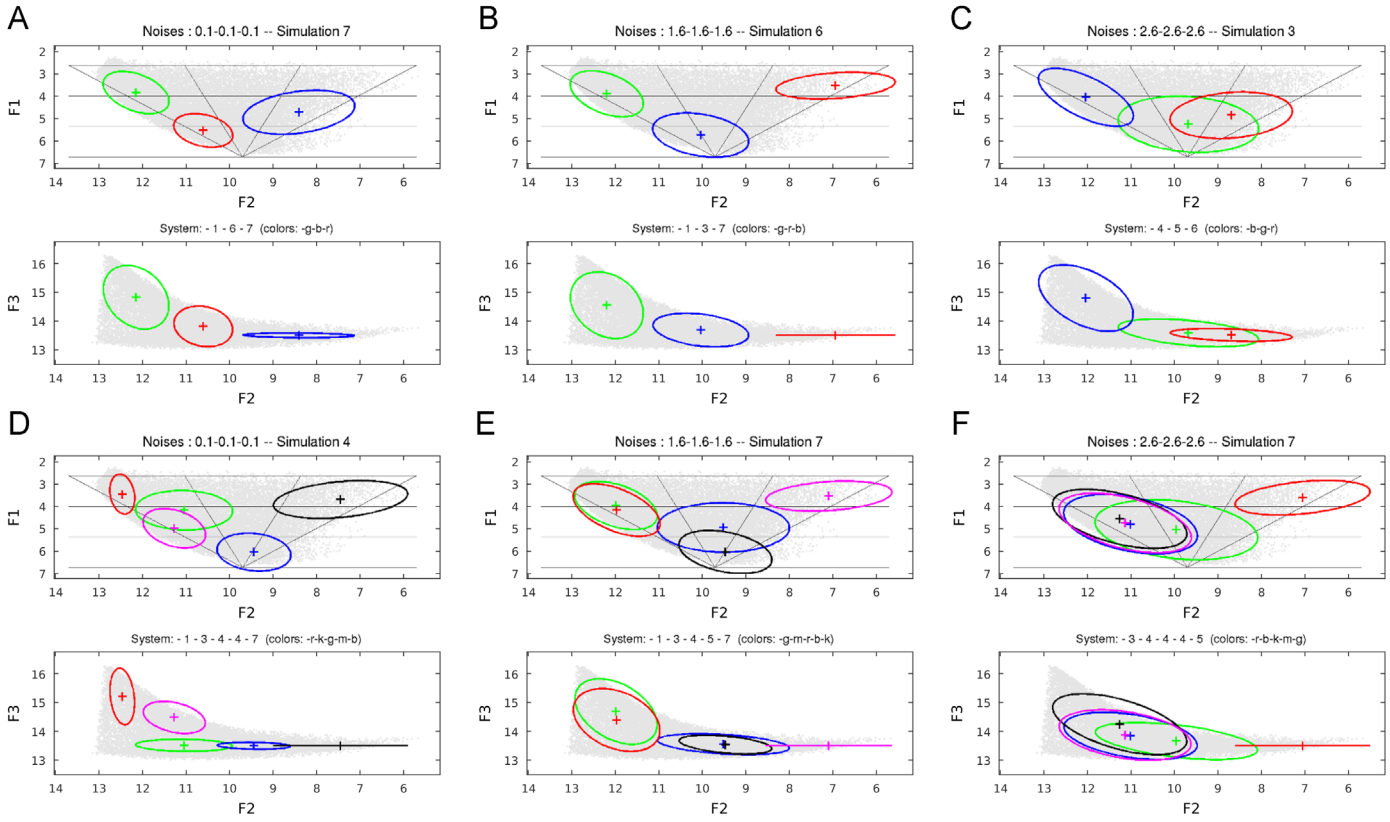


Fig. 14. Typical systems obtained in the "1–1–1" noise ratio simulations. Same convention as in Fig. 13(B). Top line (A–C): three-vowel simulations. Bottom line (D–F): five-vowel simulations. Left column (A, D): low noise level, with $\sigma_{F1} = \sigma_{F2} = \sigma_{F3} = 0.1$. Center column (B, E): medium noise level, with $\sigma_{F1} = \sigma_{F2} = \sigma_{F3} = 1.6$. Right column (C, F): high noise level, with $\sigma_{F1} = \sigma_{F2} = \sigma_{F3} = 2.6$.

Table 2
Simulation results for the "1–1–1" noise ratio experimental condition, in the three-vowel case. For each base noise value (σ_{F1}), 10 independent simulations resulted in 10 final vowel systems. The columns sum to 10.

System	Value of σ_{F1}								
	0.1	0.6	1.1	1.6	2.1	2.6	3.1	3.6	4.1
/i,a,o/	5	3	5		1				
/i,a,e/	3	3					1		
/e,a,o/	2	1			1	1			
/i,a,u/		**3**	**5**	**9**	**5**				
/e,a,u/				1	2	2			
/e,e,o/					1	2	7		
/e,ə,o/						3	2		
/i, ə, o/						2			
/e,ə,ə/								5	3
/e,e,ə/								5	6
/e,e,e/									1

Table 3
Simulation results for the "1–1–1" noise ratio experimental condition, in the five-vowel case. For each base noise value (σ_{F1}), 10 independent simulations resulted in 10 final vowel systems. The columns sum to 10 except when, for clarity, systems that would appear only once have been omitted.

System	Value of σ_{F1}								
	0.1	0.6	1.1	1.6	2.1	2.6	3.1	3.6	4.1
/i,e,a,a,u/	4	1		2	2				
/i,e,e,a,u/	4	5	6	2					
/i,e,a,o,u/	**1**	**1**	**3**						
/i,e,a,ə,u/	1		1	5	2				
/i,i,e,a,u/		3		1					
/e,e,e,a,u/					2				
/e,e,e,ə,u/						2			
/e,e,ə,ə,o/						2			
/e,e,e,ə,o/						1	1		
/e,e,e,ə,ə/							7	5	3
/e,e,e,e,ə/							1	5	5

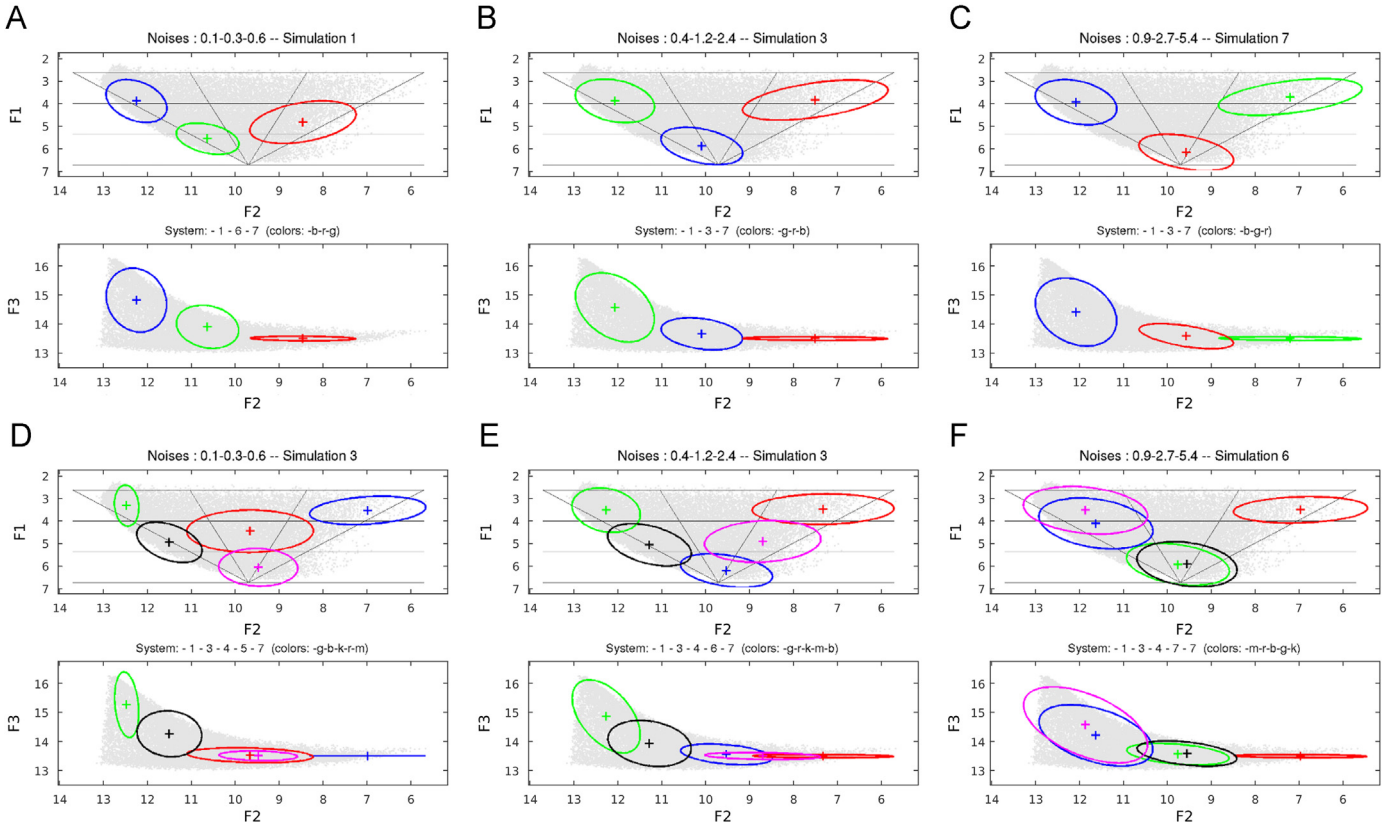


Fig. 15. Typical systems obtained in the "1–3–6" noise ratio simulations. Same convention as in Fig. 13(B). Top line (A–C): three-vowel simulations. Bottom line (D–F): five-vowel simulations. Left column (A, D): low noise level, with $\sigma_{F1}=0.1, \sigma_{F2}=0.3, \sigma_{F3}=0.6$. Center column (B, E): medium noise level, with $\sigma_{F1}=0.4, \sigma_{F2}=1.2, \sigma_{F3}=2.4$. Right column (C, F): high noise level, with $\sigma_{F1}=0.9, \sigma_{F2}=2.7, \sigma_{F3}=5.4$.

object) simulations.

6.2.3. Results

Noise ratios "1–1–1": For this noise ratio condition, with the same Gaussian noise $\sigma_{F1}=\sigma_{F2}=\sigma_{F3}$ in each formant dimension, σ_{F1} varies from 0.1 to 4.1 Barks in increments of 0.5, with nine possible values altogether. We capture simulation variability by running 10 independent simulations, for each of these nine noise values. We study both three-vowel and five-vowel systems, for a total of 180 experimental simulations of 150,000 deictic games. Each simulation leads to a final vowel system, characterized in our classification system by an [x y z] triplet and presented in the following results as a vowel triplet.

Typical systems obtained for three-object and five-object simulations are shown in Fig. 14, for low ($\sigma_{F1}=0.1$ Barks), medium (1.6) and high (2.6) noise values. The distribution of obtained systems over the 10 simulations per condition is reported in Tables 2 and 3.

Experimental results, both for three-vowel and five-vowel simulations, reproduce the variation of dispersion that we have already observed in Section 5.3.2. At low noise values, a small dispersion suffices to make items distinguishable, such as /i,a,o/, /i,a,e/ and

Table 4
Simulation results for the "1–3–6" noise ratio experimental condition, in the three-vowel case. For each base noise value (σ_{F1}), 10 independent simulations resulted in 10 final vowel systems. The columns sum to 10.

System	Value of σ_{F1}								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
/i,a,o/	5	8	6	2					
/i,e,o/	5								
/e,a,o/		1	1						
/i,a,u/	**1**	**3**	**7**	**7**	**10**	**6**	**9**	**9**	
/e,a,u/			1	3		4	1		
/i,ə,a/									1

Table 5
Simulation results for the "1–3–6" noise ratio experimental condition, in the five-vowel case. For each base noise value (σ_{F1}), 10 independent simulations resulted in 10 final vowel systems. The columns sum to 10.

System	Value of σ_{F1}								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
/i,e,a,a,u/	2	3					4	7	7
/i,e,e,a,u/					1	2	1	1	
/i,e,a,o,u/	**3**	**4**	**7**	**9**	**7**	**5**			
/i,e,a,ə,u/	5	3	2	1	1	3	5		
/i,i̯,e,a,u/			1						
/i,ə,a,a,u/					1				
/i,i,a,a,u/								2	3

/e,a,o/ for three-vowel systems. At intermediate noise values, the dispersion has to be higher. Finally, at high noise values, agents no longer converge to an efficient system, and dispersion decreases, such as /e,ə,ə/ or /e,e,ə/ for three-vowel systems.

We further observe that systems with suboptimal dispersion dominate at low noise levels, whereas the optimally dispersed system /i,a,u/ prevails at medium noise levels in three-vowel simulations. However, /i,e,a,o,u/ is poorly represented in five-vowel simulations. The reason is that the application of an equal noise value to each formant allows good use of the *F*3 dimension to be made. This results in other vowel systems, such as /i,e,e,a,u/ for $\sigma_{F1} = 0.1$, exploiting the dispersion in *F*3 rather than in (*F*1, *F*2) (see Fig. 14, bottom, left). Note that this is not in agreement with data from UPSID.

Noise ratios "1–3–6": Because the five-vowel simulations with a noise ratio "1–1–1" are not coherent with the preferred human system /i,e,a,o,u/, we now consider the "1–3–6" noise ratio set. In this experimental condition, the standard deviation σ_{F1} varies from 0.1 to 0.9 Barks in steps of 0.1, with $\sigma_{F2} = 3\sigma_{F1}$ and $\sigma_{F3} = 6\sigma_{F1}$.

Typical systems obtained for three-object and five-object simulations are shown in Fig. 15, for low ($\sigma_{F1} = 0.1$ Bark), medium ($\sigma_{F1} = 0.4$) and high ($\sigma_{F1} = 0.9$) noise values. The distribution of systems obtained over the 10 simulations per condition is reported in Tables 4 and 5. Note that the three-vowel simulations still yield the human-preferred three-vowel system /i,a,u/ for a large range of medium noise values.

In contrast to the "1–1–1" set, a medium noise level in the *F*1 dimension corresponds to a high noise value in the *F*3 dimension, which therefore cannot be used efficiently. For this case, dispersion occurs mainly in the (*F*1, *F*2) plane. Moreover, because *F*1 is less noisy than *F*2, items are more able to disperse in *F*1 than in *F*2. We observe that the preferred five-vowel system in world languages, /i,e,a,o,u/, emerges widely in our simulations.

6.2.4. Conclusion

Simulations with COSMO in the VLAM framework with three-dimensional motor and sensory variables appear tractable. The fact that emergent three-vowel and five-vowel systems are in line with data from UPSID is not surprising, considering that similar results had already been obtained in previous work (for example, see Berrah et al., 1996; de Boer, 2000; Oudeyer, 2005). Of interest in these simulations is their ability to specify the range of experimental parameters that lead to realistic predictions of vowel systems. We can identify two major points.

Firstly, while the /i,a,u/ system emerges for both the "1–1–1" and "1–3–6" noise conditions, only the "1–3–6" set obtains the five-vowel /i,e,a,o,u/ system, which is consistent with previous works (Schwartz et al., 1997a). This ratio will therefore be used for the next set of simulations.

Secondly, as for the unidimensional model in Section 5.3.2, the sensory–motor behavior in COSMO yields "adaptive" rather than "maximal" dispersion, with items being just sufficiently dispersed in the acoustic space that correct distinctions occur. Dispersion is function of the noise level and the number of objects. A medium value of σ_{F1} (around 0.4) produces both maximal dispersion and the most realistic simulations.

We now extend these investigations to involve less-studied phonological units, for which very few computational simulations are available; namely, stops and syllables.

6.3. Stop consonant emergence

6.3.1. Objectives

Whereas simulations of vowel systems have been widely studied, almost no quantitative simulation of the emergence of consonant systems exists in the literature. This is the focus of the present section, where we consider stops, which involve almost-closed configurations of the vocal tract. The objective is to obtain simulations of three-stop systems in which /b,d,g/ emerges as a favored system, in agreement with the statistics from UPSID presented in Section 2 – though recall our remark in Section 2 about voiced vs. unvoiced plosives: /p,t,k/ is actually the favored system, but we concentrate here on predicting the place of articulation of stop consonants, the (bilabial, coronal, velar) series being the most frequent one.

In Fig. 2, we showed that whereas /b,d,g/ are well differentiated in the ($F2$, $F3$) plane, other stops, such as the pharyngeal /ʕ/, are well contrasted in $F1$. The /b,d,g/ stops are produced with a rather high position of the jaw, whereas pharyngeals necessitate a rather low position of the jaw, which is necessary to pull the tongue toward the pharynx at the back of the vocal tract. In a recent study of stop systems, Schwartz et al. (2012b) proposed to exploit the constraint provided by FCT, in which stops correspond to the high-jaw part of the syllabic jaw cycle, with vowels corresponding to the low-jaw part. This is tested in the present study, where we contrast two sets of simulations, in which the jaw is free ("free-jaw" simulations) or biased toward high positions ("high-jaw" simulations). For the first set, pharyngeals should provide good configurations in terms of auditory dispersion, which would tend to make them prominent in simulations, contrary to UPSID data. For the second set, pharyngeals should occur only rarely, because the high position of the jaw is unfavorable to them, in line with FCT.

6.3.2. Methodology

Agent specification: In these simulations, the jaw is added to the motor space, which now contains the whole set of four variables (J, TB, TD and LH). Furthermore, to simulate stops, instead of conserving open vocal tract configurations, as for vowels simulations, we now conserve only the almost-closed vocal tract configurations from the dictionary, with a constriction area of between 0.05 and 0.15 cm^2 (a minimal area is required for VLAM to be able to compute formants).

Probability distributions: The motor subsystem becomes

$$\begin{aligned} P(M \mid O_S) &= P(J\ TB\ \ TD\ \ LH \mid O_S) \\ &= P(J \mid O_S)P(TB \mid O_S)P(TD \mid O_S)P(LH \mid O_S). \end{aligned}$$

Considering the factor $P(J \mid O_S)$, which is specific to stops, two implementations are considered. In the "free-jaw" condition, jaw configurations are allowed to vary over the whole range of the J variable. The "high-jaw" condition corresponds to the hypothesis, derived from FCT, whereby consonants are realized in the high-jaw part of the syllabic jaw cycle. In the "high-jaw" condition, we consider that $P(J \mid O_S) = P(J)$, i.e. that the jaw configuration is independent of the communication object. $P(J)$ is defined as a Gaussian distribution with a high initial mean and a low initial standard deviation.

The sensory motor subsystem $P(S \mid M)$ is learned from the dictionary as for the vowel emergence simulations, except that the dictionary now corresponds to closed configurations of the vocal tract. The auditory subsystem is learned as for the vowel simulations.

Deictic games: Deictic games function as for the vowel simulations. We run three-object simulations, with two sets of simulations: "free-jaw" and "high-jaw". Considering noise ratio conditions, in agreement with the results of vowel simulations, we keep only the "1–3–6" noise ratio set.

Evaluation: Consonants emerging from the simulations are classified according to the eight classes displayed in Fig. 2; i.e., according to their constriction places. For a given simulation, the stop associated with an object is classified according to this grid, using the mean of the X_c values (the position of the constriction inside the vocal tract, see Fig. 16, top), for the stimuli produced by the whole society for that object, in the last of the deictic games. Fig. 16 (bottom) shows the acoustic space for stops computed by VLAM, with the constraints defined above and the dispersion ellipses corresponding to each class. The /b,d,g/ system corresponds to classes [1 3 5].

6.3.3. Results

As for the vowel emergence simulations, we ran 10 independent simulations for each σ_{F1} value, from 0.1 to 0.9 in 0.1 increments. We set σ_{F2} and σ_{F3} according to the "1–3–6" noise ratio. Fig. 17 shows three typical simulations, for low, medium and high noise, in both "free-jaw" and "high-jaw" conditions. Tables 6 and 7 show the complete results, including the consonant systems that emerged from the 180 independent simulations.

We observe that in the "free-jaw" condition, /b,d,g/ emerges at a low noise level (notice that the dental /d̪/ may appear instead of the alveolar /d/, but the difference between dentals and alveolars is seldom made in phonological descriptions) but pharyngeal stops tend to appear at medium and high levels, as expected, considering the high $F1$ values that make them excellent candidates for acoustic distinctiveness. In the "high-jaw" condition, the /b,d,g/ consonant system is strongly preferred, with pharyngeal consonants being discarded from the simulations by the "high-jaw" constraint.

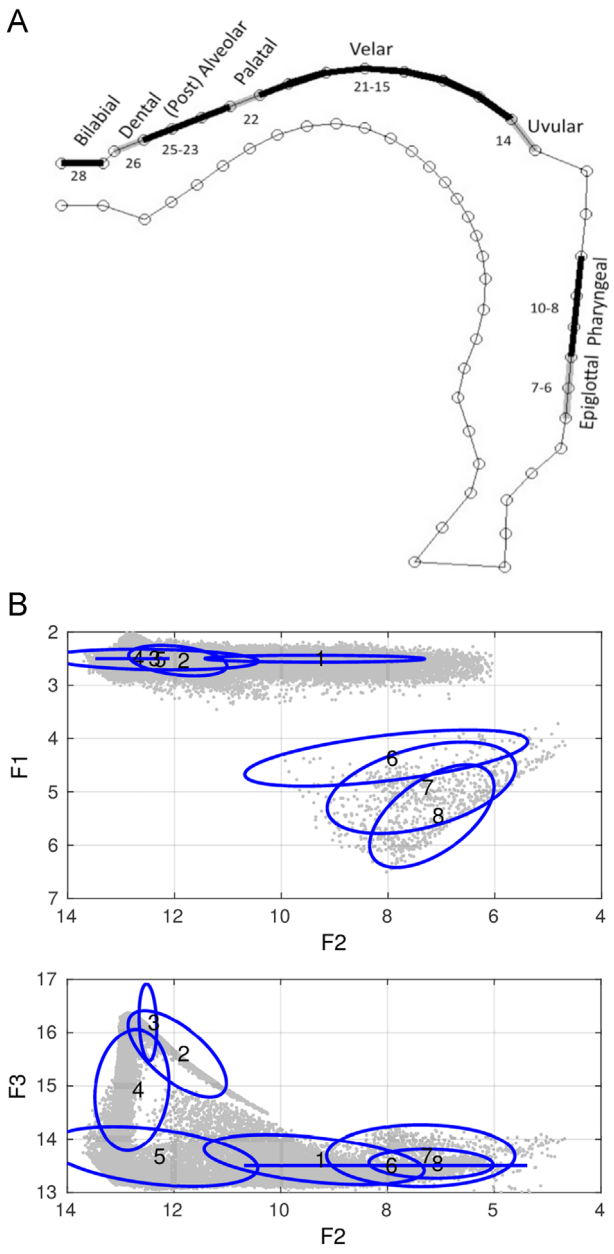


Fig. 16. Stop classification. (A) Stop classification, according to the constriction position in the vocal tract. Adapted from Fig. 5a of Schwartz et al. (2012b). (B) Acoustic consequences of the eight stop classes in the (*F*1, *F*2) and (*F*2, *F*3) planes used in our classification scheme. 1: bilabials /b/, 2: dentals /d̪/, 3: alveolars /d/, 4: palatals /ɟ/, 5: velars /g/, 6: uvulars /ɢ/, 7: pharyngeals /ʕ/ and 8: epiglottals /ʡ/.

6.3.4. Conclusion

Whereas vowel simulation results were in line with previous studies, our simulations of the emergence of stops are, to our knowledge, a novel contribution. These simulations involve a realistic model of orofacial control of the vocal tract, and they lead to the emergence of /b,d,g/, provided that a constraint on the jaw position, derived from FCT, is introduced into the model.

Of major interest in these simulations is their production via a model that used the same parameters as for the vowel emergence simulations, in terms of the noise levels and their ratios. Globally, the whole computational architecture involving COSMO, VLAM and FCT, with the same range of experimental parameters (including the communication noise added in *F*1, *F*2 and *F*3), provides coherent simulations for both vowel and stop systems. We shall now present preliminary experiments dealing with the association of stops and vowels in syllables.

6.4. Syllable emergence

6.4.1. Objectives

The two previous studies showed that plausible vowel and stop systems emerge from deictic games involving sensory–motor COSMO agents equipped with a VLAM articulatory-to-acoustic model, within the framework of FCT, and with a coherent set of

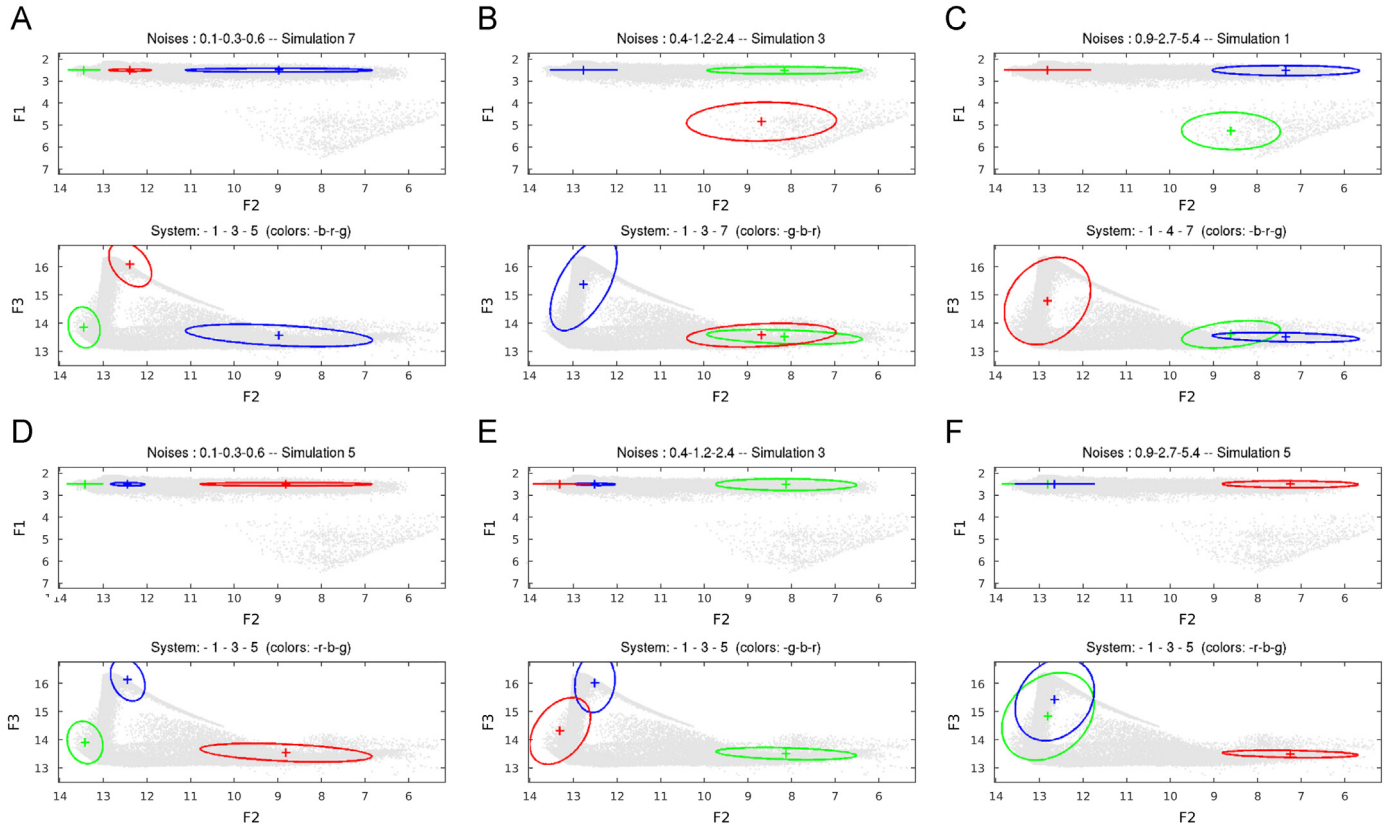


Fig. 17. Typical systems obtained in stop simulations. Same convention as in Fig. 13 B. Top line (A, B, C): "free-jaw" simulations. Bottom line (C, D, E): "high-jaw" simulations. Left column (A, D): low noise level, with $\sigma_{F1} = 0.1$, $\sigma_{F2} = 0.3$, $\sigma_{F3} = 0.6$. Center column (B, E): medium noise level, with $\sigma_{F1} = 0.4, \sigma_{F2} = 1.2, \sigma_{F3} = 2.4$. Right column (C, F): high noise level, with $\sigma_{F1} = 0.9, \sigma_{F2} = 2.7, \sigma_{F3} = 5.4$.

Table 6
Simulation results for stop simulations, in the "free-jaw" condition. For each base noise value (σ_{F1}), 10 independent simulations resulted in 10 final consonant systems. The columns sum to 10.

System	Value of σ_{F1}								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
/b,d,g/	**4**	**6**		**1**	**1**	**1**		**1**	**1**
/b, d̪, g/	2	1	2		2	3	1	1	
/b,b,d/	2	2	1				1		
/b,b, d̪/	1								
/b,d, ɟ/	1		1				1		1
/b,d, ʡ/		1	3	4	1	2		4	2
/b,g, ʡ/			3		1		1	2	2
/b,b,g/				2					
/b, ɟ, ʡ/				2	1			1	3
/b, d̪, ʡ/				1	2	1			1
/b,d,d/					1				
/b, d̪, d/					1	3	3		
/b, d̪, ɟ/							2	1	
/b, d̪, d̪/							1		

experimental parameters (including communication noise). In this final study, we test the emergence of systems of stop/vowel syllables as a whole.

The first challenge is theoretical. The objective here is to study whether syllabic sequences could lead directly to the emergence of stops and vowels compatible with existing data about stop and vowel systems, rather than performing simulations for the separate emergence of vowels and stops as in the previous simulations.

The second challenge is computational. Mixing stops and vowels induces a higher level of complexity, related to the increased dimensionality of the syllable motor and sensory spaces. The study of syllable emergence is a first step toward dealing with complex sound sequences in COSMO.

Table 7
Simulation results for stop simulations, in the "high-jaw" condition. For each base noise value (σ_F1), 10 independent simulations resulted in 10 final consonant systems. The columns sum to 10.

System	Value of σ_F1								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
/b,d,g/	**7**	**6**	**5**	**6**	**4**	**2**		**1**	**4**
/b,b,d̪/	1	1	2	1	2				
/b,d,ʄ/	1						1	1	
/b,b,d/	1	3	1	2	1	1	1		
/b,b,g/			1	1		2	2		
/b,b,ʄ/			1						
/b,d̪,g/					2	3	2	3	1
/b,ʄ,g/					1	1		1	2
/b,d̪,ʄ/						1		2	
/b,g,g/							2	1	2
/b,d̪,d/							1		
/b,d,d/							1		
/b,ʄ,ʄ/								1	
/b,d̪,d̪/									1

Of course, it is possible to conceive syllables as sequences of completely independent sensory–motor units. In this case, independence would simply remove the complexity and lead to simulations in which syllabic systems are just the combination of pure vowel and stop systems. That is, the study of a three-syllable system would lead to arbitrary combinations of the stop set /b,d,g/ with the vowel set /i,a,u/ (e.g. /bi,da,gu/ or /bu,da,gi/).

On the contrary, we should assume that there are links between stops and vowels, as described by FCT. FCT postulates that syllables emerge from cyclic movements of the jaw, producing alternations of stops when the jaw is up and vowels when the jaw is low. In the first stage of FCT, articulators other than the jaw do not move much within the jaw cycle. This results in preferred combinations called "co-occurrences", in which the stop and the vowel have no specific displacement of the tongue (central vowels with labial consonants, as in /ba/), or both are front oriented (front vowels with coronal consonants, as in /di/) or both are back oriented (back vowels with dorsal consonants, as in /gu/). MacNeilage and Davis (2000) show that these co-occurrences are favored in human languages (and also in infants' first words – though see Giulivi, Whalen, Goldstein, Nam, & Levitt, 2011; Nam, Goldstein, Giulivi, Levitt, & Whalen, 2013). We therefore aim to test this in our final set of simulations.

6.4.2. Methodology

Agent specification: The sensory–motor variables in the syllabic model are simply the concatenation of the stop and vowel variables. In the motor space, this leads to four articulatory variables for the stops $M_C = (J_C, TB_C, TD_C, LH_C)$, and three for the vowels $M_V = (TB_V, TD_V, LH_V)$ (the jaw parameter J_V is set to a neutral position as for the vowel model). In the auditory space, this leads to three variables each for the stops ($F1_C$, $F2_C$, $F3_C$), and vowels ($F1_V$, $F2_V$, $F3_V$).

Probability distributions: In agreement with FCT, we specify a "syllable production" model in which the stop motor configuration depends on the vowel motor configuration (see Fig. 18). To produce a syllable, we first reuse the vowel production model (see Section 6.2). That is, given an object o_i, a motor gesture m_V is selected according to the sensory–motor behavior. This motor gesture is fed as an input to a modified version of the consonant production model, where the motor model $P(M_C \mid O_S)$ is refined into $P(M_C \mid M_V O_S)$, so that the gesture chosen to produce the preceding vowel influences the consonant gesture. The mathematical details behind this syllable production model are provided in Appendix A.5.

We describe the knowledge expressed in $P(M_C \mid M_V O_S)$ intuitively. The jaw is probabilistically constrained so that it tends to be in a neutral position for vowels and in a closed position for consonants (as in the "high-jaw" condition). This leads to a jaw half-cycle that is in line with FCT. The other articulators are constrained so that large movements during the jaw half-cycle are less probable than small movements.

Although this model appears to describe a vowel-consonant succession, we should note that probabilistic dependencies do not correspond to any assumption about temporal properties.

Deictic games: Deictic games function as for the vowel and stop simulations. We ran three-object simulations. Considering noise ratio conditions, and in agreement with the results of the vowel and stop simulations, we used only the "1–3–6" noise ratio set with medium noise, thereby allowing a good dispersion for vowels and consonants ($\sigma_{F1} = 0.4$).

Evaluation: Syllable system classification corresponds to the concatenation of vowel and consonant classifications, as described in the corresponding Sections 6.2.2 and 6.3.2. A syllabic system is therefore composed of three elements, one per object, with each associating one of seven vowel classes (see Fig. 13) with one of eight consonant classes (see Fig. 16). Therefore, 56 configurations are possible for each syllable in the simulated system.

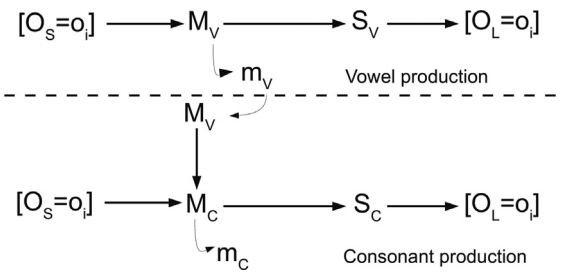


Fig. 18. Sensory–motor behavior of syllable production. The internalization of the communication condition $[C = 1]$ is represented in an equivalent way by $[O_S = o_i]$ and $[O_L = o_i]$. Top: vowel production model of Section 6.2. Bottom: consonant production of Section 6.3, except for the motor subsystem receiving an additional input from vowel production, leading to a $P(M_C | M_V O_S)$ conditional distribution.

Table 8
Three-syllable simulations, for a "1–3–6" noise ratio. Syllable system occurrences emerging from our simulations with a motor constraint between the vowel and consonant realizations.

System	Number of occurrences
/da,di,bu/	13
/ba,di,bu/	7
/ba,di,gu/	5
/da,de,bu/	2
/da,bi,gu/	2
/ba,de,gu/	1
/da,bi,bu/	1
/ga,di,bu/	1
/ba,di,go/	1

6.4.3. Results

We ran 33 independent simulations[4], which resulted in various final syllable systems. We show, in Table 8, their relative frequency of occurrence.

The results do show co-occurrence effects, because associations between stops and vowels in emergent syllabic systems are not random (see Table 8). Note that whereas stop configurations are defined by the place of articulation in the vocal tract, the corresponding formant patterns may differ because of vowel context, as shown in Fig. 19.

The /ba,di,gu/ system, described by MacNeilage and Davis (2000) as having a preferred "co-occurrence" pattern, is well represented in the simulations.

However, the co-occurrences obtained are not wholly satisfactory. Indeed, the /da,di,bu/ system is predominant. This system is predicted by a total dependence model for which the syllable is obtained by a pure movement of the jaw for each vowel within /i,a,u/. In fact, for the VLAM model, it has been shown (Serkhane, 2005; Serkhane, Schwartz, & Bessiere, 2005, 2007; Vilain, Abry, Badin, & Brosda, 1999) that with a pure upward jaw movement, an /a/ configuration leads more often to a /d/ than a /b/, that an /i/ clearly leads to a /d/, and that a /u/ leads more often to a /b/ than a /g/ (the lips being quite close during /u/). These results would appear to contradict the predictions of FCT. However, Vilain et al. (1999) show that patterns of co-occurrences can be highly dependent on the vocal tract morphology for a given speaker.

In these simulations, the dispersion force induced by the sensory–motor behavior is insufficient to overcome the motor constraint. This results in a /da,di,bu/ system that is well dispersed with respect to its vowel but less so with respect to its consonant. Reaching a /g/ configuration from an /a/, which would allow the emergence a well-dispersed system, appears to be impeded by the motor constraint, because going from one to the other necessitates a large movement by the tongue dorsum (*TD*). We also observe the absence in the simulations of two syllables, /du/ and /gi/, which involve large movements of the tongue body *TB* in the front–back dimension. Their absence is therefore in line with FCT.

Globally, the lack of independence between vowels and stops in the simulations, caused, in the model, by the initial strong dependence between motor parameters for the vowel and the stop (and expressed by the factor $P(M_C | M_V O_S)$), results in a pattern of dependence that is well captured by the relationship between the *F*2 values for the stop and the vowel. This relationship is displayed in Fig. 20, in which we plot the *F*2 value for the vowel and for the consonant for the final configurations obtained in the whole set of the 33 simulations, that is grouping altogether 99 "syllabic" plosive-vowel configurations. This relationship is considered in classical studies as a good indicator of the "coarticulation" mechanism according to which stops and vowels in adult speech are always produced in a coordinated way (Sussman, Fruchter, Hilbert, & Sirosh, 1998). Therefore, this pattern of dependence is in agreement with real data.

[4] Clearly, this is insufficient for complete characterization of these simulations, which are therefore considered as preliminary. However, they give an outline of the global statistical tendencies for the way that syllable systems can be shaped by the perceptuo-motor properties of the COSMO agents.

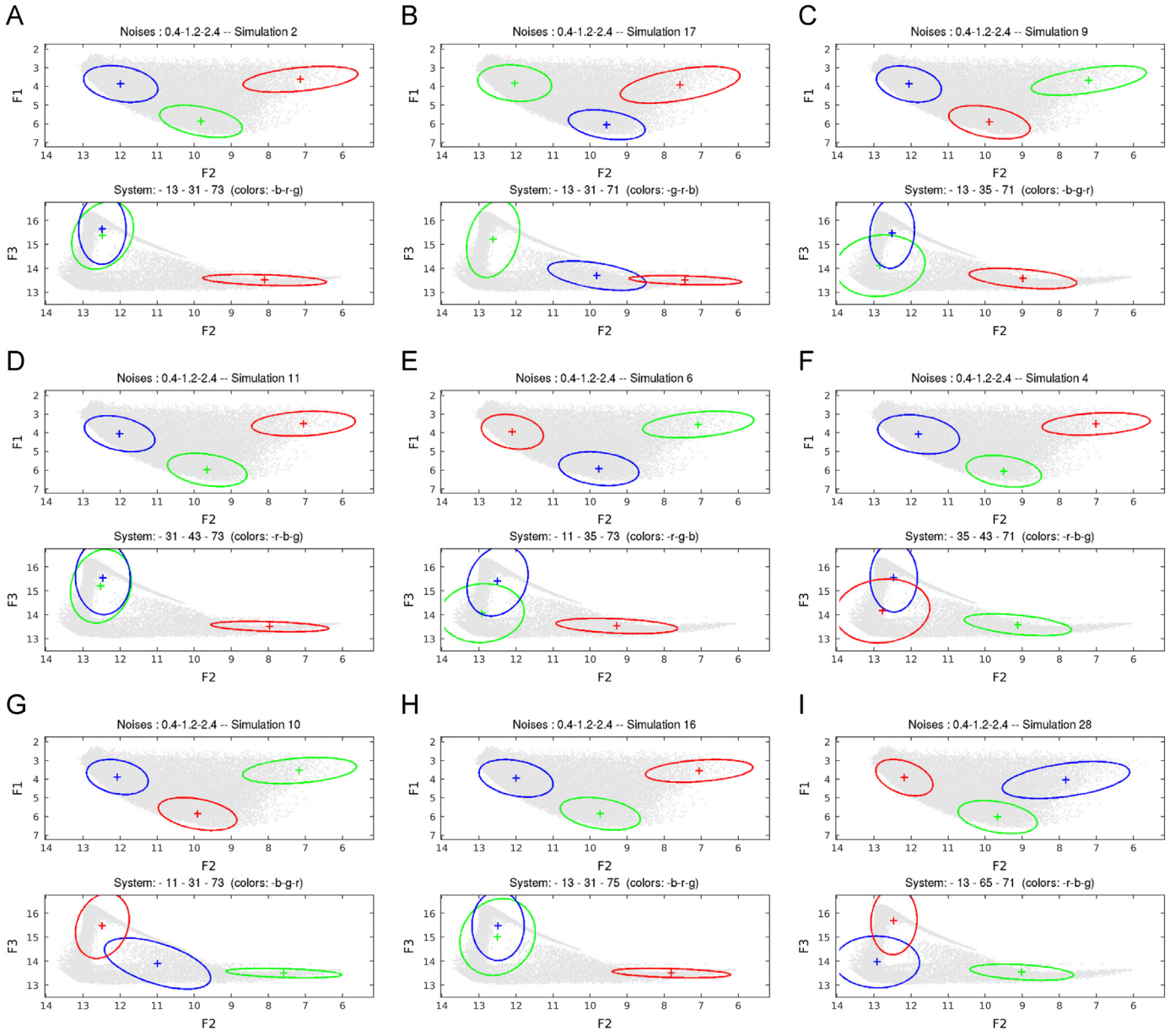


Fig. 19. Syllable systems observed in our simulations, shown in the (*F*1, *F*2) plane for vowels and the (*F*2, *F*3) plane for consonants. Same convention as in Fig. 13 B. (A) /da,di,bu/. (B) /ba,di,bu/. (C) /ba,di,gu/. (D) /da,de,bu/. (E) /da,bi,gu/. (F) /ba,de,gu/. (G) /da,bi,bu/. (H) /ga,di,bu/. (I) /ba,di,go/.

Finally, it must be stressed that acoustic dispersion in the stop space does play a part, in addition to motor constraints, in well-represented systems, including not only /ba,di,gu/, of course, but also /ba,di,bu/. In fact, for /ba,di,bu/, the two /b/ exemplars correspond to different acoustic realizations (see Fig. 19, top, middle, (*F*2, *F*3) panel).

6.4.4. Conclusion

This section has shown preliminary results on syllable emergence using the COSMO model. This is the first time that such systems have been simulated on VLAM. We show that it is possible to combine motor dynamic principles inspired by FCT with acoustic dispersion principles implied by the sensory–motor production behavior.

The results are not wholly satisfactory, because the simulations do not give /ba,di,gu/ as the preferred set. There are several possible reasons for this result, including not only the particular morphology of the VLAM simulator (possibly favoring /da/ instead of /ba/, Vilain et al., 1999) but also a suboptimal compromise between acoustic dispersion and motor constraints resulting from modeling choices.

However, the simulations do provide systems that are dispersed (in terms of vowels, certainly, and with a reasonable dispersion in stops) and that display patterns of co-occurrence more or less in line with data from the literature. Of course, much work remains to be done, following these preliminary simulations, to provide a complete understanding of the emergence of syllable systems.

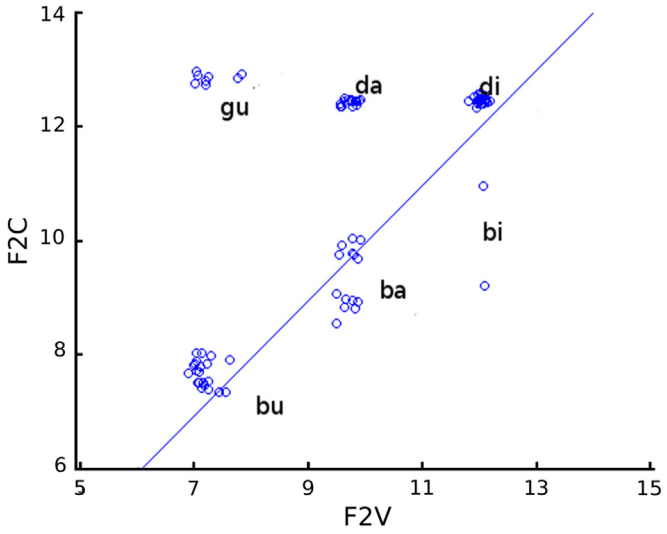


Fig. 20. Plot of the second formant of the stop $F2_C$ as a function of the second formant of the vowel $F2_V$, in the syllable systems resulting from our simulations. The diagonal line is not a linear model obtained from regression but is the identity function, which would indicate complete dependence. Overall, this point cloud shows the pattern of dependence between $F2_V$ and $F2_C$.

7. General discussion

7.1. Contributions

Let us recall the two major objectives of this work: (1) attempt to simulate in a single integrated framework universals of vowel systems, plosive systems and their associations in CV syllables; and (2) connect within this framework theories of speech sound systems (e.g. the dispersion theory or the quantal theory) and theories of speech communication (e.g. auditory, motor or perceptuo-motor theories). These objectives have been reasonably well fulfilled, thanks to three major contributions.

Our first contribution (Section 3) is mainly theoretical. We extract three basic requirements from a speech-communication prototypical scenario, related to reference, parity and adequacy. The proposed COSMO model is based on an "internalization hypothesis" that produces a unified framework for describing speech production and perception in a global cognitive architecture that associates motor and auditory subsystems through a sensory–motor link. In this framework, the so-called "motor" and "auditory" theories of speech communication can be seen as simplifications of the general COSMO sensory–motor theory that internalizes the whole communication situation. We show that the internalization hypothesis can be formalized in a Bayesian model that allows a probabilistic expression of the computational issues involved in the main theoretical trends in the speech production and perception literature (see Table 1).

Our second contribution (Section 5) deals with "parity" in speech communication, by extending the model to include a referential evolutionary bootstrap for studying speech code emergence in societies of interacting COSMO agents. We use deixis as a plausible prelinguistic communicative behavior. A comparison of simulation results based on the three COSMO versions (motor, auditory and sensory–motor), implemented using simplified articulatory–acoustic spaces, confirms previous theoretical propositions about the need to take auditory expectations into account if adequate phonological system predictions (Schwartz et al., 2007; Studdert-Kennedy & Goldstein, 2003) are to be obtained. We show that within this framework, societies of "pure motor agents" do not allow the emergence of an efficient and structured speech code, unlike auditory and sensory–motor agents. In addition, the sensory–motor COSMO version leads to quicker convergence and "better" speech codes in terms of sound distinctiveness. Furthermore, Lindblom's Dispersion Theory and Stevens' Quantal Theory appear to be grounded in more general principles of communication, in the sense that neither distance optimization nor nonlinearity detection processes were included explicitly in the computational architecture.

Our third contribution (Section 6) deals with "adequacy" in speech communication and is based on realistic predictions about the phonological systems of human languages in the "sensory–motor" COSMO model. Firstly, we confirm previous results about adequate formant weights for auditory distances to obtain accurate predictions (Schwartz et al., 1997a). Although these weights were proposed for vowel systems, we extend them to stops, showing that COSMO simulations, incorporating a jaw constraint in line with FCT, also provide adequate predictions for stop systems. Conjoint predictions about vowel and consonant systems in a unified model, i.e., one that uses the same structure and parameters, is a novel contribution to knowledge. Using this unified framework, we extend COSMO to sequences in which vowels and consonants are interwoven into syllables, showing that a dynamical motor constraint expressed as a probabilistic dependency between vowel and consonant productions leads to plausible, even if preliminary, predictions about "co-occurrence" and coarticulation effects.

Altogether, these three contributions show the value of proposing a unified formal and simulation framework that can deal with both online speech communication and the emergence of sound systems for human languages, thereby enabling comparisons between evolutionary hypotheses and speech theory implementations with respect to world-language data and associated phenomenological laws.

7.2. Perspectives

Although the present implementation of COSMO capitalizes on specific theories about the emergence of speech and language, such as the role of deixis for reference, the use of a mirror system for parity and the use of a frame-then-content scenario for adequacy, the COSMO conception and formalization is independent of these specific assumptions, which should encourage both extensions and alternatives. We now consider some of these possibilities.

7.2.1. Sensory–motor spaces and compositionality

One perspective is to consider extending the domain of sensory–motor variables M and S, both spatially (the number and control precision of the articulators involved) and temporally (the sequences within syllables or words).

In the articulatory domain, adding parameters for the tip of the tongue or the lip protrusion in VLAM would provide more accurate predictions, including improving the production of rounded vowels such as /u/ or frontal consonants such as /d/. In the auditory domain, considering other acoustic features such as signal intensity, acoustic burst or turbulence detection might allow the emergence of vowel and consonant categories, without having to specify them explicitly in the model. It would also enable predictions about other phonological categories, such as fricative sounds.

Any extension to manage temporal sequences, as we outlined for syllables in Section 6.4, involves consideration of compositionality, which is not yet included in the model (De Boer & Zuidema, 2010). Indeed, simulations of syllabic systems should take into consideration the fact that in human languages, vowel and consonants are combined into syllables in a compositional way, in the sense that each vowel can be associated with each consonant of a given phonological system. Compositionality also seems to appear at the phonetic level (Ohala, 1979; Clements, 2003a, 2003b), put forward as a so-called principle of the “Maximal Use of Available Features”. According to this principle, if a given set of features has emerged, then all combinations of each feature value are generally available in the system.

Such extensions will have to deal with the computational complexity implied by Bayesian inference. One way to manage the curse of dimensionality is to involve motor synergies and motor primitives. Motor synergies allow some articulatory dimensions to be merged into a single parameter, such as associating lip height and lip protrusion into a single lip-rounding parameter, as in Moulin-Frier et al. (2012). Motor primitives enable a complex temporal trajectory, often highly dimensional, to be defined as an appropriately parameterized function of time (Konczak, 2005). Further extensions could involve adding vision to audition in the sensory domain, and hand to mouth in the motor domain. The role of vision in the emergence of phonological systems has been mentioned in previous studies (Schwartz et al., 2007). Manual gestures could also play a crucial role in the emergence of reference, as we will discuss in the next section.

7.2.2. Reference and syntax

In their present state, our simulations consider that reference is provided solely by deixis, which allows us to refer to objects in the world. These objects are typically referred to by nouns, which are considered as pure phonological units in the present paper. However, hand gestures can provide much more information than pure deixis, as illustrated by Arbib in his expanding spiral model (Arbib, 2005a). A second set of perspectives might then extend the domain of reference variables O_S and O_L, and their probabilistic link through the fusion variable C.

The “Mirror System Hypothesis” (Rizzolatti & Arbib, 1998) is a relevant framework of interest, which suggests that language has evolved from the ability to understand another's actions as if they were one's own. This would provide a basis for verbs rather than nouns. Further extensions of this hypothesis (Roy & Arbib, 2005) argue for a “syntactic motor system”, given that some syntactic properties are already present in manual actions toward objects, such as “taking an object” or “threatening a congener”. Including this proposition in the COSMO framework would require definition of the objects of communication as associations between a physical object in the environment and a manual action through the concept of affordance. Again, the issue of compositionality is invoked here to combine object and action denotation.

7.2.3. Interaction paradigms

COSMO agents can be tested inside other interaction paradigms. In a recent paper (Moulin-Frier et al., 2012), we used the COSMO framework to propose a computational comparison of the motor, auditory and sensory–motor versions of the model with respect to speech perception tasks. The interaction paradigm was different, in that agents were considered to have already acquired a speech code, and we studied the robustness of each COSMO version to adverse perception conditions (noise and speaker variability). This led to the determination of “perfect” conditions, in which motor and auditory theories are in fact indistinguishable (and the corollary, by which they can be distinguished under degraded conditions). In the context of the active, although long-standing, debate between supporters of different theories about speech perception, the comparative computational approach is increasingly used in cognitive psychology. The same approach might be proposed for speech production and speech development (Moulin-Frier & Oudeyer, 2012, 2013).

A number of questions remain open at the present stage of our work. We can immediately identify two questions:

- What is the role of feedback in interaction? In the present state of our simulations, the agents are supposed to know that they face the same object and to assume that communication is successful. Feedback could be provided at some stage, as is the case for

some other multi-agent simulations (for example, see de Boer, 2000). This would modify the learning processes, and it is certainly a crucial process in the course of language development.

- Can dialects within subgroups of agents emerge from simulations? So far, we have considered only a few agents, which are assumed to represent the whole society. Of course, for more realistic simulations involving a large number of agents, the diffusion process of phonological systems might become quite complex and might lead to inhomogeneous distributions via the constitution of subgroups, with “dialects” and local accents (for example, see Berrah et al., 1996; Berrah, 1998).

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152 - “Speech Unit(e)s”). A secondary source of founding was the ERC Starting Grant EXPLORERS 240 007 which partially financed the first author during the revision process.

The authors would also like to thank Louis-Jean Boë, Raphaël Laurent and Pascal Perrier for countless helpful discussions, as well as the Explora-Doc program of the French Rhône-Alpes region.

Appendix A

This appendix describes the Bayesian Programming framework we use for the COSMO model and provides some details about the variations we made in the syllable model of Section 6.4.

We first define the notations used in this paper in more detail. Second, we explain how the joint probability distribution of the COSMO model is obtained from conditional independence hypotheses. Third, we describe how Bayesian inference on the joint distribution is used to express speech production and perception tasks. Fourth, we explain how learning occurs by updating the terms of the joint distributions during deictic games. Fifth, we provide details about how consonant production is influenced by vowel production in the syllable model of Section 6.4.

A.1. Notations

The notations and probabilistic principles we used in this paper are inspired by Jaynes (2003) and Lebeltel et al. (2004), Bessiére et al. (2013).

Upper case A denotes a probabilistic variable, defined by its domain $\mathcal{D}(A)$, corresponding to the set of discrete values that the variable can take (we limit our modeling to the discrete case here). The conjunction of two variables $A \wedge B$ can be defined as a new variable C with domain $\mathcal{D}(A) \times \mathcal{D}(B)$. Lower case a denotes a particular value of the domain $\mathcal{D}(A)$. $P(A)$ is the probability distribution over A and $P(AB)$ is the probability distribution over $A \wedge B$. $P(A \mid [B = b])$ is the conditional distribution over A, given a particular value b of variable B (also noted $P(A \mid b)$ when there is no ambiguity on variable B). For simplicity in the paper, we generally do not distinguish a variable and its domain, thus allowing notations like $a \in A$.

A.2. Joint probability distribution decomposition

In the Bayesian Programming framework, an agent cognitive architecture is defined as the decomposition of a joint probability distribution over variables of interest. These latter are typically motor, sensory and internal cognitive variables.

Mathematically, the joint probability distribution over a variable conjunction, $V = V_1 \wedge \ldots \wedge V_n$, can be expressed as the product of simpler distributions using Bayes rule. For instance

$$P(V) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_1V_2)\ldots P(V_n \mid V_1\ldots V_{n-1}). \tag{8}$$

This is called a decomposition. Note that many possible decompositions can be chosen. For example, Bayes rule can also yield

$$P(V) = P(V_3)P(V_1 \mid V_3)P(V_2 \mid V_1V_3)\ldots P(V_n \mid V_1\ldots V_{n-1}),$$

and so on.

In practice, such a decomposition can be simplified by considering conditional independence hypotheses between variables. Let us consider for example that the modeler knows that each probability distribution on variable V_i only depends on variable V_{i-1}. Then (8) can be simplified into

$$P(V) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2)\ldots P(V_n \mid V_{n-1}). \tag{9}$$

This allows the complexity of the joint probability distribution computation to be considerably reduced.

In COSMO, the joint probability distribution deals with five variables of interest: $P(O_S\ M\ S\ O_L\ C)$. The chosen decomposition is

$$P(O_S\ M\ S\ O_L\ C) = P(O_S)P(M \mid O_S)P(S \mid M)\ P(O_L \mid S)P(C \mid O_S\ \ O_L). \tag{10}$$

Therefore, COSMO assumes the following conditional independence hypotheses:

- $P(S \mid O_S\ M) = P(S \mid M)$: given the motor command, the auditory stimulus is independent of the speaker object;
- $P(O_L \mid O_S\ M\ S) = P(O_L \mid S)$: given the auditory stimulus, the listener object is independent of the speaker object and motor command;
- $P(C \mid O_S\ M\ S\ O_L) = P(C \mid O_S\ O_L)$: given the speaker and listener objects, the communication success is independent of the motor command and auditory stimulus.

In other words, COSMO assumes that the motor command only depends on the speaker object, the auditory stimulus only depends on the motor command, the listener object only depends on the auditory stimulus and the communication success only depends on the speaker and listener objects.

Such a decomposition and the conditional independence hypotheses can be represented by a graph with no oriented cycle, as in Fig. 4, thus providing a graphical representation of the chosen cognitive architecture.

A.3. Inference

In the Bayesian Programming framework, behaviors correspond to Bayesian inferences from the joint distribution.

A Bayesian inference is the computation of a conditional probability distribution $P(Se \mid Kn)$, where Se and Kn are disjoint subsets of the variables of interest, called the *searched* and *known variables*, respectively (with $Se \neq \varnothing$). Using Bayes rule and the marginalization rule, any such conditional probability distribution can be computed from the joint probability distribution by the following formula:

$$P(Se \mid Kn) = \frac{\sum_{Fr} P(Se\ Kn\ Fr)}{\sum_{Se,Fr} P(Se\ Kn\ Fr)}, \tag{11}$$

where Fr are the variable of interest which are neither in Se nor in Kn, called the *free variables*. The denominator being constant when Kn is known, it can be considered as a normalization term, assumed to be computed afterward, thus simplifying inference to

$$P(Se \mid Kn) = \frac{1}{Z} \sum_{Fr} P(Se\ Kn\ Fr). \tag{12}$$

This is also noted $P(Se \mid Kn) \propto \sum_{Fr} P(Se\ Kn\ Fr)$, with the $\propto$ symbol denoting proportionality.

A conditional probability distribution of the form $P(Se \mid Kn)$ is called a *question* to the joint distribution: knowing a particular value k of Kn (the *known* variables), what is the probability distribution over Se (the *searched* variables)? The answer consists in effectively computing the distribution, using (12) and, possibly, using efficient computational techniques.

In COSMO, a production behavior is modeled by a question of the form $P(M \mid [O_S = o_i][C = 1])$: given that the agent wants to communicate ($[C = 1]$) object o_i, what is the probability distribution about motor commands M? Using (12) on the joint distribution of Eq. (10), the answer is

$$\begin{aligned} &P(M \mid [O_S = o_i][C = 1]) \\ &\quad \propto P(M \mid [O_S = o_i]) \\ &\qquad \sum_{S,O_L} P(S \mid M) P(O_L \mid S) P([C = 1] \mid [O_S = o_i]\ O_L) \\ &\quad \propto P(M \mid [O_S = o_i]) \\ &\qquad \sum_{S} P(S \mid M) P([O_L = o_i] \mid S). \end{aligned} \tag{13}$$

The first derivation comes from the application of Eq. (12) on the joint distribution of Eq. (10) with $Se = M$, $Kn = O_S \wedge C$ and $Fr = S \wedge O_L$, considering the known values $[O_S = o_i]$ and $[C = 1]$. The second derivation simplifies the sum over O_L given that $[C = 1]$ if and only if $O_S = O_L$ (see Section 3.2.1), i.e. $O_L = o_i$ in this particular case where $O_S = o_i$.

Using similar reasoning, a perception behavior in COSMO is modeled by a question of the form $P(O_L \mid [S = s][C = 1])$ (or, equivalently, $P(O_S \mid [S = s][C = 1])$. The computation is

$$\begin{aligned} &P(O_L \mid [S = s][C = 1]) \\ &\quad \propto P(O_L \mid [S = s]) \\ &\qquad \sum_{M} P(M \mid [O_S = O_L]) P([S = s] \mid M). \end{aligned} \tag{14}$$

Production and perception behaviors defined in (13) and (14), respectively, both involve the entire COSMO architecture: the motor subsystem $P(M \mid O_S)$, the auditory subsystem $P(O_L \mid S)$, the sensory–motor subsystem $P(S \mid M)$ and the fusion subsystem $P(C \mid O_S O_L)$. This full sensory–motor version of COSMO can be simplified into a motor or an auditory version of the model, by considering that the auditory or the motor subsystems are deactivated, respectively. Deactivation is modeled in probabilistic terms by considering the corresponding distribution as uniform, as explained in Section 3.3.

A.4. Distribution updating

During the agent interactions in deictic games (Section 4.2), some terms of the joint probability distribution maintained by each agent are updated online according to the new information available. During a particular deictic game, the speaker agent observes a

new association between the motor command m it produced and the object o_i in front of it. The listener agent observes a new association between the auditory stimulus s it perceived and the object o_i in front of it.

These new associations are used to update the motor subsystem $P(M \mid O_S)$ of the speaker and the auditory subsystem $P(O_L \mid S)$ of the listener. $P(M \mid O_S)$ is a family of Gaussian probability distributions over M (one for each value of O_S), whereas $P(O_L \mid S)$ is inferred from Gaussian probability distributions over S (see (5)). Each $P(M \mid [O_S = o_i])$ is updated by computing the mean and covariance matrix of M in the last 200 deictic games played by the agent as a speaker and where $[O_S = o_i]$. Similarly, each $P(S \mid [O_L = o_i])$ (used to infer the auditory subsystem $P(O_L \mid S)$) is updated by computing the mean and covariance matrix of S in the last 200 deictic games played by the agent as a listener and where $[O_L = o_i]$. Those updates are triggered each 200 deictic games in which a particular agent is involved.

A.5. Detail on the syllable motor subsystem

In our simulations of syllable system emergence in Section 6.4, the motor subsystem $P(M \mid O_S)$ is modified compared to other simulations, as mentioned in Section 6.4.2.

We consider that the consonant is dependent on the vowel:

$$P(M \mid O_S) = P(M_C M_V \mid O_S) = P(M_V \mid O_S) P(M_C \mid M_V O_S)$$

The vowel motor subsystem $P(M_V \mid O_S)$ corresponds to the one presented in Section 6.2.2.

The consonant motor subsystem is extended to be influenced by the vowel:

$$\begin{aligned} P(M_C \mid M_V O_S) &= P(J_C TB_C TD_C LH_C \mid J_V TB_V TD_V LH_V O_S) \\ &= P(J_C) P(TB_C TD_C LH_C \mid J_V TB_V TD_V LH_V O_S) \end{aligned} \tag{15}$$

$$P(M_C \mid M_V O_S) = P(J_C) P(TB_C \mid TB_V O_S) P(TD_C \mid TD_V O_S) P(LH_C \mid LH_V O_S) \tag{16}$$

In the first derivation, the consonant jaw configuration is assumed to be independent of any other variable: it is set in a high position, as in the consonant model of Section 6.3.2 ("high-jaw" condition). In the second derivation, the other articulator configurations are assumed to depend only on the vowel configuration for the same articulator (in addition to the speaker object O_S).

These probability distributions $P(X_C \mid X_V O_S)$, $X \in \{TB, TD, LH\}$ are initially defined as Gaussian probability distributions centered on the vowel configuration X_V, with a small variance. More precisely, $P(X_C \mid [X_V = x] O_S)$ is initially set as a Gaussian probability distribution of mean x and standard deviation set to a sixth of the X_V range (whatever the value of O_S).

Finally, the update rule is similar to the one described in the previous section, except that associations involved in the update of $P(X_C \mid [X_V = x][O_S = o_i])$ are the last 200 deictic games where the agents produced $[X_V = x]$ in front of $[O_S = o_i]$.

Altogether, consonant production is initially constrained by the vowel production but this constraint can be relaxed as learning occurs during deictic games.

References

Abry, C., Ducey Kaufmann, V., Vilain, A., & Lalevée, C. (2008). When the babble syllable feeds the foot in a point. In B. Davis, K. Zajdo (Eds.), *The syllable in speech production: Perspectives on the frame content theory* (pp. 460–472). Erlbaum, https://hal.archives-ouvertes.fr/hal-00264464.

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: Social Behaviour and Communication in Biological and Artificial Systems*, *5*(3), 313–325.

Arbib, M. A. (2005a). From monkey-like action recognition to human language: *An evolutionary framework for neurolinguistics. Behavioral and Brain Sciences*, *28*, 105–167.

Arbib, M. A. (2005b). Interweaving protosign and protospeech: *Further developments beyond the mirror. Interaction Studies*, *6*, 145–171.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science*, *332*(6027), 346–349.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. Cambridge: MIT Press.

Berrah, A.-R. (1998). *Evolution d'une société artificielle d'agents de parole: un modèle pour l'émergence des structures phonétiques* (Ph.D. thesis). Institut National Polytechnique de Grenoble - INPG.

Berrah, A.-R., Glotin, H., Laboissiére, R., Bessiere, P., & Boë, L.-J. (1996). From form to formation of phonetic structures: An evolutionary computing perspective. In T. Fogarty, & G. Venturini (Eds.), *ICML '96 workshop on evolutionary computing and machine learning* (pp. 23–29), Bari.

Bessiére, P., Laugier, C., & Siegwart, R. (Eds.) (2008). *Probabilistic reasoning and decision making in sensory–motor systems, Springer tracts in advanced robotics* (Vol. 46). Berlin: Springer-Verlag.

Bessiére, P., Mazer, E., Ahuactzin, J.-M., & Mekhnacha, K. (2013). *Bayesian programming*. Chapman and Hall/CRC , https://www.crcpress.com/Bayesian-Programming/Bessiere-Mazer-Ahuactzin-Mekhnacha/9781439880326.

Boë, L.-J. (1999). Vowel spaces of newly-born infants and adults consequences for ontogenesis and phylogenesis. In *The 14th international congress of phonetic sciences* (pp. 2501–2504).

Boë, L.-J., Badin, P., Ménard, L., Captier, G., Davis, B., Macneilage, P., Sawallis, T. R., & Schwartz, J.-L (2013). Anatomy and control of the developing human vocal tract: *A response to Lieberman. Journal of Phonetics*, *41*(5), 379–392.

Boë, L.-J., Bessière, P., Ladjili, N., & Audibert, N. (2008). Simple combinatorial considerations challenge Ruhlen's mother tongue theory. In B. L. Davis, & K. Zajdo (Eds.), *The syllable in speech production* (pp. 63–92). New York: Laurence Erlbaum Associates.

Boë, L.-J., Heim, J.-L., Honda, K., Maeda, S., Badin, P., & Abry, C. (2007). The vocal tract of newborn humans and neanderthals: *Acoustic capabilities and consequences for the debate on the origin of language. a reply to Lieberman. Journal of Phonetics*, *35*(4), 564–581.

Boë, L.-J., Vallée, N., Badin, P., Schwartz, J.-L., & Abry, C. (2000). Tendencies in phonological structures: *The influence of substance on form. Les Cahiers de l'ICP. Bulletin de la communication parlée*, *5*, 35–55.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(02), 201–251.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: *An overview. Phonetica*, *49*(3–4), 155–180.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. In *Phonology Yearbook* (Vol. 3, pp. 219–252).

Carlson, R., Granström, B., & Klatt, D. (1979). Vowel perception: *The relative salience of selected acoustic manipulations. STL-QPSR*, *34*, 19–35.

Cheney, D. L., & Seyfarth, R. M. (1982). How vervet monkeys perceive their grunts: *Field playback experiments. Animal Behaviour*, *30*(3), 739–751.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clements, N. (2003a). Feature economy as a phonological universal. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 371–374), Barcelona.

Clements, N. (2003b). Feature economy in sound systems. *Phonology*, *3*, 287–333.

Corballis, M. C. (2002). *From hand to mouth: The origins of language*. Princeton, NJ: Princeton University Press.
De Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, *28*(4), 441–465.
De Boer, B., & Zuidema, W. (2010). Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior*, *18*(2), 141–154.
Demange, S., & Ouni, S. (2013). An episodic memory-based solution for the acoustic-to-articulatory inversion problem. *Journal of the Acoustical Society of America*, *133*(4), 2921–2930.
Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179.
Dominey, P. F. (2007). Towards a construction-based framework for development of language, event perception and social cognition: *Insights from grounded robotics and simulation*. *Neurocomputing*, *70*(13), 2288–2302.
Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: *A magnetic stimulation study*. *Journal of Neurophysiology*, *73*(6), 2608–2611.
Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*(1), 3–28.
Gell-Mann, M., & Ruhlen, M. (2011). The origin and evolution of word order. *Proceedings of the National Academy of Sciences*, *108*(42), 17290–17295.
Gentilucci, M., & Corballis, M. C. (2006). From manual gesture to speech: *A gradual transition*. *Neuroscience and Biobehavioral Reviews*, *30*, 949–960.
Gilet, E., Diard, J., & Bessiére, P. (2011). Bayesian action-perception computational model: *Interaction of production and recognition of cursive letters*. *PLoS One*, *6*(6), e20387.
Giulivi, S., Whalen, D., Goldstein, L. M., Nam, H., & Levitt, A. G. (2011). An articulatory phonology account of preferred consonant-vowel combinations. *Language Learning and Development*, *7*(3), 202–225.
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). Mahwah, NJ: Lawrence Erlbaum Associates.
Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.
Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*(5), 350–365.
Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*(4), 611–633.
Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346.
Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: *What is it, who has it, and how did it evolve*. *Science*, *298*(5598), 1569–1579.
Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, *77*(2), 187–222.
Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press , http://www.cambridge.org/us/academic/subjects/physics/theoretical-physics-and-mathematical-physics/probability-theory-logic-science?format=HB.
Kemp, C., & Tenenbaum, J. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10687–10692.
Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. In *IEEE International Conference on acoustics, speech, and signal processing, ICASSP'82*,(Vol. 7, pp. 1278–1281). IEEE, http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1171512&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D1171512 ..
Konczak, J. (2005). On the notion of motor primitives in humans and robots. In L. Berthouze, F. Kaplan, H. Kozima, H. Yano, J. Konczak, G. Metta, et al. (Eds.), *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems* (Vol. 123, pp. 47–53). Lund University Cognitive Studies.
Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One*, *2*(9), e943.
Leavens, D. A., & Bard, K. A. (2011). Environmental influences on joint attention in great apes: *Implications for human cognition*. *Journal of Cognitive Education and Psychology*, *10*(1), 9–31.
Lebeltel, O., Bessiere, 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), 1–36.
Liberman, A. M., Mattingly, I. G., et al. (1989). A specialization for speech perception. *Science*, *243*(4890), 489–494.
Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187–196.
Lieberman, P. (1984). *The biology and evolution of language*. Harvard University Press.
Lieberman, P. (2012). Vocal tract anatomy and the neural bases of talking. *Journal of Phonetics*.
Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: *The role of perceptual contrast*. *Language*, *48*(4), 839–862.
Lindblom, B. (1984). Can the models of evolutionary biology be applied to phonetic problems. In *Proceedings of the 10th international congress of phonetic sciences* (pp. 67–81). Foris Pubns USA.
Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Ohala, & J. Jaeger (Eds.), *Experimental phonology* (pp. 13–44). Orlando, FL: Academic Press.
Lindblom, B. (1990). Explaining phonetic variation: *A sketch of the H&H theory*. In W. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer.
MacNeilage, P., & Davis, B. (2001). Motor mechanisms in speech ontogeny: *Phylogenetic, neurobiological and linguistic implications*. *Current Opinion in Neurobiology*, *11*, 696–700.
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.
Maddieson, I. (1984). *Patterns of sounds*. Cambridge University Press.
Maddieson, I. (2001). Typological patterns-geographical distribution and phonetic explanation. In *Conference on the phonetics–phonology interface*.
Maddieson, I., & Precoda, K. (1989). Updating UPSID. *The Journal of the Acoustical Society of America*, *86*(S1), S19.
Maeda, S. (1989). Compensatory articulation during speech: *Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model*. *Speech Production and Speech Modelling*, 131–149.
Manser, M. B., & Fletcher, L. B. (2004). Vocalize to localize: *A test on functionally referential alarm calls*. *Interaction Studies*, *5*(3), 327–344.
Moore, R. K. (2007). Spoken language processing: *Piecing together the puzzle*. *Speech Communication*, *49*(5), 418–435.
Moulin-Frier, C. (2011). *Rôle des relations perception-action dans la communication parlée et l'émergence des systémes phonologiques: étude, modélisation computationnelle et simulations* (Ph.D. thesis). Université de Grenoble.
Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., & Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: *An exploratory Bayesian modeling study*. *Language and Cognitive Processes*, *27*(7–8), 1240–1263.
Moulin-Frier, C., & Oudeyer, P.-Y. (2012). Curiosity-driven phonetic learning. In *ICDL-Epirob—International conference on development and learning*, *Epirob*, San Diego, États-Unis.
Moulin-Frier, C., & Oudeyer, P.-Y. (2013a). Exploration strategies in developmental robotics: A unified probabilistic framework. In *International conference on development and learning, ICDL-Epirob*, Osaka, Japan, in press.
Moulin-Frier, C., & Oudeyer, P.-Y. (2013b). The role of intrinsic motivations in learning sensorimotor vocal mappings: A developmental robotics study. In *Proceedings of Interspeech*, Lyon, France.
Moulin-Frier, C., Schwartz, J., Diard, J., & Bessière, P., (2008). Emergence of a language through deictic games within a society of sensori-motor agents in interaction. In *The eighth international seminar on speech production, ISSP'08*, Strasbourg, France.
Moulin-Frier, C., Schwartz, J., Diard, J., & Bessière, P. (2010). A unified theoretical bayesian model of speech communication. In *The first conference on Applied Digital Human Modeling*, Miami, USA.
Moulin-Frier, C., Schwartz, J., Diard, J., & Bessière, P. (2011). Emergence of articulatory–acoustic systems from deictic interaction games in a "Vocalize to Localize" framework. In *Primate communication and human language: Vocalisations, gestures, imitation and deixis in humans and non-humans. Advances in interaction studies series*. John Benjamins Pub. Co.
Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499.
Nam, H., Goldstein, L. M., Giulivi, S., Levitt, A. G., & Whalen, D. (2013). Computational simulation of {CV} combination preferences in babbling. *Journal of Phonetics*, *41*(2), 63–77.
Ohala, J. (1979). Moderator's introduction to symposium on phonetic universals in phonological systems and their explanation. In *Proceedings of the ninth international congress of phonetic sciences* (Vol. 3, pp. 181–185).
Oliphant, M. (1996). The dilemma of Saussurean communication. *BioSystems*, *37*(1), 31–38.
Oudeyer, P. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, *233*(3), 435–449.
Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*. *Studies in the evolution of language*, Vol. 6. Oxford University Press.
Oudeyer, P.-Y. (2013). *Aux sources de la parole*. Odile Jacob.
Pickles, J. (2012). *An introduction to the physiology of hearing*. Emerald Group Publishing Limited.
Pradalier, C., Colas, F., & Bessiere, P. (2003). Expressing Bayesian fusion as a product of distributions: Applications in robotics. In *International conference on intelligent robots and systems (IROS 2003)* (Vol. 2, pp. 1851–1856). IEEE.
Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, *21*(5), 188–194.
Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research Cognitive Brain Research*, *3*(2), 131–141.
Roy, D. (2005). Semiotic schemas: *A framework for grounding language in action and perception*. *Artificial Intelligence*, *167*(1), 170–205.
Roy, A. C., & Arbib, M. A. (2005). The syntactic motor system. *Gesture*, *5*(1), 7–37.
Ruhlen, M. (1996). *The origin of language: Tracing the evolution of the mother tongue*. New York: John Wiley & Sons.

Schroeder, R., Atal, B. S., & Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, *66*(6), 1647–1652.
Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012a). The Perception-for-Action-Control Theory (PACT): *A perceptuo-motor theory of speech perception*. *Journal of Neurolinguistics*, *25*(5), 336–354.
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. S. Beddor, M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 104–124). Oxford University Press.
Schwartz, J.-L., Boë, L.-J., Badin, P., & Sawallis, T. R. (2012b). Grounding stop place features in the perceptuo-motor substance of speech communication. *Journal of Phonetics*, *40*, 20–36.
Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997a). The dispersion–focalization theory of vowel systems. *Journal of Phonetics*, *25*(3), 255–286.
Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997b). Major trends in vowel system inventories. *Journal of Phonetics*, *25*(3), 233–253.
Serkhane, J. E. (2005). Un bébé androïde vocalisant: Etude et modélisation des mécanismes d'exploration vocale et d'imitation orofaciale dans le développement de la parole (Ph.D. thesis), Grenoble, INPG.
Serkhane, J., Schwartz, J.-L., & Bessiere, P. (2005). Building a talking baby robot: *A contribution to the study of speech acquisition and evolution*. *Interaction Studies*, *6*(2), 253–286.
Serkhane, J., Schwartz, J.-L., Boë, L.-J., Davis, B., & Matyear, C. (2007). Infants' vocalizations analyzed with an articulatory model: *A preliminary report*. *Journal of Phonetics*, *35*, 321–340.
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*(10), 2387–2399.
Steels, L. (1994). The artificial life roots of artificial intelligence. *Artificial Life Journal*, *1*(1), 89–125.
Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, *1*(1), 1–34.
Steels, L. (1999). The spontaneous self-organization of an adaptive language. In S. Muggleton (Ed.), *Machine intelligence*, Vol. 15 (pp. 205–224). Oxford: Oxford University Press.
Steels, L. (2008). The symbol grounding problem has been solved. so what's next. In *Symbols and embodiment: Debates on meaning and cognition* (pp. 223–244).
Stevens, K. (1972). The quantal nature of speech: *Evidence from articulatory-acoustic data*. In E. David, & P. Denes (Eds.), *Human communication: A unified view* (pp. 51–66). McGraw-Hill.
Stevens, K. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*(1), 3–45.
Stevens, K., & Keyser, S. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics*, *38*(1), 10–19.
Studdert-Kennedy, M., & Goldstein, L. (2003). Launching language: *The gestural origin of discrete infinity*. In M. Christiansen, & S. Kirby (Eds.), *Language evolution: The states of the art*. Oxford University Press.
Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: *The orderly output constraint*. *Behavioral and Brain Sciences*, *21*(2), 241–259 discussion 260–299.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: *Statistics, structure, and abstraction*. *Science*, *331*(6022), 1279–1285.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: *The origins of cultural cognition*. *Behavioral and Brain Sciences*, *28*(5), 675–690.
Vallée, N. (1994). Systèmes vocaliques: de la typologie aux prédictions. Grenoble, Université Stendhal: Thèse de Doctorat en Sciences du Langage.
Vallée, N., Rossato, S., & Rousset, I. (2009). Favoured syllabic patterns in the world's languages and sensorimotor constraints. In F. Pellegrino, E. Marsico, I. Chitoran, & C. Coupe (Eds.), *Approaches to phonological complexity* (pp. 111–139). Berlin: Mouton de Gruyter.
Vilain, A., Abry, C., Badin, P., & Brosda, S. (1999). From idiosyncratic pure frames to variegated babbling: Evidence from articulatory modelling. In *Proceedings of the 14th International congress of phonetic sciences* (Vol. 3, pp. 2497–2500).
Volterra, V., Caselli, M. C., Capirci, O., & Pizzuto, E. (2005). Gesture and the emergence and development of language. In M. Tomasello, & D. Slobin (Eds.), *Beyond nature-nurture: Essays in honor of Elizabeth Bates* (pp. 3–40). Mahwah, NJ: Lawrence Erlbaum Associates.
Zuidema, W., & Westermann, G. (2003). Evolution of an optimal lexicon under constraints from embodiment. *Artificial Life*, *9*(4), 387–402.