Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: an exploratory Bayesian modeling study

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

Title:
Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: an exploratory Bayesian modeling study

Journal:
Language and Cognitive Processes (in press)

Authors:
Moulin-Frier, C. (1), Laurent, R. (1,2), Bessière, P. (2,3), Schwartz, J.L. (1), Diard, J. (4)

Affiliations:
(1) GIPSA-Lab, UMR 5216, CNRS-Grenoble University, France
(2) LIG, UMR 5217, CNRS-Grenoble University, France
(3) LPPA, UMR 7152, CNRS-Collège de France, Paris, France
(4) LPNC, UMR 5105, CNRS-Grenoble University, France

Abstract:
In this paper, we put forward a computational framework for the comparison between motor, auditory and perceptuo-motor theories of speech communication. We first recall the basic arguments of these three sets of theories, either applied to speech perception or to speech production. Then we expose a unifying Bayesian model able to express each theory in a probabilistic way. Focusing on speech perception, we demonstrate that under two hypotheses, regarding communication noise and inter-speaker variability, providing perfect conditions for speech communication, motor and auditory theories are indistinguishable. We then degrade successively each hypothesis to study the distinguishability of the different theories in "adverse" conditions. We first present simulations on a simplified implementation of the model with monodimensional sensory and motor variables, and secondly we consider a simulation of the human vocal tract providing more realistic auditory and articulatory variables. Simulation results allow us to emphasize the respective roles of motor and auditory knowledge in various conditions of speech perception in adverse conditions, and to suggest some guidelines for future studies aiming at assessing the role of motor knowledge in speech perception.

Keywords:
Auditory, Motor and Perceptuo-motor theories of speech communication, Bayesian modeling, speech perception in adverse conditions, model distinguishability.

To cite this article:
C. Moulin-Frier, R. Laurent, P. Bessière, J. L. Schwartz & J. Diard (2012): Adverse conditions improve distinguishability of auditory, motor, and perceptuo-motor theories of speech perception: An exploratory Bayesian modelling study, Language and Cognitive Processes, DOI:10.1080/01690965.2011.645313

To link to this article:
http://dx.doi.org/10.1080/01690965.2011.645313

1. Introduction

A central issue in speech science concerns the nature of representations and processes involved in communication. Three major sets of theories have been widely argued for and against in this long-standing debate: motor, auditory and perceptuo-motor theories. Arguments have so far mostly been based on experimental data about variability and invariance (coarticulation phenomena supposedly in favor of motor theories: Galantucci et al., 2006; or motor equivalence principles supposedly in favor of auditory theories: e.g. Guenther et al., 1998, Diehl et al., 2004; or co-structuration of the perceptual and motor repertories in the Perception-for-Action-Control Theory, PACT, Schwartz et al., 2010).

However, none of these observed isolated properties and their associated arguments are decisive, and the theoretical debate appears to be stagnating. It is our belief that mathematical modeling of these theories could provide breakthroughs in this regard. More precisely, we propose that casting these theories into a single, unified mathematical framework would be the most efficient way of comparing the theories and their properties in a systematic manner.

Bayesian modeling is a mathematical framework that precisely allows such comparisons. The trick is that the same tool, namely probabilities, can be used both for defining the models and for comparing them. 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 (Körding et al., 2007), or on theoretical comparison of memory models (Myung and Pitt, 2009). Moreover, the use of a unified framework implies that common hypotheses would have common mathematical translations. This also helps toward more principled studies of the competing theories.

In this paper, we thus cast the motor, auditory and perceptuo-motor theories into three instances of the same Bayesian model, and compare them in varied speech perception situations. Using simulations, we study their behavior both in nominal and adverse conditions. Adverse conditions in this respect could involve various dimensions such as communication in noisy or multi-speaker environments, multisensory binding in complex or incoherent scenes, communication in foreign languages or between various accents, sensory or cognitive deficits. Here, we focus on the effect of noise and speaker variability on performance in speech perception. This allows us to study both the level of performance predicted by each theory but also, and surely more importantly, how the level of noise influences the distinguishability of models. It suggests some predictions for future behavioral studies aiming at assessing the role of motor processes in speech perception.

The rest of this paper is structured as follows. We first recall various illustrations of auditory, motor and perceptuo-motor theories of speech perception and speech production, and organize them into a single unifying Bayesian framework in which they all appear as instances of various questions asked to a single probabilistic communicating agent model. Then, we present the detailed implementation of auditory, motor and perceptuo-motor speech perception models. We present theoretical evidence that the auditory and motor models are indistinguishable in perfect conditions. Finally, we provide a series of simplified simulations of the recognition of speech in adverse conditions, enabling to better assess what could be the respective roles of auditory processing and motor knowledge.

2. A unified theoretical Bayesian framework for speech communication theories

2.1 Theories of speech perception and production

A major question in speech communication studies concerns the nature of the reference frame. Surprisingly, the question is generally asked independently in the speech production and speech perception domains, while the perceptuo-motor link is central in the theoretical debate. Indeed, we can find in the literature various occurrences of motor, auditory and perceptuo-motor theories of both speech production and speech perception (see Table 1; and Moulin-Frier et al., 2010).

2.1.1 Motor theories

Motor theories consider the reference frame of speech communication as gestures. In the context of speech production, Articulatory Phonology (Browman and Goldstein, 1989) models speech motor control as scores of overlapping gestures, able to express the context-dependent variability of speech, without taking explicitly into account the auditory consequence of a motor event. Concerning speech perception, the Motor Theory (Liberman and Mattingly, 1985) proposes that perceiving speech amounts to perceiving gestures. A main argument is coarticulation-driven signal variability, which makes the auditory content of a given phoneme dependent on the phonetic context (see a review in Galantucci et al., 2006).

The interest for the Motor Theory of Speech Perception was recently renewed by the discovery of mirror neurons in monkeys (see e.g., Rizzolatti et al., 1996) and of a "mirror system" in humans (Fadiga et al., 2002).

2.1.2 Auditory theories

Auditory Theories consider that the reference frame for speech is auditory. In the case of speech production, the target would be a region in the auditory space (Guenther et al., 1998). The main argument is motor equivalence, showing that various articulatory configurations are used for achieving the same auditory goal, as shown for instance in perturbation experiments. In the case of speech perception, proponents of auditory theories consider that speech perception involves auditory or multisensory representations and processing, with no reference to speech production (Diehl et al., 2004).

In terms of neuroanatomical correlates, it remains a question to know if the sensorimotor connection between temporal auditory and audiovisual regions, parietal somesthesic/proprioceptive areas and frontal motor and premotor zones inside a dorsal cortical network plays or not a significant role in speech comprehension (Hickok & Poeppel, 2007; Scott et al., 2009).

2.1.3 Perceptuo-motor theories

Perceptuo-motor theories have recently emerged for both speech perception (*e.g.*, Skipper et al., 2007, Schwartz et al., 2010) and production (*e.g.*, Guenther, 2006). They generally consider auditory frames as the core for communication, but they include the perceptuo-motor link inside the global architecture. They claim that in normal conditions, production involves cortical motor (frontal) areas and perception involves cortical auditory (temporal) ones, but that the perceptuo-motor (dorsal) link, necessary for speech acquisition, could also play a role in adverse conditions.

Regarding speech production, the DIVA model (Guenther, 2006) combines a feedforward control sub-system for on-line production, and a feedback control sub-system when the auditory consequence of a gesture is not congruent. In a similar way, perceptuo-motor theories of speech perception argue for a core auditory (or audio-visual) system for speech perception, enhanced by motor processes in complex conditions such as noise, through "binding" (Schwartz et al., 2010) or "prediction" (Skipper et al., 2007).

2.2 Communicating agents based on internalization of the communication loop

We propose here to model a general interaction process, in which a speaker and a listener communicate about an object of the environment. For this aim, the speaker, willing to communicate about the object O^S, performs a gesture M producing a sound S enabling the listener to understand and recover an object O^L (Fig. 1a). Efficient communication can be assessed by an external validation system (success vs. failure indicator C_{env}), whether it is an outside oracle (as in a supervised learning stage, for instance), or a shared attention mechanism.

A central hypothesis of the general model is that the communication loop (Fig. 1a), in which a speaker interacts with a listener through the environment, can be internalized and emulated into the brain of a single agent (Fig. 1b). Firstly, the agent can take both roles, listener and speaker, and thus contains both subsystems, motor and sensory. Secondly, the agent has 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 sensory consequences of motor gestures. Finally, the external validation system is also internalized; in other words, the agent has two internal representations of objects, linked by a system that verifies whether they refer to the same object.

This internalization hypothesis could be discussed in the framework of general cognitive theories of social communication and human evolution (e.g. Baron Cohen, 1995; Tomasello et al., 2005; see also Moore, 2007, for similar views about internalization, expressed in a control theory framework).

2.3 Bayesian model of communicating agents

We now propose a computational model of the communicating agent defined according to the internalized communication loop in Fig. 1b. The model is built using Bayesian Programming (Lebeltel et al., 2004; Bessiere et al., 2008). Bayesian Programming aims at defining models of reasoning agents using probability distributions to represent knowledge and Bayesian inference to manipulate knowledge in a mathematically principled manner. This view of probabilities as states of knowledge allows one to formally represent different preliminary knowledge (Colas et al., 2010). In the remainder, we use symbols to identify different models (*e.g.*, in π_{Env}, the model of the simulated environment, some noise is represented, which may or may not be properly captured in some agent's internal model π_{Ag}).
In this context, the model of a perceptuo-motor agent is defined as a joint probability distribution over variables of interest (typically motor, sensory and internal variables). This joint distribution is broken down as a product of simpler distributions, using Bayes rule and conditional independence hypotheses. Using this knowledge, a behavior is then defined as a conditional probability distribution computed from the joint distribution (for example: "given values of some sensory variables, what is the probability distribution over the speech objects?"),

called a *question* to the model, and is solved using Bayesian inference.

2.3.1 Variables

Our model of a general communicating agent is based on five probabilistic variables, which are a direct translation of our conceptual model of the communication situation:

- M_{Ag}: the agent motor gesture,
- S_{Ag}: the agent sensory representation,
- O^S_{Ag}, O^L_{Ag}: the object of communication, when the agent respectively takes the speaker's and the listener's point of view,
- C_{Ag}: the internalization of the communication validation system. C is a Boolean variable, and is true (value 1) when $O^S_{Ag} = O^L_{Ag}$.

Here it must be acknowledged that an "object" may take a very wide spectrum of definitions. Ideally, it refers to the level at which the link between phonology and semantics takes place in linguistic communication: typically the word. In the remainder of this paper, it however refers to any phonological category shared by the speaker and the listener in the speech communication process (*e.g.*, a syllable or a phoneme). We do not consider in this paper all the (many) problems associated with defining the adequate level. We just assume that at some stage, a given phonological unit can be successfully exchanged between a speaker and a listener.

2.3.2 Joint probability distribution

There are five subsystems to define. The first three depend on the precise definition of the sensory and motor variables: they are introduced here and made precise in Section 3.2.

- The motor subsystem is defined as a conditional probability distribution $P(M_{Ag} \mid O^S_{Ag})$: given an object to communicate, what is the probability distribution over motor gestures?
- The sensory (or auditory) subsystem is defined as a conditional probability distribution $P(O^L_{Ag} \mid S_{Ag})$: given a sensory input, what is the probability distribution over the perceived objects?
- The perceptuo-motor subsystem is defined as a conditional probability distribution $P(S_{Ag} \mid M_{Ag})$: given a motor gesture, what is the probability distribution over the sensory inputs?

The last two systems are independent of sensory and motor variables, and defined as follows.

- The internalized communication validation subsystem is defined as a conditional probability distribution $P(C_{Ag} \mid O^S_{Ag}\ O^L_{Ag})$: given objects in the motor and sensory subsystems, C_{Ag} is true (=1) when both refer to the same object. It is defined as a Dirac probability distribution such that:
  $$P([C_{Ag} = 1] \mid [O^S_{Ag} = X]\ [O^L_{Ag} = Y]) = \begin{cases} 1 \text{ if } X = Y \\ 0 \text{ otherwise} \end{cases}.$$
  Technically, C_{Ag} is a coherence variable (Gilet et al., 2011), which allows connecting or disconnecting the sensory and motor routes in the agent's cognitive architecture.
- Finally, the object prior $P(O^S_{Ag})$, is defined as a Uniform probability distribution: the objects are assumed to be present with the same frequency in the environment.

Therefore, the general model of a communicating agent π_{Ag} is the following joint probability distribution, illustrated in Fig. 2:

$$P(O^S_{Ag}\ M_{Ag}\ S_{Ag}\ O^L_{Ag}\ C_{Ag})$$
$$= \quad P(O^S_{Ag})P(M_{Ag} \mid O^S_{Ag})P(S_{Ag} \mid M_{Ag})P(O^L_{Ag} \mid S_{Ag})P(C_{Ag} \mid O^S_{Ag}\ O^L_{Ag}) \ .$$

2.3.3 Bayesian inference for simulating speech perception and production tasks

From the joint probability distribution $P(O^S_{Ag}\ M_{Ag}\ S_{Ag}\ O^L_{Ag}\ C_{Ag})$ we can apply Bayesian inference in order to simulate speech perception and production tasks, which appear as probabilistic questions addressed to the general distribution. Perception tasks can be simulated by computing probability distributions over objects, given an input sensory signal. Production tasks can be simulated by computing probability distributions over motor gestures, given an object to communicate about.

The driving reasoning in our Bayesian modeling is that motor vs. auditory theories can be defined in reference to the pivot role provided to O^S_{Ag} vs. O^L_{Ag} in the probabilistic reasoning.

In motor theories, the speaker is the pivot, and the direct connection between S_{Ag} and O^L_{Ag} is disabled.
In terms of speech production, this provides a simple probabilistic question to solve for speech motor control, that is $P(M_{Ag} \mid O^S_{Ag})$: what is the adequate action for a given speaker?
But in terms of speech perception, this leads to a "motor theory of speech perception" in which the question to solve is $P(O^S_{Ag} \mid S_{Ag})$: knowing the sensory input, what is the object which was in the speaker's mind? Bayesian inference yields:
$$P(O^S_{Ag} \mid S_{Ag}) = \sum_{M_{Ag}} P(M_{Ag} \mid O^S_{Ag})P(S_{Ag} \mid M_{Ag}) \ .$$

This question can be interpreted as a motor inference: indeed, motor terms are involved in the equation, whereas the sensory system $P(O^L_{Ag} \mid S_{Ag})$ is marginalized away. In this equation, the $P(S_{Ag} \mid M_{Ag})$ term expresses the search for motor values able to lead to the perceived sensory input (this is classically referred to as "inversion" or "analysis by synthesis"). The $P(M_{Ag} \mid O^S_{Ag})$ factor can be conceived as an "articulatory decoder", assuming that invariance rather lies in motor than in auditory cues.

In auditory theories, the listener is the pivot, and the direct connection between O^S_{Ag} and M_{Ag} is disabled.
This leads to auditory theories of speech perception through the question $P(O^L_{Ag} \mid S_{Ag})$ (direct inference without any motor knowledge, typical of these theories).
It leads to auditory theories of speech production through $P(M_{Ag} \mid O^L_{Ag})$ attempting to estimate through auditory inference what gestures should the speaker produce to make the listener perceive the adequate object for the listener. Bayesian inference yields:
$$P(M_{Ag} \mid O^L_{Ag}) \propto P(M_{Ag}) \sum_{S_{Ag}} P(S_{Ag} \mid M_{Ag})P(O^L_{Ag} \mid S_{Ag}) \ .$$

This corresponds to associating auditory targets, defined by the term $P(O^L_{Ag} \mid S_{Ag})$, with forward models, defined by the term $P(S_{Ag} \mid M_{Ag})$, for estimating the adequate motor command.

Finally, perceptuo-motor theories take into account the information provided by both motor and sensory subsystems. Furthermore, inference is performed under the assumption that both

subsystems are coherent, that is, $C_{Ag}=1$. For example, for speech perception this yields: $P(O^L_{Ag} \mid S_{Ag}\ [C_{Ag}=1]) \propto P(O^L_{Ag} \mid S_{Ag}) \sum_{M_{Ag}} P(M_{Ag} \mid O^S_{Ag}) P(S_{Ag} \mid M_{Ag})$.

Note that $P(O^L_{Ag} \mid S_{Ag}\ [C_{Ag}=1]) = P(O^S_{Ag} \mid S_{Ag}\ [C_{Ag}=1])$: both O^L_{Ag} and O^S_{Ag} can interchangeably be considered pivot during this perceptuo-motor inference.

This can be seen as a combination of the two previous inferences. The probability of the object O_L directly inferred from the sensory input S through an auditory theory is modified by the knowledge that S was also produced from the object O_S (= O_L) using motor variables M. In other words, the coherence variable effect can be seen here as forcing a fusion between purely perceptual and purely motor inferences. In our case, this perceptuo-motor fusion takes the form of a product between both processes.

A similar result is obtained for speech production.

This results in the taxonomy of Table 2 in which each of the speech perception and production theories displayed in Table 1 actually corresponds to different inferences in the same probabilistic model, or, alternatively, processes involving different portions of a unique knowledge set. The unifying process in this framework is that both perception and production models derive from a single internalized communication model, in which the coherence variable C_{Ag}, expressing the hypothesis of internal coherence, yields a fusion between the motor and perceptual branches for both perception and production. Pure auditory and motor theories just consist in cutting this connection and keeping only either the perceptual or the motor branch for both speech production and perception.

3. Studying perception tasks using the Bayesian model of communicating agent

We now narrow down our focus to the study of perception tasks. We consider an environment, populated with several agents (see Fig. 3). We take the point of view of one of the agents, *Ag*, which performs perception tasks. It interacts with another agent, *Master*, which provides it with both learning and test signals. Both agents are instances of the communicating agent model we defined (Section 2).

Our goal is to compare the three auditory, motor and perceptuo-motor variants of speech perception, considered as probabilistic questions and displayed in the right column of Table 2, asked to the agent *Ag*.

We consider in Section 5 two agent models of increasing complexity and realism (in terms of the forms of all variable distributions). The first one, however, already contains all properties that are needed to justify our first two theoretical results, and is also the basis of the first simulation experiment in Section 5.1. The other model can thus be seen as a variant of this first, core model. Therefore, we now provide its full detailed definition, and only present the variant as needed in Section 5.2.

3.1. Variable domains

The domains of M_{Ag}, M_{Master}, S_{Ag} and S_{Master} vary in our different experiments. In the first, simple model, *they* are mono-dimensional, discrete variables (with values regularly distributed between -10 and 10). The object variables O^S_{Ag}, O^S_{Master}, O^L_{Ag} and O^L_{Master} each denote two possible speech objects: *o*+ and *o*-.

3.2. Probability distribution forms

The model structure being set according to Section 2.3.2, we define each of the probabilistic terms of interest that are implied by Fig. 3.

3.2.1. Object prior $P(O^S_{Ag})$

The first term of the decomposition is a prior probability distribution over objects. In our first experiment, the set of possible objects contains two values, and we define both $P(\times)$ and $P(O^S_{Master})$ as uniform probability distributions over this set. In other words, *o*+ and *o*- each have a prior probability of 0.5 to be the communication object.

3.2.2. Motor models $P(M_{Ag} \mid O^S_{Ag})$ and $P(M_{Master} \mid O^S_{Master})$

The terms $P(M_{Ag} \mid O^S_{Ag})$ and $P(M_{Master} \mid O^S_{Master})$ are probability distributions over motor gestures, one for each possible communication object. Their definitions depend on the dimensionality and nature of the motor variable. In our first experiment, motor variables are discrete and mono-dimensional; therefore they are defined as 1-D Gaussian probability distributions, approximated over the discrete domain of M (Fig. 4).

3.2.3. Motor to acoustic mappings $P(S_{Ag} \mid M_{Ag})$ and $P(S_{Ag} \mid M_{Master})$

The relation between motor gestures and the resulting speech sounds, characterized by the formants of the acoustic signal, is known to feature non-linear events. Proponents of sensory theories of speech perception argue that processing the acoustic signal is easier in the acoustic domain than back in motor space, because non-linearities would naturally structure the acoustic domain into plateaus and boundaries (Quantal Theory, Stevens, 1972, 1989: see Fig. 5a).

In our model, terms of the form $P(S \mid M)$ encode the articulatory-to-acoustic transformation. They are defined as sets of mono-dimensional Gaussian probability distributions, one for each discrete motor value, defined by parameters μ and σ. μ is given by a sigmoid function of M, defined by

$$\mu(M) = b\frac{\arctan(aM)}{\arctan(ab)} \ ,$$

and σ models a dispersion around the central value predicted by the sigmoid. The parameter a controls the linear vs. nonlinear nature of the μ(M) function (Fig. 5b), whereas b controls the range of the function.

Two articulatory-to-acoustic mappings are to be defined. The first, $P(S_{Ag} \mid M_{Master})$, is the one performed by the simulated environment model π_{Env}, of parameters μ_{Env}, σ_{Env}: σ_{Env} models the noise in the simulated environment. The second, $P(S_{Ag} \mid M_{Ag})$, is the one internalized in the agent model π_{Ag}, of parameters μ_{Ag}, σ_{Ag}: σ_{Ag} models the uncertainty the agent has about the articulatory-to-acoustic transformation (*e.g.*, when $\mu_{Env} = \mu_{Ag}$, and $\sigma_{Env} = \sigma_{Ag}$, the agent has perfect knowledge about the environment characteristics).

3.2.4. Auditory model $P(O^L_{Ag} \mid S_{Ag})$

In our framework, the auditory model $P(O^L_{Ag} \mid S_{Ag})$ of the agent *Ag* is assumed to be learned in a supervised manner.

The master agent provides object-to-acoustic signal couples to the learner using the following algorithm:

- we first choose a random object *o* in the domain of O^S_{Master} to communicate,
- we compute, in model π_{Master}, a motor gesture *m* by drawing at random according to the master agent's production model $P(M_{Master} \mid [O^S_{Master}=o])$,
- we compute, in model π_{Env}, from *m*, a resulting auditory signal *s*, by drawing at random according to the environment articulatory-to-acoustic model $P(S_{Ag} \mid [M_{Master}=m])$,
- the object *o* is then transferred without error to the learner, that is, $C_{Env}=1$, assuming a perfect shared-attention mechanism, outside of spoken communication, that allows to agree on the object of interest. The learning agent takes the object *o* as an O^L_{Ag} value.

Thus the learner can, using a history of such couples <*s, o*>, identify the parameters of the probability distributions of $P(O^L_{Ag} \mid S_{Ag})$.

4. Perfect communication leads to theoretical equivalence between motor and auditory models

In this section, we consider the motor and auditory processes of speech perception, as defined above. We put forward two hypotheses on the learner agent knowledge, under which motor and auditory speech perception models cannot be distinguished.

Let us consider the purely auditory speech perception of the learner. As defined before, it is performed by simply accessing the auditory model $P(O^L_{Ag} \mid S_{Ag})$. But this model, when learned in the supervised manner we defined, is based on <*s, o*> couples that have been generated by the master agent, using its motor model (and the model of acoustic propagation through the environment). Mathematically, our learning algorithm of Section 3.2.4 performs the following computation:

$$P(O^L_{Ag} \mid S_{Ag}) \propto \sum_{M_{Master}} P(M_{Master} \mid O^S_{Master}) P(S_{Ag} \mid M_{Master}) \ .$$

We now add two hypotheses:

- H1: $P(S_{Ag} \mid M_{Ag}) = P(S_{Ag} \mid M_{Master})$: we assume that the learner perfectly captured the environment articulatory-to-acoustic transformation;
- H2: $P(M_{Ag} \mid O^S_{Ag}) = P(M_{Master} \mid O^S_{Master})$: the learner and master agent have the same motor model.

Under these hypotheses, we obtain:

$$P(O^L_{Ag} \mid S_{Ag}) \propto \sum_{M_{Ag}} P(M_{Ag} \mid O^S_{Ag}) P(S_{Ag} \mid M_{Ag}) \ .$$

In the last equation, the left member is the expression of an auditory theory of speech perception, while the right member is the expression of a motor theory of speech perception (see Table 2). This shows that auditory and motor models of speech perception are indistinguishable

in "perfect" communication conditions, that is with identical motor models in the agents, and perfect learning of the environment properties. This is the first result of our Bayesian approach.

The models can be distinguished only if either H1 or H2 is wrong: this is what we consider here as "adverse conditions". Departure from H1 (imperfect forward model) can be produced in various ways. The learner could have a limited learning capability, with not enough memory or inadequate expressiveness. Or it could have an inadequate model of the articulatory-to-acoustic transformation, for example in cases where the noise level assumed by the learner is different from the noise level during subsequent perception. Departure from H2 with different motor models in the master and the learner could be due to any biomechanical or idiosyncrasy differences. Of course, speech impediment or accented speech would provide other, more severe, discrepancies.

In summary, this theoretical result shows that, outside of adverse conditions (incomplete or imperfect learning, noisy environment, degraded or accented speech), auditory and motor processes of speech perception cannot be distinguished in the Bayesian framework presented here[1].

5. Simulating speech perception in adverse conditions

To achieve distinguishability, we consider two kinds of degradation of the communication paradigm, which correspond to departures from the two hypotheses H1 and H2.

The first one involves communication noise in the environment. To do so, we vary the value of σ_{Env} which controls dispersion in the $P(S_{Ag} \mid M_{Master})$ articulatory-to-acoustic transformation (Section 3.2.3). If σ_{Env} equals zero, $P(S_{Ag} \mid M_{Master})$ becomes deterministic, which means that S_{Ag} is completely determined by M_{Master} in a communication occurrence. Increasing σ_{Env} increases the uncertainty on S_{Ag}, which simulates communication noise. The second one is based on imposing differences in motor prototypes between π_{Ag} and π_{Master} in the $P(M \mid O^S)$ distributions (Section 3.2.2).

We have five expectations in these simulations.

(1) According to H1, incorrect representation of the noise will induce distinguishability between models.

(2) According to H2, difference between learner and master agent motor prototypes will also induce distinguishability between models.

(3) Applying a nonlinear $P(S \mid M)$ articulatory-to-acoustic transformation results in a natural organization of the S space in two categories (see Fig. 5). An auditory model of speech perception is naturally suited to this structure. On the contrary, a motor model first inverses the sensory to motor representations and hence loses this structure (this is the basic argument which led Schwartz et al. (2010) to discard a purely motor theory of speech perception). Therefore, at small to moderate noise levels, the auditory model of speech perception should surpass the motor model of speech per-

[1] Notice that the perceptuo-motor model is actually different from both the auditory and the motor models here. Indeed, since it is based on a fusion of the perceptual and motor inputs, it questions twice the classifier $P(O^L \mid S)$ (equal to $P(O^S \mid S)$ in this situation) and keeps only coherent answers, thus decreasing the error rate.

ception.

(4) Noise should degrade the auditory model. However, motor knowledge represented in $P(M_{Ag} \mid O^S_{Ag})$ contains some information that could be likely to compensate – to a certain extent – for the degradation. Therefore, it may be hypothesized that at high noise levels, the motor model of speech perception should surpass the auditory model of speech perception (this is the basic argument which led Schwartz et al. (2010) towards a perceptuo-motor theory of speech perception, PACT).

(5) The perceptuo-motor model, combining auditory and motor branches for a global decision process (Section 2.3.3), should surpass both the auditory and the motor model at all noise levels.

We first test these expectations on the core 1D model, and then consider stimuli generated by a realistic articulatory-to-acoustic model of the human vocal tract.

5.1. 1D core model

Let us recall that in this first variant, sensory and motor variables are one-dimensional varying between arbitrary values -10 and 10, and perception deals with two speech objects *o*+ and *o*-.

5.1.1. Varying communication noise

In this first simulation, we test how auditory, motor and perceptuo-motor models of speech perception perform at various levels of communication noise (σ_{Env} varying from 0 to large values) for three different values of the slope of the $P(S_{Ag} \mid M_{Master})$ transform, producing transforms from quasi linear to highly nonlinear (Fig. 5b).

For each model:

- Objects are drawn with equal probabilities from both categories *o*+ and *o*-.
- From these objects, values for S_{Ag} are drawn according to the master production system $P(M_{Master} \mid O^S_{Master})$ followed by the articulatory-to-acoustic transformation of the environment $P(S_{Ag} \mid M_{Master})$ (including the noise σ_{Env}).
- They are given as input to the probabilistic question of the model (a motor, sensory, or sensorimotor question, corresponding to probability distributions displayed in the right column of Table 2).
- From the comparison of the probability distribution of the answers with the known category of the object for all the drawn values of S_{Ag}, a confusion matrix is built, from which we define the correct recognition rate as the sum of the diagonal terms. This score describes the probability that the model recognizes the right category for any input S_{Ag}.

The results are displayed on Figure 6.

We observe the following results.

- The superiority of the motor or sensory model over the other one depends on both the value of the noise and the amplitude of the nonlinearity in the $P(S_{Ag} \mid M_{Master})$ transform.
- As predicted, without noise (*i.e.*, when $\sigma_{Env} = \sigma_{Ag}$, that is, hypothesis H1 is true) and since motor prototypes are equal (hypothesis H2) the auditory and motor models perform exactly the same.
- In the linear case, the sensory model is always poorer than the motor one. Nonlinearity

induces a range (for small noise) in which the sensory model performs better than the motor one. A high value for the slope parameter (highly nonlinear transform) makes the sensory model better than the motor model on a larger range.

- The motor model becomes better than the sensory model when noise increases.
- The sensorimotor model has always equal or better recognition scores than the two other models.

5.1.2. Varying motor prototypes between master and learner agent

In this second simulation with a constant small communication noise (σ_{Env} = 1), we considered differences in master and learner agent motor prototypes by applying different values of means μ of the Gaussian probability distributions $P(M_{Master} \mid O^{S}_{Master})$ in the master (used for learning the $P(O^{L}_{Ag} \mid S_{Ag})$ distribution, see Section 3.2.4) and $P(M_{Ag} \mid O^{S}_{Ag})$ in the learner. We note μ^{+}_{Master}, μ^{-}_{Master} and μ^{+}_{Ag} and μ^{-}_{Ag} the means for the two categories *o*+ and *o*- respectively. We set μ^{+}_{Master}=5 and μ^{-}_{Master}=-5, and we suppose that the motor prototypes for the Agent differ from them by a δ value: $\mu^{+}_{Ag} = \mu^{+}_{Master} + \delta$, $\mu^{-}_{Ag} = \mu^{-}_{Master} - \delta$. A negative (resp. positive) value of the δ parameter means that the learner's motor prototypes are closer (resp. further) than the master's. When $\delta=0$, motor prototypes of *Master* and *Ag* are identical (*i.e.*, hypothesis H2 is true). Applying a delta bias between the means of the motor prototypes in the master and the learner could be interpreted as a motor idiosyncrasy or an accent. Once again, we considered three different values for the slope parameter of the $P(S_{Ag} \mid M_{Master})$ transform, from quasi linear to highly nonlinear. The results are displayed on Fig. 7, with the same presentation as in Fig. 6.

These show that:

- The superiority of the motor or sensory model over the other one depends on the discrepancy between motor prototypes in the master and the agent.
- The sensory model does not depend on values of the agent motor prototypes (since sensory inputs are the same in all cases) while the motor model improves when the distance between the agent motor prototypes increases (better internal separation between motor prototypes). Interestingly, this shows that in a motor theory of speech perception, the best perceptual performances are obtained for an agent dispersing its motor prototypes as much as possible, rather than perfectly fitting with its interlocutor.
- A high value for the slope parameter (highly nonlinear transform) makes the sensory model better than the motor model on a larger range.
- As in the previous simulation, the sensorimotor model has always equal or better recognition scores than the two other models.

5.2. Simulations on stimuli provided by a realistic model of the human vocal tract

We now extend the previous simulations to more realistic articulatory-acoustic configurations generated by a realistic model of the human vocal tract. This model, developed by Maeda (1989), delivers sagittal contours and lip shapes from seven input parameters driving the jaw, tongue and lips, which are interpretable in terms of phonetic commands and are closely related to muscle commands (Maeda & Honda, 1994). After the control parameters are entered, the area function of the vocal tract is estimated, from which the transfer function and the formants are calculated (Badin & Fant, 1984).

In the following, we replicated the communication noise experiment (described in Section 5.1.1) on two vowel corpora produced with this vocal tract model. For each corpus, a single articulatory parameter controls the acoustic (F1, F2, F3, F4) trajectory between two vowels. Then, F2, F3 and F4 are summarized by a "perceptual second formant" F'2 computed by an analytical formula derived from experimental data on vowel perception (Schwartz et al., 1997). Therefore, each corpus associates a single articulatory variable to a two-dimensional (F1, F'2) perceptual space. Objects all along Section 5.2 are phonemes (*i.*e. vowel categories).

5.2.1. An [i]-[y] corpus with highly nonlinear articulatory-to-acoustic transform

The first corpus is generated by varying the lip shape (by an adequate combination of lip height and lip protrusion) from the high front unrounded [i] to the high front rounded [y]. This produces a well-known trajectory in which the resonances of the back cavity do not change while only the resonance of the front cavity decreases from a high F3 value close to F4, to a low F2 value. F'2 enhances and summarizes this perceptual variation by providing a quasi quantal variation (Fig. 8a).

We generated with the vocal tract model two sets of vocalic configurations respectively around [i] and [y], also displayed on Fig. 8a, and we applied the auditory, motor and perceptuo-motor speech recognition models to these data, in the same conditions as in Section 5.1.1 – including communication noise σ_{Env} – though sensory variables were now bi-dimensional (F1, F'2) rather than mono-dimensional. The results are displayed in Fig. 8b. They are quite similar with those of Fig. 6f, with a large range of noise values for which the sensory model performs much better than the motor one, while the perceptuo-motor is always the best one.

5.2.2. An [y]-[a] corpus with quasi-linear articulatory-to-acoustic transform

The second corpus is generated by varying the jaw and tongue configuration from high to low (by an adequate combination of so-called "mandible" and "tongue body" parameters) from the high front rounded [y] to the low [a]. This produces a trajectory in which F1 increases and F2 decreases, while F3 does not change much, resulting in a smooth F'2 decrease (Fig. 9a).

Once again, we generated with the vocal tract model two sets of vocalic configurations respectively around [y] and [a], also displayed on Fig. 9a, and we applied the auditory, motor and perceptuo-motor speech recognition models to these data, in the same conditions as in Section 5.1.1 – including communication noise σ. The results are displayed in Fig. 9b. They are quite similar to those of Fig. 6b, the sensory model performing always worse than the motor one, which is almost equal to the perceptuo-motor one.

6. General discussion

6.1. Perceptuo-motor interactions conceived as a fusion problem

The integrative approach that we propose here, in which the basic ingredients of contrastive theories of speech perception and speech production are embedded in the same Bayesian framework for better comparative assessment, seems highly productive, and already results in four important achievements in this paper.

(1) Auditory, motor and perceptuo-motor theories of speech perception and production are described, within a general model of speech communication (Fig. 1), and formalized

through a set of probabilistic equations (Table 2) allowing further quantitative tests in various paradigms.

(2) In this probabilistic framework, focusing on speech perception theories, it appears that in perfect communication conditions (without any kind of noise and with perfect agreement between perceptuo-motor properties of communicating agents) auditory and motor theories of speech perception provide exactly the same outputs and are hence experimentally indistinguishable.

(3) If communication noise or inter-agent variability of any kind is introduced, auditory and motor theories become distinguishable. For small amounts of noise, the auditory model performs better than the motor one. The noise range where this happens is increased when the articulatory-to-acoustic transform is nonlinear. This may be linked to Stevens' quantal theory. On the contrary, for large noise levels, the motor model performs better than the auditory one.

(4) The perceptuo-motor model of speech perception, based on a multiplicative fusion of the auditory and motor models, performs always better than both the auditory and motor models whatever the experimental conditions.

The conception of perceptuo-motor models of speech perception as operating a fusion between an auditory and a motor sub-model provides an interesting perspective for a new paradigm in the study of speech perception. Multisensory fusion in audiovisual speech has generated a large number of experimental data (typically on conflicting auditory and visual inputs and on speech in noise) and computational models and simulations (typically varying the kind of fusion models and the consequence on predictions and performances). The present work suggests that the question of *perceptuo-motor fusion* could be set at the center of the agenda of future experimental and computational studies on auditory vs. motor theories of speech perception. We elaborate a little more on this in the next sections.

6.2. Implications and predictions for further experimental tests

This general portrait sheds some light on the debate between theories of speech perception. To begin with, it provides a formal basis for better understanding why it is so difficult to disentangle auditory from motor theories: the motor knowledge is stored to a certain extent in the auditory model $P(O^L \mid S)$, hence if the learning and communication conditions are "perfect" in the sense of H1 and H2, experimental data cannot conclude. This formalizes in a rigorous way an argument explicitly stated by auditory theorists, *e.g.*, Diehl et al. (2004): "listeners do not recover gestures, but they do perceive the acoustic consequences of gestures. Any regularities of speech production (*e.g.*, context dependencies) will be reflected in the acoustic signal, and, through general mechanisms of perceptual learning, listeners come to make use of the acoustic correlates of these production regularities in judging the phonemic content of speech signals."

It also suggests some directions for further experimental tests comparing theories. Let us mention some of these proposals derived from testable predictions suggested by the results in Section 5. Firstly, linear vs. nonlinear configurations produce very different predictions in auditory vs. motor models: while the degradation with noise is similar in both cases in the motor model, it produces very different effects in the auditory model, with a larger plateau and a steeper decrease in the nonlinear case (compare Fig. 6f with 6b, or 8b with 9b). Comparing degradation of phoneme categorization with noise in linear vs. nonlinear cases could provide some hints in the debate between auditory vs. motor theories of speech perception.

Secondly, the role of motor processes should be easier to display for noisy communication rather than for clear speech: this is actually in line with some recent data assessing how perturbations of the motor system intervene in speech perception, and displaying effects only when speech stimuli are contaminated with noise (*e.g.*, Meister et al., 2007; d'Ausilio et al., 2009; Sato et al., in press). Interestingly, comparisons of perceptuo-motor vs. auditory models in Fig. 6, 8 and 9 show that the role of motor processes in noise appears sooner in linear than in nonlinear configurations: that is, if motor processes do intervene in *e.g.* vowel perception, lip perturbation in assessing vowel rounding in an [i]-[y] task should produce less effect than tongue or jaw perturbation in assessing vowel height in an [i]-[e] or [y]-[a] task.

6.3. Perspectives

A perceptuo-motor theory such as developed in the Perception-for-Action-Control Theory seems to provide a quantitatively efficient computational compromise between properties of the auditory system (non-linear shaping) and exploitation of motor procedural knowledge in perception (see Viviani & Stucchi, 1992, for the first introduction of the concept of "motor procedural knowledge" in human perception). This also fits an increasing number of such proposals, developed in the context of both behavioral and neuroanatomical data (e.g. Skipper et al., 2007).

The work presented in this paper can be extended and developed in various directions. Firstly, we are presently beginning to test more realistic and complex speech perception tasks, involving complete vowel, consonant or syllabic sets associated with multi-dimensional articulatory and perceptual spaces defined on the vocal tract model presented in Section 5.2. We expect that contextual effects should enable to better display the role of articulatory knowledge in speech perceptual processing, typically concerning coarticulation. For example, in syllables involving the same plosive embedded in various vowel contexts, the motor information on the plosive, naturally captured in the motor sub-system, should result in enhancing the plosive recognition score in a perceptuo-motor model compared with a pure auditory model.

Secondly, more sophisticated paradigms and models should be explored in various directions. Processing of atypical stimuli (*e.g.*, produced by a speaker with a specific idiosyncrasy, or a foreign accent) will involve three-agent paradigms in which there is a perceiving agent, a master agent (providing prototypical stimuli in learning) and a speaker agent, all possibly different. Comparison of perceptuo-motor fusion models could be also explored in various directions. Since fusion is basically obtained in the perceptuo-model by the coherence variable C set to 1 by a Dirac probability distribution, other types of fusion could be obtained by relaxing this Dirac into a softer fusion mechanism, possibly driven by additional information such as noise or context (see Schwartz et al., 1998, for a discussion on fusion models in the context of audiovisual speech perception).

Thirdly, we shall design auditory, motor and perceptuo-motor models of speech production associated with the equations in the left column of Table 2, and study their ability to deal with classical speech production questions such as motor equivalence, coarticulation, perturbation, or learning.

The models should also be extended to questions about the behavioral and neuroanatomical correlates of each component of the Bayesian implementation. This relates to such questions as how and where are represented sensory and motor variables, how are they linked in perception and production, how can the potential role of auditory knowledge in speech production

and motor knowledge in speech perception be assessed, are the phonological input (related to O^L) and output (related to O^S) systems equivalent and/or identical (see evidence for a distinction between these systems in Jacquemot et al., 2007). Notice that the computational Bayesian models of speech communication introduced in this paper can also be extended towards models of the emergence of language in societies of computational sensori-motor agents in interaction (Moulin-Frier et al., 2011). This should enable to better link what we know about online speech communication with our knowledge about the shape and evolution of sound systems in human languages (*e.g.*, Schwartz et al., 1997).

References

Badin, P., & Fant, G. (1984). Notes on vocal tract computations, *STL-QPSR,* **2-3**, 53–108.

Baron-Cohen, S, (1995). *Mindblindness: an essay on autism and theory of mind*. MIT Press/Bradford Books.

P. Bessière, C. Laugier, R. Riegwart, (Eds). *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of Springer Tracts in Advanced Robotics. Springer-Verlag, Berlin (2008).

Browman, C.P., & Goldstein, L. (1989). Articulatory Gestures as Phonological Units. *Phonology*, 6, 201–251.

Colas, F., Diard, J., & Bessière, P. (2010). Common bayesian models for common cognitive issues. *Acta Biotheoretica*, 58(2-3):191–216.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, 19, 381–385.

Diehl, R.L., Lotto, A.J. & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*, 74 , 431–461.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur J Neurosci*, 15, 399–402.

Galantucci, B., Fowler, C.A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13, 361–377.

Gilet , E., Diard, J., & Bessière, P. (2011). Bayesian Action–Perception computational model: interaction of production and recognition of cursive letters. *PLoS ONE*, 6(6), e20387.

Guenther, F.H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105, 611–633.

Guenther, F.H., (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39, 350–365.

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

Jacquemot, C., Dupoux, E., & Bachoud-Levi, A.C. (2007). Breaking the mirror: Asymmetrical disconnection between the phonological input and output codes. *Cogn Neuropsychol*, *24*, 3–22.

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.

Lebeltel, O., Bessière, P., Diard, J., & Mazer, E. (2004). Bayesian Robots Programming. *Autonomous Robots*, 16(1), 49–79

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

Maeda, S. (1989) Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model. In *Speech Production and Modelling*, W.J. Hardcastle, A. Marchal (Eds), Academic Publishers, Kluwer, 131–149.

Maeda, S., & Honda, K. (1994) From EMG to formant patterns of vowels: the implication of vowel spaces. *Phonetica*, **51**, 17–19.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology*, 17, 1692–1696.

Moore R K. (2007). Spoken Language Processing: Piecing Together the Puzzle. *Speech Communication*, 49, 418-435.

Moulin-Frier, C., Schwartz, J.L., Diard, J., & Bessière, P. (2011). Emergence of articulatory-acoustic systems from deictic interaction games in a "Vocalize to Localize" frame work. In *Primate Communication and Human Language - Vocalisation, gestures, imitation and deixis in humans and non-humans,* A. Vilain, J.L. Schwartz, C. Abry J. Vauclair (Eds.), John Benjamins Publishing Company, 191–218.

Moulin-Frier, C., Schwartz, J.L., Diard, J., & Bessière, P. (2010). *A unified theoretical Bayesian model of speech communication.* 1st conference on Applied Digital Human Modeling, Miami, USA.

Myung, J.I., & Pitt, M.A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116(3), 499–518.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Res Cogn Brain Res*, 3(2), 131–141.

Sato, M., Grabski, K., Glenberg, A., Brisebois, A., Basirat, A., Ménard, L. & Cattaneo, L. (in press). Articulatory bias in speech categorization: evidence from use-induced motor plasticity. *Cortex*.

Schwartz, J.L., Basirat, A., Ménard, L., & Sato, M. (2010). The perception-for-action-control theory (pact): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics* (in press).

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

Schwartz, J.L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield... A taxonomy of models for audiovisual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (eds.) *Hearing by eye, II. Perspectives and directions in research on audio-visual aspects of language processing* (pp. 85-108). Hove (UK) : Psychology Press.

Scott, S. K., Mcgettigan, C., & Eisner, F. (2009). A little more conversation, a little less action: candidate roles for motor cortex in speech perception. *Nature Reviews Neuroscience*, 10(4), 295–302.

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

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

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

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Science*s, 28, 675–691.

Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 603–623.

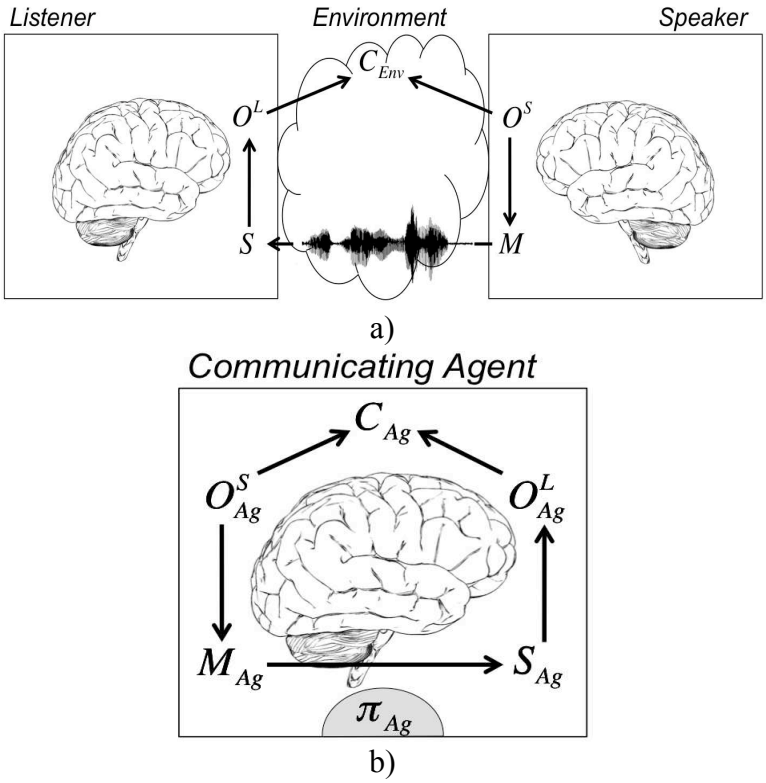


FIGURE 1. Conceptual model of a communication interaction. **(a)** A speaker and a listener are in presence of an object and, using acoustic signals, designate it. **(b)** The whole communication loop is internalized in the "brain" of each agent.

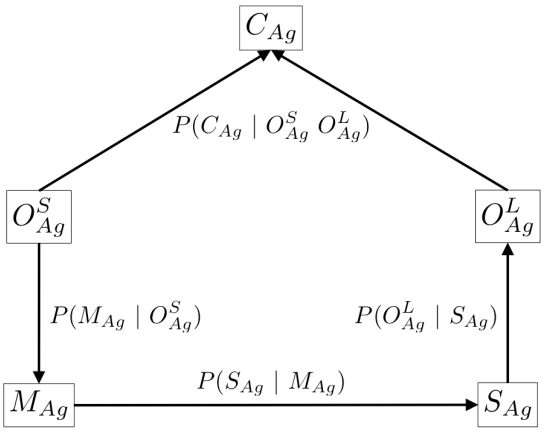


FIGURE 2. Structure of the communicating agent model, represented with a Bayesian network. Nodes correspond to variables, and arrows represent probability distributions, which illustrate the dependency structure.

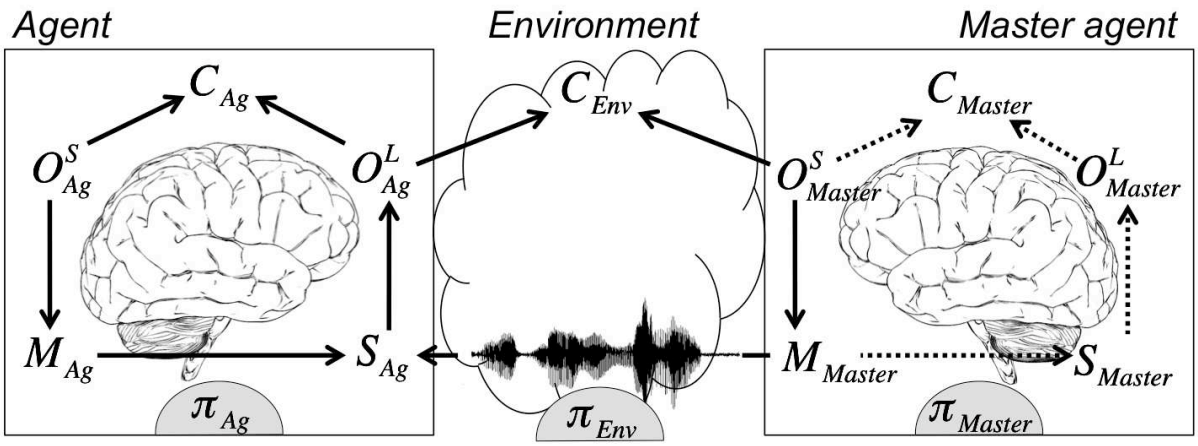


FIGURE 3. Experimental situation. Several agents populate the environment. On the left, the agent *Ag* is the focus of interest in our experiments. The other, *Master*, interacts with it and provides acoustic signals in order to designate an object. Both are instances of the Bayesian communicating agent model. A simulated environment mediates their interaction.

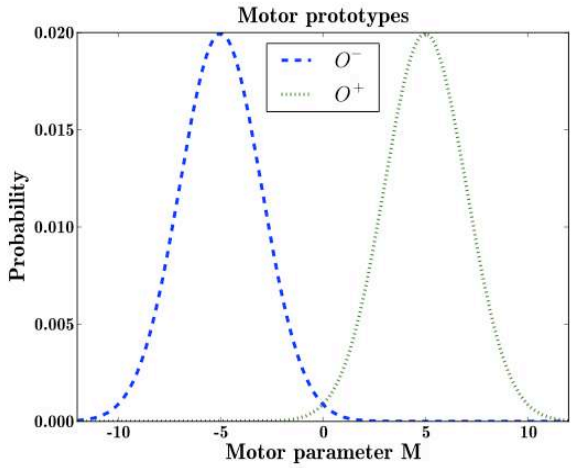
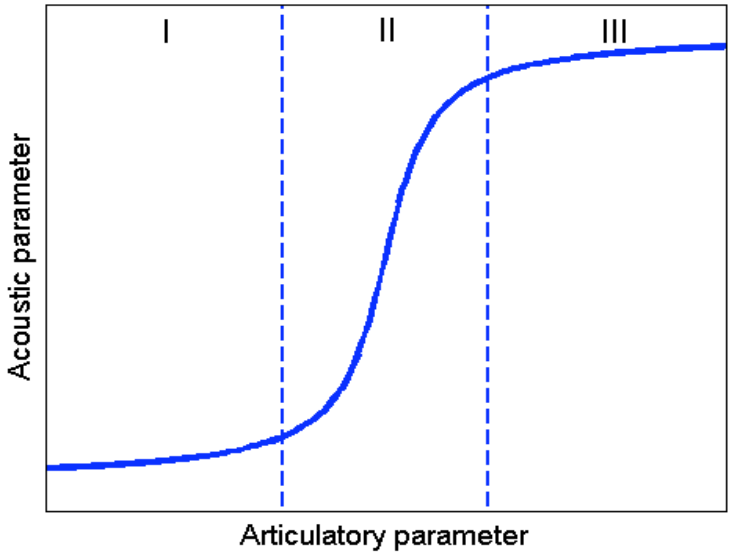
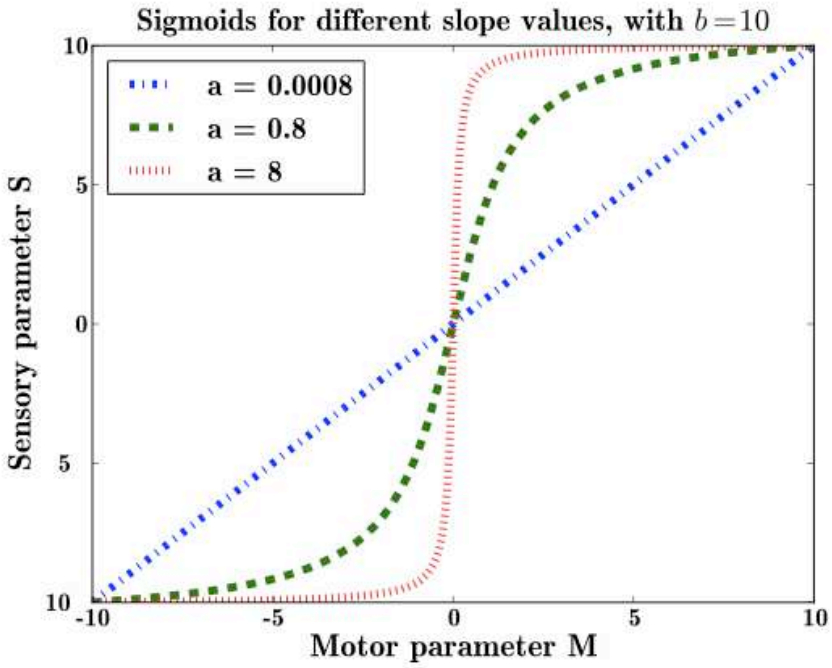


FIGURE 4. Motor prototype models. The left curve is $P(M \mid [O_S=o-])$, the right curve is $P(M \mid [O_S=o+])$.



a)



b)

FIGURE 5. Nonlinearities in articulatory-to-acoustic transforms. **(a)** The articulatory-to-acoustic transform according to the Quantal theory. This transform would include stability regions where variations of articulatory parameters have little or no influence (zones I and III) vs. instability regions around the nonlinearities, where small variations of articulatory parameters have a large influence (zone II). **(b)** Sigmoid functions used for computing the mean acoustic *S* signal (y-axis) resulting from a motor gesture *M* (x-axis). For each curve, the inflexion point is at *M=0*, and the slope at this point varies. Note that the chosen functional form allows studying both quasi-linear (for small *a* values) and strongly nonlinear (for large *a* values) articulatory-to-acoustic relationships.

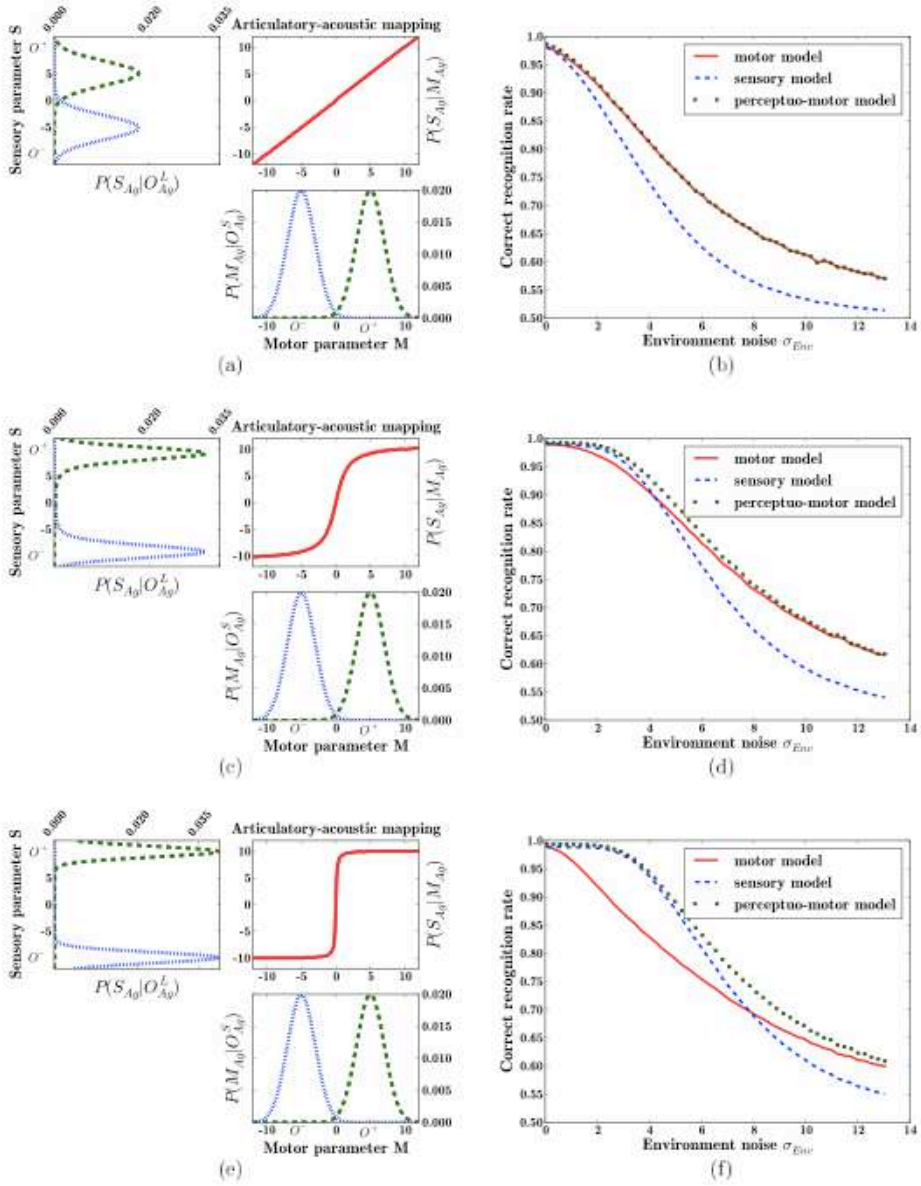


FIGURE 6. Experimental results for the 1D simulation of communication noise. See text for details. The plots of the left column show the motor prototypes *P(M | O_S)* in the bottom right corner, the articulatory-to-acoustic link *P(S | M)* (in terms of mean of the sensory variable for a given motor variable) in the top right corner, and the auditory prototypes *P(S | O_L)* in the top left corner. The plots on the right column show the corresponding variations of the correct recognition rate for the different models, when the environment noise σ_{Env} varies. Top row: linear case; middle row: nonlinear case; bottom row: strongly nonlinear case.

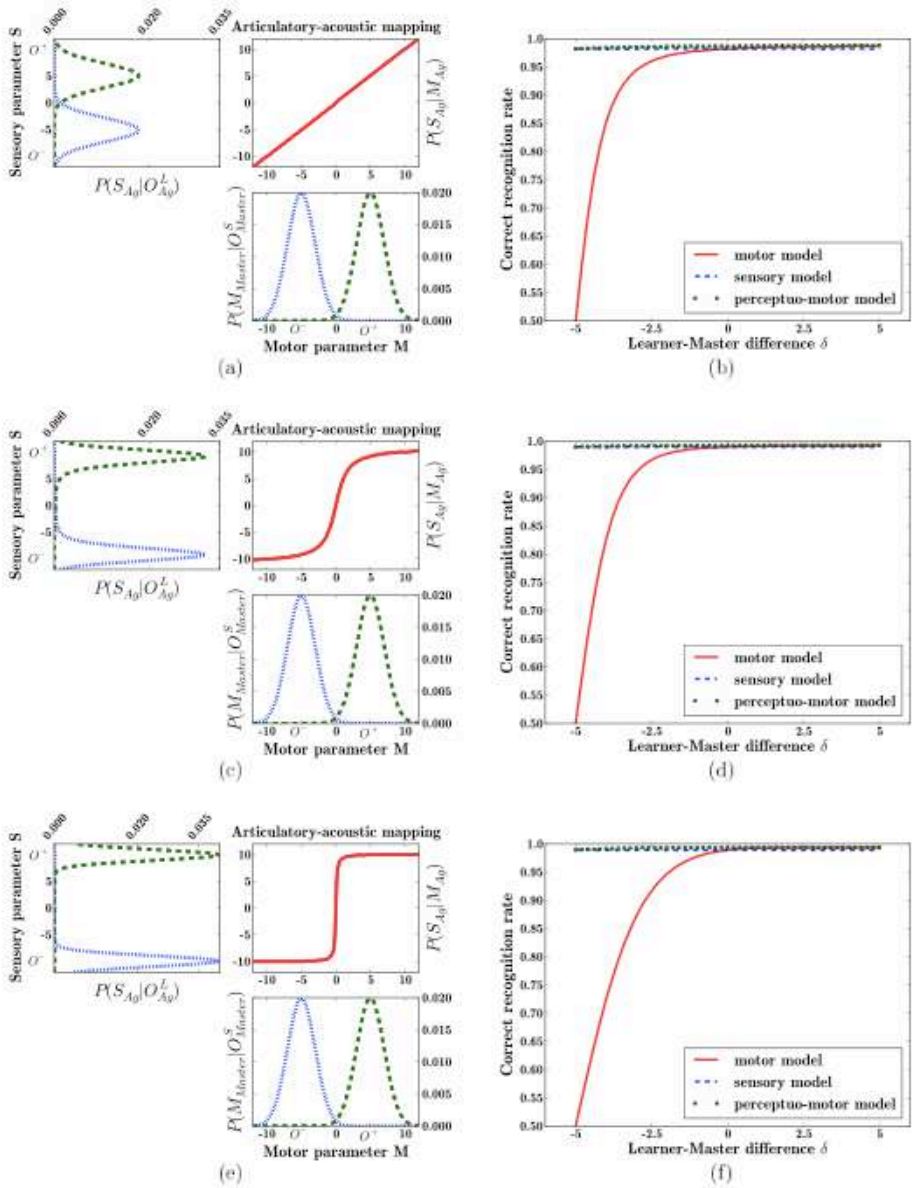
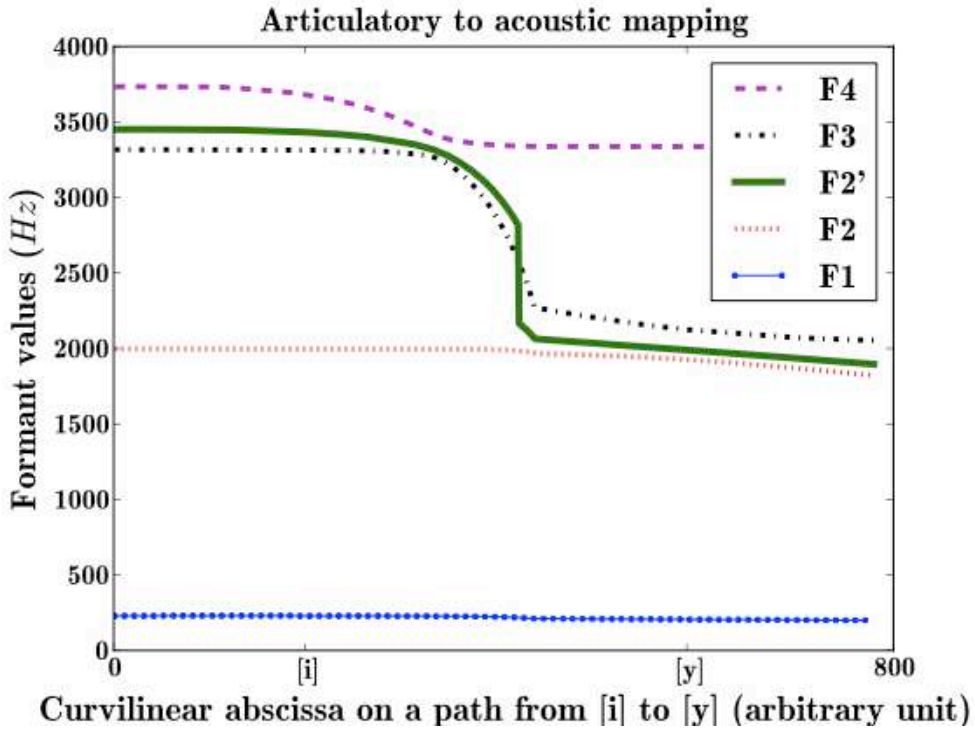
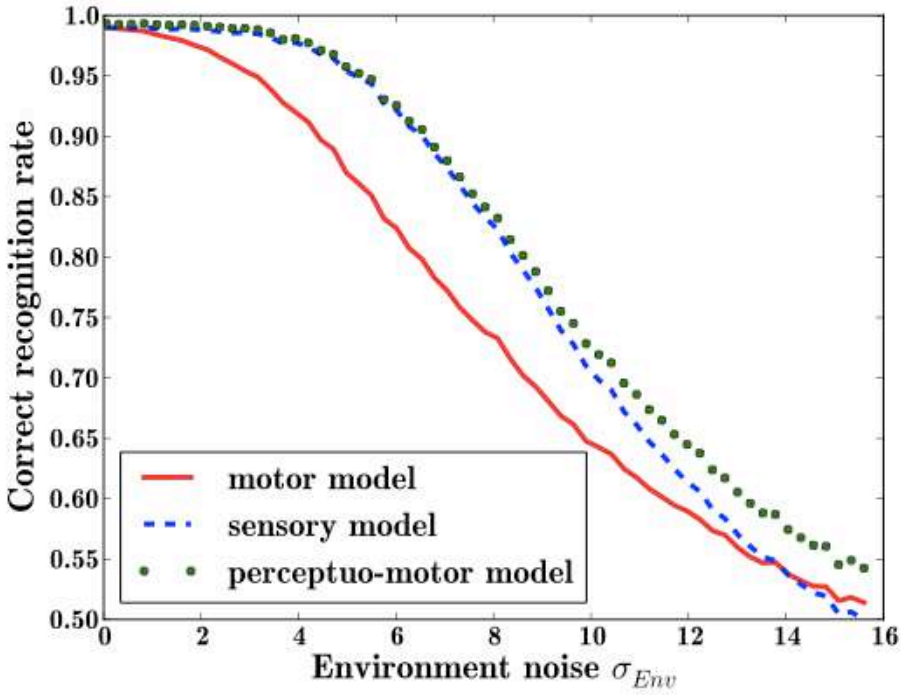


FIGURE 7. Experimental results for the 1D simulation of variation in motor prototypes between the learner agent *Ag* and the master agent *Master*. Same presentation as in Fig. 6. In the right column, the x-axis corresponds to the departure δ between the *Master* and *Ag* agent motor prototypes (0 means equal prototypes in the *Master* and *Ag* models).

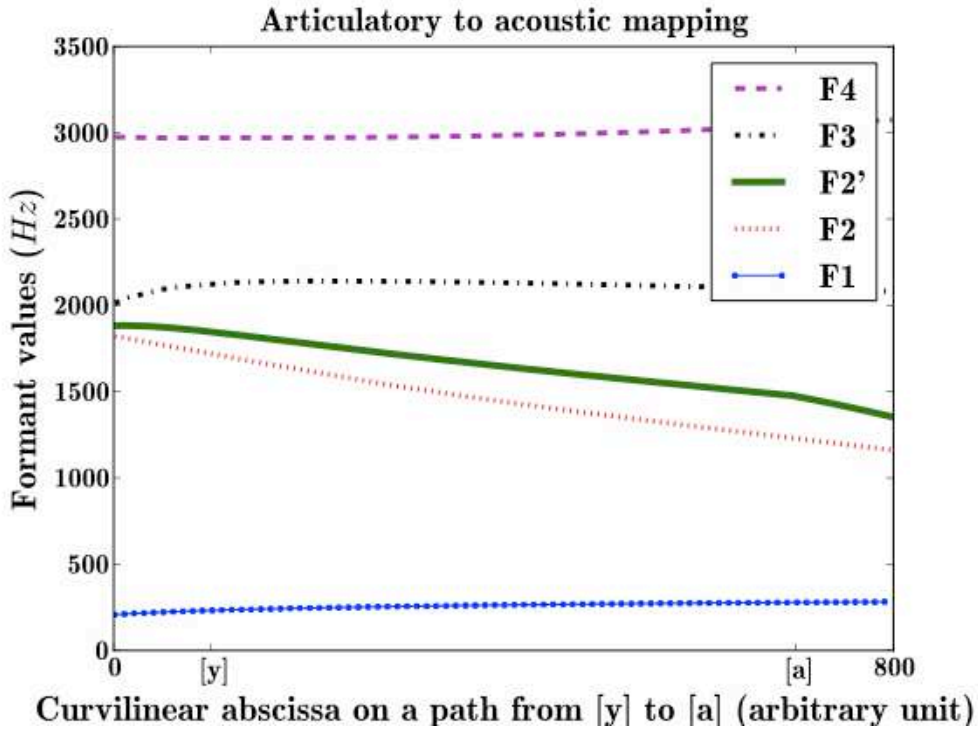


a)

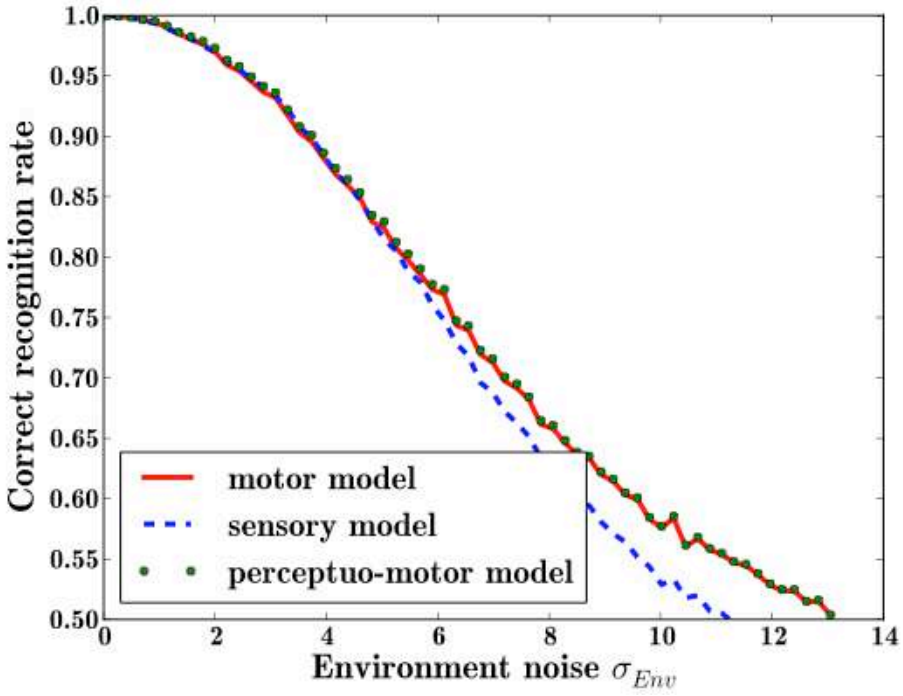


b)

FIGURE 8. Experimental results for the first 2D simulation: [i]-[y] perception with a highly non-linear articulatory-to-acoustic transform. **(a)** Variations of formants (F1, F2, F3, F4) and effective second perceptual formant F'2 when lip rounding varies from [i] (on the left) to [y] (on the right). **(b)** Correct recognition scores for the three models for varying communication noise.



a)



b)

FIGURE 9. Experimental results for the second 2D simulation: [y]-[a] perception with an almost linear articulatory-to-acoustic transform. **(a)** Variations of formants (F1, F2, F3, F4) and effective second perceptual formant F'2 when mouth opening varies from [y] (on the left) to [a] (on the right). **(b)** Correct recognition scores for the three models for varying communication noise.

Table 1. Taxonomy of speech production and perception theories and models

Task / Theory	**Production**	**Perception**
Motor	Articulatory Phonology (Browman and Goldstein, 1989)	Motor Theory (Liberman and Mattingly, 1985)
Auditory	Auditory reference frames for speech planning (Guenther, 1998)	Auditory theories (Diehl et al., 2004)
Perceptuo-motor	DIVA model (Guenther et al., 2006)	Perception for Action Control Theory (Schwartz et al., 2010)

Table 2. Model Taxonomy

Task / Theory	**Production** Inference of the form *P(M | O)*	**Perception** Inference of the form *P(O | S)*
Motor Object of interest is O^S	$P(M_{Ag} \mid O^S_{Ag})$	$P(O^S_{Ag} \mid S_{Ag}) = \sum_{M_{Ag}} P(M_{Ag} \mid O^S_{Ag}) P(S_{Ag} \mid M_{Ag})$
Auditory Object of interest is O^L	$P(M_{Ag} \mid O^L_{Ag}) \propto P(M_{Ag}) \sum_{S_{Ag}} P(S_{Ag} \mid M_{Ag}) P(O^L_{Ag} \mid S_{Ag})$	$P(O^L_{Ag} \mid S_{Ag})$
Perceptuo-motor Both O^S and O^L must be equal ($C=1$)	$P(M_{Ag} \mid O^L_{Ag}\ [C_{Ag}=1]) \propto P(M_{Ag} \mid O^S_{Ag}) \sum_{S_{Ag}} P(S_{Ag} \mid M_{Ag}) P(O^L_{Ag} \mid S_{Ag})$	$P(O^L_{Ag} \mid S_{Ag}\ [C_{Ag}=1]) \propto P(O^L_{Ag} \mid S_{Ag}) \sum_{M_{Ag}} P(M_{Ag} \mid O^S_{Ag}) P(S_{Ag} \mid M_{Ag})$