



HAL
open science

The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception

Raphaël Laurent, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière,
Julien Diard

► To cite this version:

Raphaël Laurent, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière, Julien Diard. The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. *Psychological Review*, 2017, 124 (5), pp.572-602. 10.1037/rev0000069 . hal-01484383v2

HAL Id: hal-01484383

<https://hal.science/hal-01484383v2>

Submitted on 17 Mar 2017

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.



AMERICAN PSYCHOLOGICAL ASSOCIATION

1

2 <http://www.apa.org/pubs/journals/rev/>

3 "This article may not exactly replicate the final version published in the APA journal. It is not

4 the copy of record."

5

6 |
7 The complementary roles of auditory and motor information evaluated in a Bayesian
8 perceptuo-motor model of speech perception

9 Raphaël Laurent, Marie-Lou Barnaud, Jean-Luc Schwartz

10 GIPSA-Lab

11 Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France

12 CNRS, GIPSA-Lab, F-38000 Grenoble, France

13 Pierre Bessière

14 CNRS, Sorbonne Universités - Université Pierre et Marie Curie

15 Institut des Systèmes Intelligents et de Robotique, Paris, France

16 Julien Diard

17 Laboratoire de Psychologie et NeuroCognition

18 Univ. Grenoble Alpes, LPNC, F-38000 Grenoble, France

19 CNRS, LPNC, F-38000 Grenoble, France

20
21 Authors' Note

22 The research leading to these results received funding from the European Research
23 Council under the European Community's Seventh Framework Programme (FP7/2007-2013
24 Grant Agreement no. 339152, "Speech Unit(e)s", J.-L. Schwartz PI). This work includes
25 research conducted as part of Raphaël Laurent's PhD thesis, defended at the Université de
26 Grenoble on October, 8th, 2014 (manuscript in French: "COSMO : un modèle bayésien des
27 interactions sensori-motrices dans la perception de la parole"). Portions of this work were
28 also presented at the 9th International Seminar on Speech Production (ISSP 11) in
29 Montréal, Canada (June 20-23, 2011), at the 14th Annual Conference of the International
30 Speech Communication Association (Interspeech 2013) in Lyon, France (August 25-29,
31 2013) and at the Workshop on Infant Language Development (WILD 15) in Stockholm,

32 Sweden (June 10-12, 2015). The authors wish to thank Clément Moulin-Frier and Louis-
33 Jean Boë for support and inspiration. Correspondence concerning this paper should be
34 addressed to J.-L. Schwartz, Université Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble,
35 France; CNRS, GIPSA-Lab, F-38000 Grenoble, France; 11 Rue des Mathématiques, 38400
36 Saint-Martin-d'Hères, France. Email: Jean-Luc.Schwartz@gipsa-lab.grenoble-inp.fr, Phone:
37 (+33)4 76 57 47 12, Fax: (+33)4 76 57 47 10.

38 Abstract

39 There is a consensus concerning the view that both auditory and motor representations
40 intervene in the perceptual processing of speech units. However, the question of the functional
41 role of each of these systems remains seldom addressed and poorly understood. We
42 capitalized on the formal framework of Bayesian Programming to develop COSMO
43 (Communicating Objects using Sensory-Motor Operations), an integrative model that allows
44 principled comparisons of purely motor or purely auditory implementations of a speech
45 perception task and tests the gain of efficiency provided by their Bayesian fusion.
46 Here, we show three main results. (i) In a set of precisely defined “perfect conditions”,
47 auditory and motor theories of speech perception are indistinguishable. (ii) When a learning
48 process that mimics speech development is introduced into COSMO, it departs from these
49 perfect conditions. Then auditory recognition becomes more efficient than motor recognition
50 in dealing with learned stimuli, while motor recognition is more efficient in adverse
51 conditions. We interpret this result as a general “auditory-narrowband vs. motor-wideband”
52 property. (iii) Simulations of plosive-vowel syllable recognition reveal possible cues from
53 motor recognition for the invariant specification of the place of plosive articulation in context,
54 that are lacking in the auditory pathway. This provides COSMO with a second property,
55 where auditory cues would be more efficient for vowel decoding and motor cues for plosive
56 articulation decoding. These simulations provide several predictions, which are in good
57 agreement with experimental data and suggest that there is natural complementarity between
58 auditory and motor processing within a perceptuo-motor theory of speech perception.

59 *Keywords:* Speech perception, computational modeling, sensory-motor interactions,
60 adverse conditions, plosive invariance

86 Skipper, Devlin & Lametti, 2017). Evidence emerged in two steps. Firstly, neuroanatomical
87 studies repeatedly showed that parietal and frontal brain areas associated with speech
88 production were consistently stimulated upon speech perception tasks (e.g. Fadiga, Craighero,
89 Buccino, & Rizzolatti, 2002; Pulvermüller et al., 2006; Watkins, Strafella, & Paus, 2003;
90 Wilson, Saygin, Sereno, & Iacoboni, 2004). This was particularly shown in non-standard
91 conditions involving noise (Binder, Liebenthal, Possing, Medler, & Ward, 2004; Zekveld,
92 Heslenfeld, Festen, & Schoonhoven, 2006), non-native stimuli (Callan, Callan, & Jones,
93 2014; Callan, Jones, Callan, & Akahane-Yamada, 2004; Wilson & Iacoboni, 2006), or
94 conflicting audiovisual inputs (Jones & Callan, 2003; Ojanen et al., 2005; Skipper, van
95 Wassenhove, Nusbaum, & Small, 2007). Secondly, behavioral studies looked for a causal role
96 of motor areas in speech perception by altering or modulating the potential efficiency of
97 speech motor centers, by Transcranial Magnetic Stimulation (TMS), repeated TMS or motor
98 perturbations. Such studies have shown small but consistent perceptual effects in
99 categorization or discrimination of speech stimuli, in ambiguous or noisy conditions (e.g.,
100 d'Ausilio et al., 2009; d'Ausilio, Bufalari, Salmas, & Fadiga, 2012; Grabski, Tremblay,
101 Gracco, Girin, & Sato, 2013; Ito, Tiede, & Ostry, 2009; Meister, Wilson, Deblieck, Wu, &
102 Iacoboni, 2007; Möttönen, Dutton, & Watkins, 2013; Möttönen & Watkins, 2009; Rogers,
103 Möttönen, Boyles, & Watkins, 2014; Sato, Tremblay, & Gracco, 2009; Sato et al., 2011;
104 Shiller, Sato, Gracco, & Baum, 2009).

105 In this context, the strong “auditory” vs. “motor” controversy about invariance at the
106 crossroads of phonetics and phonology that prevailed until the end of the 1980s was almost
107 completely replaced since the beginning of the 1990s by an integrative view from cognitive
108 neuroscience, assuming that the motor and auditory systems collaborate in speech perception.
109 This has the merits of taking into account new experimental insights, but its drawback is that
110 the question of the respective functions of sensory and motor systems has almost completely

111 disappeared from the literature. However, if both auditory and motor processes do intervene
112 in speech perception ⁽¹⁾, the potential specificity and complementarity of these two systems
113 within a perceptuo-motor speech perception architecture becomes essential. How could it be
114 useful for speech perception to capitalize on two different systems? How could the motor
115 system be more helpful in adverse conditions? What specific aspects of computation, for what
116 kind of information extraction, are respectively implemented by the motor and auditory (if not
117 visual or somatosensory) components of the speech perception system?

118 These are the questions we address in the theoretical framework of the “Perception-
119 for-Action-Control Theory” (PACT). PACT is a perceptuo-motor theory of speech perception,
120 connecting perceptual shaping and motor procedural knowledge in a principled way, in
121 speech multisensory processing within the human brain (Schwartz, Basirat, Ménard, & Sato,
122 2012a; Schwartz, Boë, & Abry, 2007). PACT considers that perceptual knowledge is involved
123 in both speech comprehension and speech control, in a communicative process. The
124 communication unit through which parity may be achieved, is neither a sound, nor a gesture,
125 but a perceptually-shaped gesture, that is a perceptuo-motor unit characterized both by its
126 articulatory coherence, provided by its gestural nature and its perceptual value, necessary for
127 function. Motor processes could be associated with multisensory processes through audio-
128 visuo-motor binding, enabling a better extraction of adequate cues for further categorization
129 processes (Basirat, Schwartz, & Sato, 2012; see also Skipper, van Wassenhove, Nusbaum &
130 Small, 2007). Furthermore, perceptual categorization would benefit from motor information
131 in addition to auditory and possibly visual clues. This would, improve variability processing
132 and the extraction of invariance (Schwartz, Abry, Boë, & Cathiard, 2002; Schwartz et al.,
133 2007, 2012a).

134 In PACT, it is also acknowledged that perception and action are co-structured in the
135 course of speech development, which involves both producing and perceiving speech items.

136 The schedule of perceptuo-motor development in the first few years of age is important in this
137 context, and seems to incorporate several major steps (Kuhl, 2004; Kuhl et al., 2008). First,
138 auditory processes mature, enabling categorization of many phonetic contrasts almost from
139 birth (e.g., Bertoncini, Bijeljac-Babic, Blumstein, & Mehler, 1987; Eimas, Siqueland,
140 Jusczyk, & Vigorito, 1971; Jusczyk & Derrah, 1987), with an early focus on the sounds of the
141 infant's language. This can be as early as 6 months old for vowels (Kuhl, Williams, Lacerda,
142 Stevens, & Lindblom, 1992) and 10 months old for consonants (Werker & Tees, 1984). Motor
143 processes evolve later and more slowly, beginning by articulatory exploration of the possible
144 vocal repertoire, with canonical babbling at around 7 months of age (Davis, MacNeilage, &
145 Matyear, 2002; MacNeilage, 1998). This continues with a later focus on the sounds of the
146 phonological system from the end of the first year and through the following ones.
147 Importantly, canonical babbling, sometimes considered as a purely endogenous process
148 enabling infants to extensively explore the possibilities of their vocal tracts, seems to be
149 influenced since its very beginning by the language heard in the surrounding environment.
150 Such "babbling drift" has been displayed in a number of experiments concerning vowel
151 formants, consonant-vowel associations and prosodic schemes (e. g. de Boysson-Bardies,
152 1993; de Boysson-Bardies, Hallé, Sagart, & Durant, 1989; de Boysson-Bardies, Sagart, &
153 Durant, 1984).

154 Auditory perception is hence mature and focused before orofacial control occurs.
155 Furthermore, the connection between the speech perception system and the motor system
156 through the parieto-frontal dorsal pathway in the cortex does not seem to be completely
157 mature at birth, but rather evolves throughout the first year of life (Dehaene-Lambertz,
158 Dehaene, & Hertz-Pannier, 2002; Dehaene-Lambertz et al., 2006; vs. Kuhl, Ramírez,
159 Bosseler, Lotus Lin, & Imada, 2014; Imada et al., 2006). Consequently, motor information
160 would not be mature, focused or fully available for perception until the end of the first year.

161

162 **Computational models of auditory vs. motor theories of speech perception**

163 In the context of debates concerning the potential role of auditory and motor systems
 164 in speech perception, computational models are likely to shed light on them by enabling
 165 quantitative evaluation of some of the theoretical arguments in relation to experimental data.
 166 There are already many computational models of auditory theories of speech perception.
 167 Many, if not all of acoustic speech recognition systems can be construed as such, as they
 168 involve the best statistical analyses of the acoustic content of large speech corpora for speech
 169 understanding (see recent reviews in e.g. Hinton et al., 2012; Huang & Deng, 2010). They
 170 also often incorporate more or less sophisticated computational models of the auditory
 171 analysis of acoustic stimuli in the human brain (e.g. Hermansky, 1998; Deng, 1999). Auditory
 172 theory models also include computational psycholinguistic models of cognitive speech
 173 processing (e.g. *Trace*: McClelland & Elman, 1986; the *Distributed Cohort Model*: Gaskell &
 174 Marslen-Wilson, 1997; *Parsyn*: Luce, Goldinger, Auer & Vitevitch, 2000; see also
 175 Scharenborg, Norris, Ten Bosch & McQueen, 2005).

176 A widespread mathematical framework, in this domain, is probabilistic modeling,
 177 where a generative, predictive model associates probable acoustic signals with linguistic
 178 categories. Then, perception is cast as a categorization process, in which Bayes theorem is
 179 used to infer the most likely linguistic category given some acoustic stimulus:

$$180 \quad P([O = o_i]|S) = \frac{P(S|[O=o_i])P([O=o_i])}{\sum_j P(S|[O=o_j])P([O=o_j])},$$

181 where $P(S|[O = o_i])$ expresses the probability distribution of acoustic cues for a given
 182 category and $P([O = o_i])$ defines prior probabilities of each category.

183 The origin of such models can be traced back, historically, to Signal Detection Theory
 184 (Tanner & Swets, 1954; Green & Swets, 1966; more recent references include Dayan &
 185 Abbott, 2001; Rouder & Lu, 2005) and its multi-dimensional generalization, the General

186 Recognition Theory (Ashby & Townsend, 1986; Ashby & Perrin, 1988). Recent years saw the
187 resurgence and spread of Bayesian models of speech perception, that consider the
188 categorization process above to model the optimal, ideal acoustic (or possibly visual, see
189 footnote 1) information processing system, without any reference to motor processes. Such
190 “optimal” models include ideal listener models (Feldman, Griffiths & Morgan, 2009;
191 Sonderegger & Yu, 2010) and ideal adapter models (Clayards, Aslin, Tanenhaus & Jacobs,
192 2007; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Kleinschmidt & Jaeger, 2011, 2015),
193 where categories are either learned in a batch manner (De Boer & Kuhl, 2003; Dillon, Dunbar
194 & Idsardi, 2013) or acquired and adapted incrementally (McMurray, Aslin & Toscano, 2009;
195 Vallabha, McClelland, Pons, Werker & Amano, 2007). In this trend, many extensions were
196 proposed, for instance for dealing with multiple cues (Toscano & McMurray, 2008, 2010) or
197 with higher-level structures above phonemic representations (e.g., words, syllables), in
198 hierarchical models (Norris & McQueen, 2008; Feldman, Griffiths & Morgan, 2009b; Kiebel,
199 Daunizeau, & Friston, 2009; Feldman, Griffiths, Goldwater & Morgan, 2013).

200 In comparison, there are many less computational motor theory models, basically
201 because of the lack of easily available articulatory or motor data required for the training of
202 such models. A few automatic speech recognition systems attempt to introduce articulatory
203 data into their statistical processes (e.g. Deng & Ma, 2000; Deng, Ramsay, & Sun, 1997;
204 Frankel, Richmond, King & Taylor, 2000; Sun & Deng, 2002) and a recent series of machine
205 learning models based on artificial neural networks and applied to articulatory data recorded
206 through electromagnetic articulography aimed to show the efficiency of articulatory inputs for
207 phonetic decoding (e.g. Canevari, Badino, d'Ausilio, Fadiga, & Metta, 2013; Castellini et al.,
208 2011). A few variants of Bayesian models of speech perception, because they consider
209 computations of the speaker’s intentions through motor inversion, can be construed as
210 involving motor knowledge during perception, although it was not their initial purpose, as

211 they were developed instead to model perceptual magnet effects (Feldman & Griffiths, 2007;
212 Feldman, Griffiths & Morgan, 2009). Other authors attempted to develop formal models
213 without real articulatory ground truth data to evaluate the possibility of implementing motor
214 or perceptuo-motor theories of speech perception (e.g. Kröger, Kannampuzha, & Kaufmann,
215 2014; Kröger, Kannampuzha, & Neuschaefer-Rube, 2009; Moore, 2007).

216 However, while all these developments basically aim to demonstrate that articulatory
217 or motor speech decoding is indeed feasible and potentially efficient, none of this research
218 attempts to really evaluate why and how motor information could be relevant for speech
219 decoding. Furthermore, it is always difficult in these models to precisely disentangle what
220 comes from the distribution of articulatory information and what comes from specific choices
221 in the computational implementation.

222 This is why the Bayesian implementation of an instance of motor theory
223 (implementing motor decoding in a Bayesian framework) or perceptuo-motor theory
224 (including the fusion of auditory and motor information) could enable the functional role of
225 the motor system to be assessed more clearly and rigorously. This is the objective of the
226 COSMO model (Communicating Objects using Sensory-Motor Operations) that will be
227 presented in the next section.

228

229 **COSMO, a Bayesian computational framework for assessing the functional role of** 230 **auditory vs. motor systems**

231 To attempt to better understand the function of motor information in speech
232 perception, we have developed over recent years a computational Bayesian framework called
233 COSMO. This model enables auditory, motor and perceptuo-motor theories of speech
234 communication to be implemented and compared in a coherent set of simple probabilistic
235 equations and processes, based on Bayesian modeling. COSMO was initially developed to

236 deal with the emergence of sound systems in human languages (Moulin-Frier, Diard,
237 Schwartz, & Bessière, 2015; Moulin-Frier, Schwartz, Diard, & Bessière, 2011) and was then
238 adapted to the study of speech perception in adverse conditions (Moulin-Frier, Laurent,
239 Bessière, Schwartz, & Diard, 2012). The present article greatly expands the initial study of
240 Moulin-Frier et al. (2012), which attempted to clearly assess when and how motor
241 information could be useful for phonetic decoding.

242 A first part will present the COSMO model, together with an initial crucial result we
243 obtained with COSMO, which we called the indistinguishability theorem. This theorem shows
244 that in a set of precisely defined “perfect conditions”, auditory and motor theories of speech
245 perception are indistinguishable (Moulin-Frier et al., 2012). We will present this theorem in
246 detail, since it is of great theoretical importance, providing a landmark for any further
247 comparison of auditory and motor models of speech perception.

248 Indeed, distinguishing the functional roles of auditory and motor systems for speech
249 perception can only be achieved by departing from these perfect conditions. This can occur in
250 one of two major ways, providing the two major contributions of the present paper to the
251 subject.

252 Firstly, the learning process can differentiate the auditory and motor systems. We
253 claim that these two systems evolve differently during learning. The auditory system could
254 focus rapidly and precisely on the set of learning stimuli provided by the environment, leading
255 to a system finely tuned to this learning set. This would provide the auditory system with a
256 “narrow-band” specificity with respect to the learning data. In contrast, the motor system
257 would “wander” more through the sensory-motor space during its exploration stage, because
258 of the complexity of the task at hand. Hence it would evolve more slowly and focus less
259 efficiently on the set of learning stimuli provided by the environment, in agreement with the
260 developmental timeline described previously. However, it would be able to process a wider

261 set of stimuli thanks to the “wandering” phenomenon. This would provide the motor system
262 with a “wide-band” specificity, making it poorer for learned stimuli, but better at generalizing
263 about adverse conditions involving unlearned stimuli. This will be developed in Part 2,
264 together with two predictions associated with this “auditory-narrow, motor-wide” property,
265 that will be compared to the available experimental data.

266 Secondly, the two systems can be differentiated in terms of the nature and complexity
267 of their internal representations, possibly leading to different processing of variability of the
268 phonological units. Considering simulations of the recognition of plosive-vowel sequences,
269 we explore the assumption that motor recognition might provide clues as to the invariant
270 specification of the place of articulation of plosives in context, which is lacking in the
271 auditory pathway, while the auditory categorization of vowels would be more straightforward
272 than its motor counterpart. Altogether, this suggests that there should be a natural
273 complementarity between auditory and motor systems within a perceptuo-motor theory of
274 speech communication. This will be developed in Part 3, together with two other predictions
275 that will be discussed in light of available experimental data.

276 Following the important simulation contributions and predictions, we will end this
277 paper with a review of some major perspectives and challenges associated with the
278 development of COSMO, in relation to cognitive processes involved in speech
279 communication.

280

281 **Part 1 – COSMO and the indistinguishability theorem**

282

283 In this first part we will introduce the two major pieces of our computational framework.

284 Firstly, we will present and describe COSMO together with its mathematical specification and

285 the way it enables modeling of auditory, motor or perceptuo-motor theories of speech

286 perception. Secondly, we will derive the indistinguishability theorem, already published by

287 Moulin-Frier et al. (2012), but which will be explained more precisely in the present paper.
288 This will provide us with the crucial landmark that will serve for further simulations presented
289 in parts 2 and 3.

290

291 **The COSMO model**

292 COSMO stems from the analysis of spoken communication, which can be broken
293 down into a minimal set of variables, with a high level of abstraction.

294 ***FIGURE 1 ABOUT HERE***

295 As is shown in the upper part of Figure 1, to communicate about an object O_S , the
296 *Speaker* Agent performs a motor gesture M resulting in a sensory input S from which the
297 *Listener* Agent retrieves the object O_L ⁽²⁾. The variable C^{Env} is a Boolean variable assessing
298 the communication success: it is *True* when $O_S = O_L$.

299 Our work is based on the hypothesis that the Communicating Agent, is able to act both
300 as a speaker and a listener, and has internal representations of the whole communication loop,
301 as shown in the lower part of Figure 1. Consequently, the Communicating Agent model is
302 made up of (i) a *motor system* associating motor representations M to the object O_S to be
303 produced and of (ii) a *sensory system* associating the perceived object O_L to the sensory
304 representation S , which are linked by (iii) a *sensory-motor system* that allows the
305 consequences of motor commands M in terms of sensory inputs S to be predicted. At this
306 stage, M , S and O are still generic variables, in order not to lose generality. They will be
307 instantiated for experiments and made more precise later in this paper.

308 The model contains two different variables O_S and O_L , one for the intention of the
309 speaker, the other for the perception of the listener. They are also useful to avoid a directed
310 loop from a single variable O to itself through motor and perceptual variables M and S .
311 Indeed, such loops are not compatible with straightforward application of Bayes theorem;

312 duplication of the variables is a classic solution to circumvent this technical problem (e.g.
 313 state variables are replicated over time for temporal series models). Such duplication between
 314 phonological codes for production and perception, linked by conversion mechanisms, is
 315 compatible with neuropsychological data (Jacquemot, Dupoux, & Bachoud-Lévi, 2007).
 316 Coherence between variables O_S and O_L is imposed by (iv) a Boolean variable C when it is set
 317 to *True*. The variable C can also be conceived as the internalization of the C^{Env} variable
 318 assessing communication success.

319 This Communicating Agent model has a name, COSMO, that also happens to recall
 320 the model variables. COSMO is formally defined within the framework of Bayesian
 321 Programming (Bessière, Laugier, & Siegwart, 2008; Bessière, Mazer, Ahuactzin-Larios, &
 322 Mekhnacha, 2013; Lebeltel, Bessière, Diard, & Mazer, 2004) by probability distributions,
 323 which encode the subjective knowledge that the agent has about the relations between its
 324 internal representations. The COSMO model is thus defined by one mathematical object, the
 325 joint probability distribution over all its variables: $P(C O_L S M O_S)$. It contains all the
 326 information the agent has about its internal variables and can be shown to be sufficient to
 327 perform any inference task about these variables, whatever the form. In other words, any
 328 conditional probability over some of these variables, knowing some others, can be computed
 329 from the joint probability distribution. We chose to decompose and simplify this joint
 330 probability distribution as follows (Moulin-Frier et al., 2012):

$$331 \quad P(C O_L S M O_S) = \underbrace{P(O_S)}_{\text{prior}} \underbrace{P(M | O_S)}_{\substack{\text{motor} \\ \text{repertoire}}} \underbrace{P(S | M)}_{\substack{\text{forward} \\ \text{model}}} \underbrace{P(O_L | S)}_{\substack{\text{sensory} \\ \text{classifier}}} \underbrace{P(C | O_S O_L)}_{\substack{\text{communication} \\ \text{success}}} . (1)$$

332 Various tasks can then be carried out by asking questions to the model, by computing
 333 conditional probability distributions of the form $P(\text{SEARCHED} | \text{OBSERVATIONS})$: What is
 334 the probability distribution over the *SEARCHED* variables, knowing the value of some
 335 *OBSERVATIONS*? In the COSMO framework, a speech production task amounts to

336 computing a conditional distribution of the form $P(M | O)$: What is the probability
337 distribution over motor commands M corresponding to the object O to be communicated? A
338 speech perception task amounts to computing a conditional distribution of the form $P(O | S)$:
339 What is the probability distribution over perceived objects O , given the sensory input S ?

340 Within the framework of *COSMO*, these questions can be instantiated in three
341 different ways: (i) by replacing O by O_S we implement a motor theory focused on the
342 speaker's perspective, (ii) by replacing O by O_L we implement an auditory theory focused on
343 the listener's perspective, (iii) by indifferently using either O_S or O_L and by further
344 conditioning the computed distribution with the constraint $C = True$, we implement a
345 perceptuo-motor theory that ensures the coherence of both representations. This is the
346 equivalent of explicitly setting the communication success as a goal of the task considered.
347 Bayesian inference provides a way to compute the conditional probability distributions
348 corresponding to all these tasks from the joint probability distribution that defines the
349 *COSMO* model (Equation (1)). Figure 2 shows the results of these computations and how they
350 can be interpreted. We now explain further the results of these Bayesian inferences, focusing
351 on the speech perception task.

352 ***FIGURE 2 ABOUT HERE***

353 As can be seen in Figure 2, the implementation of an auditory theory of perception
354 consists of a direct computation of $P(O | S) = P(O_L | S)$, with no intervention of motor
355 variables. This is consistent with classical proposals about auditory theories, which deny the
356 role of motor knowledge in speech perception and consider that it is based exclusively on the
357 set of auditory processing and categorization mechanisms available in the human brain (e.g.
358 Diehl et al., 2004). We note that this portion of our model is equivalent to many preceding
359 models of acoustic categorization, including the Ideal Adapter model of Kleinschmidt &
360 Jaeger (2015) mentioned previously.

361 Likewise, the implementation of a motor theory of perception, i.e., computing
 362 $P(O | S) = P(O_S | S) \propto \sum_M (P(M | O_S) P(S | M))$ (Feldman & Griffiths, 2007; Feldman,
 363 Griffiths & Morgan, 2009; Moulin-Frier et al., 2012), is consistent with the view that speech
 364 perception occurs by retrieving the intended motor gestures of the speaker (Liberman &
 365 Mattingly, 1985). Indeed, the motor variable M now plays a role in the inference. The
 366 deterministic two-stage process posited by motor theories begins with the retrieval of M from
 367 S through an inverse model, which is followed by the categorization process estimating O_S
 368 from M through a motor decoder. In the Bayesian framework, these are replaced by the
 369 computation of the sum over the possible values of the variable M , weighted by the
 370 probability that they have the sensory consequence S and by the probability that they are
 371 associated with O_S the considered object. This is a Bayesian analogue to analysis-by-synthesis
 372 (Halle & Stevens, 1959; Stevens & Halle, 1967; see a review in Bever & Poeppel, 2010). The
 373 deterministic two-stage process, firstly with motor-to-sensory inversion and secondly with
 374 motor decoding, is an approximation of the summation over all M values.

375 Finally, the implementation of a perceptuo-motor theory of perception consists simply
 376 of a mere Bayesian fusion of the predictions of the sensory and motor categorization
 377 processes: $P([O = o] | S [C = True]) \propto P([O_L = o] | S) \sum_M (P(M | [O_S = o]) P(S | M))$.

378

379 **Indistinguishability of auditory and motor theories in perfect conditions of learning and** 380 **communication**

381 Although purely sensory and purely motor perceptions are described by different
 382 equations (see Figure 2), it can be proven that if three hypotheses defining a set of “perfect
 383 conditions” of learning are verified, the motor and auditory theories of perception make
 384 exactly the same predictions. Therefore, these cannot be distinguished empirically. This
 385 demonstration has been presented previously (Moulin-Frier et al., 2012), but in a less explicit

386 formulation. We will present it here again in detail, in its more rigorous form, with three
 387 hypotheses instead of the two used previously.

388 ***FIGURE 3 ABOUT HERE***

389 We consider a supervised learning scenario, shown in Figure 3, which features
 390 Learning Agents and a Master Agent, each described as a COSMO agent. To distinguish their
 391 variables, superscripts are added and variables become O_S^{Ag} , O_S^{Master} , M^{Ag} , M^{Master} , etc.

392 In the learning scenario, the Learning Agent is provided by the Master Agent with the
 393 following <object, stimulus> pairs. The Master Agent uniformly selects O_S^{Master} objects,
 394 draws corresponding M^{Master} motor commands according to the production model
 395 $P(M^{Master} | O_S^{Master})$, which are then transformed by the environment modeled by
 396 $P(S^{Ag} | M^{Master})$ and result in sensory S^{Ag} inputs. Furthermore, the variable C^{Env} , which
 397 ensures coherence between the O_S^{Master} and O_L^{Ag} objects, implements a shared attention
 398 mechanism, e.g. deixis, which allows the Learning Agent to retrieve the right objects (O_L^{Ag})
 399 from the Master to associate with the S^{Ag} stimuli in its sensory classifiers $P(O_L^{Ag} | S^{Ag})$. The
 400 Learning Agent builds its sensory classifier through successive random draws, which are
 401 mathematically expressed by the following approximation:

$$402 \quad P(O_L^{Ag} | S^{Ag}) \approx \sum_M (P(M^{Master} | O_S^{Master}) \cdot P(S^{Ag} | M^{Master})). \quad (2)$$

403 In this equation, the sign \approx expresses the fact that the set of learning stimuli (right part
 404 of the equation) has to be learned in some way from the $P(O_L^{Ag} | S^{Ag})$ distribution (left part of
 405 the equation).

406 We now define the three hypotheses used in this approach and prove that their
 407 conjunction ensures the indistinguishability of the motor and auditory theories of speech
 408 perception.

409 i. H1 (the “perfect sensory learning hypothesis”): the sensory classifier is

410 perfectly learned from the Master's productions, i.e. $P(O_L^{Ag} | S^{Ag}) =$
 411 $\sum_M (P(M^{Master} | O_S^{Master}) \cdot P(S^{Ag} | M^{Master}))$. By replacing the operator \approx of
 412 Equation (2) by an equality operator $=$, H1 explicitly states that the sensory
 413 classifier $P(O_L^{Ag} | S^{Ag})$ learned by the agent perfectly encodes all the
 414 information expressed by the combination of the probability distributions
 415 $P(M^{Master} | O_S^{Master})$ and $P(S^{Ag} | M^{Master})$. These describe the way the
 416 Master performs its motor gestures and the way they are transformed by the
 417 environment.

- 418 ii. H2 (the “perfect motor learning hypothesis”): the motor repertoire of the agent
 419 is identical to that of the Master, i.e. $P(M^{Ag} | O_S^{Ag}) = P(M^{Master} | O_S^{Master})$.
 420 iii. H3 (the “perfect sensory-motor learning hypothesis”): the agent's sensory-
 421 motor system perfectly encodes the properties of the transformation performed
 422 by the environment during the learning process, i.e. $P(S^{Ag} | M^{Ag}) =$
 423 $P(S^{Ag} | M^{Master})$.

424 The indistinguishability theorem states that if H1, H2 and H3 hold, then the motor and
 425 sensory instantiations of the speech perception task are indistinguishable.

426 The proof is straightforward. Starting from Equation (2), which states how the sensory
 427 decoder is learned along the paradigm in Figure 3, hypothesis H1 enables the learning
 428 operator \approx to be replaced by an equality operator $=$, while hypotheses H2 and H3 enable the
 429 two terms on the right hand side of Equation (2) to be replaced by $P(M^{Ag} | O_S^{Ag})$ and
 430 $P(S^{Ag} | M^{Ag})$, respectively, which yields:

$$431 \quad P(O_L^{Ag} | S^{Ag}) = \sum_M (P(M^{Ag} | O_S^{Ag}) \cdot P(S^{Ag} | M^{Ag})). \quad (3)$$

432 The right hand side of Equation (3) has now become the expression of the motor
 433 instantiation of the speech perception task, while the left hand side is the expression of the
 434 perception task instantiated within the framework of the auditory theory (see Figure 2).

435 Therefore, if these three hypotheses are verified within a set of “perfect conditions” for
436 learning, the sensory and motor models rely on the same information and make the same
437 predictions. They are thus indistinguishable, whatever the test conditions might be.

438 When the indistinguishability theorem is satisfied, information encoded in the motor
439 and sensory pathways is redundant. This shows that even when two theories or models are
440 seemingly different – as the auditory and motor theories of speech perception appear to be –
441 they may be identical with respect to the computation they perform (as conceptualized by
442 Marr, 1982, in his three-level framework, in which the same computational task can be carried
443 out by algorithmic models with different representations; see also Laurent, Schwartz,
444 Bessière, & Diard, 2013).

445 Similar arguments are sometimes invoked in papers about auditory theories (e.g. Diehl
446 et al, 2004, p. 168: “listeners do not recover gestures, but they do perceive the acoustic
447 consequences of gestures. Any regularities of speech production (e.g., context dependencies)
448 will be reflected in the acoustic signal and, through general mechanisms of perceptual
449 learning, listeners come to make use of the acoustic correlates of these production regularities
450 in judging the phonemic content of speech signals”). The indistinguishability theorem
451 provides a theoretical basis based on Bayesian modeling to explain such more or less intuitive
452 claims. More importantly, it suggests that what should drive our understanding of the
453 respective roles of the auditory vs. motor systems in speech perception is related to what we
454 are able to learn about them in the course of speech development.

455 Understanding the potential role and complementarity of the sensory and motor
456 recognition processes requires departing from the perfect conditions defined previously.
457 Given the structure of the motor and sensory models, the possible differences between their
458 predictions of perception tasks are strongly dependent on the information they encode, i.e., on
459 how they were learned.

460 The next parts of this article will introduce two sets of simulations providing two
461 directions in which auditory and motor theories depart from each other. Furthermore, some
462 fundamental sources of functional complementarity will be displayed.

463

464

465 **Part 2 – The “auditory-narrow, motor-wide” framework for speech perception**

466 In this part, we will focus on a *generic property* of COSMO that we consider largely
467 independent from the specific implementation choices and that refers to structural aspects of
468 the way auditory vs. motor decoding can be modeled in a Bayesian framework. This generic
469 property generates a natural complementarity between auditory and motor decoding
470 processes, that we summarize by the so-called “auditory-narrow, motor-wide” framework.
471 Finally we discuss the relationship between simulations and experimental data for speech
472 perception development and speech processing in noise.

473

474 **The sensory branch is narrow-band, the motor branch is wide-band: simulations within** 475 **a simplified one-dimensional sensory-motor space**

476 Equations in Figure 2 defining motor vs. sensory categorization show a major
477 structural difference between the two processes. While sensory perception implements a direct
478 association between the sensory input S and the perceived object O_L , motor perception
479 appears to be more complex. Indeed the pathway from S to O_S involves motor information, M .
480 This suggests that motor recognition might require more time or cognitive resources before
481 convergence in the learning process, compared to sensory recognition. A possible
482 consequence is that the sensory system should be able to focus more rapidly and efficiently on
483 the set of exogenous learning stimuli provided by the environment, while the motor system
484 “wanders” through the sensory-motor space and endogenously explores regions, possibly

485 different ones from the exogenous input. This would provide the sensory and motor systems
486 with what we have called a “narrow-band” vs. “wide-band” specificity with respect to the
487 learning data. The latter would be less efficient for learned stimuli, but would function better
488 in adverse conditions involving unlearned stimuli.

489 This is what we set out to demonstrate, on a highly simplified theoretical framework
490 based on 1-D motor and sensory variables linked by a sigmoid transformation. In this section
491 the variables of the *COSMO* model are constrained to be very simple and are instantiated as
492 follows: M and S are 1-D and discrete (with values regularly distributed between -15 and 15),
493 while O_S and O_L both denote two possible objects o_1 and o_2 . The Master Agent and the
494 Learning Agent correspond to two different instances of the *COSMO* model with the same
495 parametric forms (mostly Gaussian probability distributions) mathematically encoding the
496 knowledge stored in the models. The two types of agent only differ by the values of the
497 parameters of these parametric forms (for instance, means and standard deviations of the
498 Gaussian probability distributions). We consider a supervised learning situation, where the
499 parameters of the Master Agent and of the motor-to-sensory transformation performed by the
500 simulated environment are fixed and the Learning Agent determines values for its parameters
501 of internal representations through interactions with the Master according to the supervised
502 learning scenario shown in Figure 3. We now describe the probability distribution forms and
503 the parameters that are constant throughout learning. The prior objects $P(O_S)$ for both types of
504 agent are set as uniform probability distributions; objects o_1 and o_2 are produced by the
505 Master with the same frequency and the Learning Agent has no prior knowledge of the
506 frequency of object apparition.

507 For both types of agent, motor repertoire probability distributions $P(M | O_S)$ are
508 encoded as Gaussian probability distributions. For instance, to select a motor command
509 corresponding to object o_1 , the Master Agent draws a value of M^{Master} according to the

510 probability distribution $P(M^{Master} | [O_S^{Master} = o_1]) = Gauss(\mu_1^M, \sigma_1^M)$, where the mean
 511 value μ_1^M of the Gaussian probability distribution corresponds to a prototypic motor gesture
 512 and the standard deviation σ_1^M quantifies the variability of the Master Agent's production. In
 513 the Master Agent model, we set $\mu_1^M = -5$, $\mu_2^M = 5$ and $\sigma_1^M = \sigma_2^M = 1$ (see Figure 4, bottom
 514 plot).

515 The motor-to-sensory transformation $P(S | M)$ occurring in the environment is
 516 modeled as Gaussian probability distributions. More precisely, when the Master Agent issues
 517 a motor command m , the Learning Agent receives a value of the sensory input S^{Ag} drawn
 518 according to the probability distribution $P(S^{Ag} | [M^{Master} = m]) = Gauss(\mu_m^S, \sigma_m^S)$, where
 519 the value $\mu_m^S = f(m)$ is given by a function f modeling the motor-to-sensory transformation
 520 and $\sigma_m^S = \sigma^{Env} = 1$ is a constant encoding the communication noise at learning time.

521 ***FIGURE 4 ABOUT HERE***

522 Next, we consider nonlinear monotonous transformations, to keep some level of
 523 generality. Interestingly, nonlinear motor-to-sensory transformations have been exploited by
 524 the Quantal Theory of Speech (Stevens, 1972) as providing natural category boundaries for
 525 phonetic contrasts. In the Quantal Theory of Speech, it is proposed that such nonlinearities
 526 lead to the existence of articulatory plateaus, where variations in the articulation input lead
 527 essentially to no, or only small, acoustic variations. These are separated by discontinuity
 528 regions, where a small articulatory variation results in a strong acoustic jump. Stevens (1972,
 529 1989) suggested that human languages exploit such discontinuities to set universal phonetic
 530 contrasts. This principle was confirmed in COSMO simulations of the emergence of
 531 phonological systems (Moulin-Frier et al., 2015). In the present study, we define the physical
 532 link f between the motor gestures M and their sensory consequences S as a sigmoid function
 533 $f(m) = b \cdot \frac{\tan^{-1}(a \cdot m)}{\tan^{-1}(a \cdot b)}$, which is shown in Figure 4 (top left plot). Parameter a allows the slope
 534 of the sigmoid function f to be tuned and parameter b controls its range. We selected a b

535 value of 12, slightly lower than the M and S values of 15 and we set a to either 0.01 in a
536 quasi-linear case, or 0.1 to obtain a nonlinear case compatible with the Quantal Theory. The
537 corresponding probability distributions $P(S^{Ag} | [O_S^{Master} = o_1])$ and $P(S^{Ag} | [O_S^{Master} =$
538 $o_2])$ are displayed in Figure 4 (top right plot). In the nonlinear case they are naturally more
539 widely separated in sensory space than the equivalent distribution in motor space.

540 At this stage, we consider that the Master, along with the Learning Agent, have the
541 same nonlinear transformation between their motor and sensory variables. A major departure
542 from this assumption would concern differences in vocal tract shape mainly associated with
543 age and sex. We consider that all agents are equipped with a normalization mechanism
544 enabling them to transform sensory information provided by the Master into an S value
545 appropriately situated in their internal sensory space. Such normalization processes exist and
546 have been displayed since the first months of age (e.g. Kuhl, 1979, 1983; Polka, Masapollo
547 and Ménard, 2014). Once the stimuli are transmitted by the Master to the Learning Agent and
548 have been appropriately normalized, the nonlinear transformation is of no further use to the
549 Master Agent. Hence, remaining differences between such transformations in the case of the
550 Master and the Learning Agent play no role in further processing in COSMO. We will return
551 in the discussion of Part 2 to consider how realistic normalization processes could modify the
552 present simulations.

553 In the computer simulations presented below, all Gaussian probability distributions are
554 truncated: we define a baseline value $\varepsilon = 10^{-5}$ and probability values below this threshold
555 are set to ε ; the probability distribution is normalized afterward⁽³⁾. This avoids cognitively
556 implausible numerical precision of probability distributions, which would yield unwanted side
557 effects. For instance, in classification tasks, when comparing the predictions of Gaussian
558 models too far from their mean values, an infinitely precise model generalizes too well and
559 behaves like an analytical model, with a precise classification frontier and abruptly changing

560 responses. In the present simulations, probability distributions, truncated in this manner,
561 degenerate outside of their “competence domains” and classification responses behave
562 according to chance when exotic stimuli are presented.

563

564 **Learning in *COSMO***

565 Hypotheses H1, H2 and H3 are at the basis of the indistinguishability theorem,
566 expressing unrealistic perfect learning conditions. We now add a plausible learning algorithm
567 to the hypotheses and we will describe how the result departs from ideal learning conditions
568 and ultimately enables sensory and motor recognition processes to be distinguished.

569 Learning follows the interaction paradigm introduced as Figure 3. To recapitulate the
570 learning scenario, both the Master and Learning Agent interact in a simulated environment.
571 Probability distributions defining the Master Agent and motor-to-sensory transformation of
572 the environment are set constant during learning. The Master Agent provides the Learning
573 Agent with <object, stimulus> pairs (later referred to as < o, s >). The Learning Agent
574 ascertains from this data both its sensory and motor classification systems; more precisely, it
575 identifies parameters for its sensory prototypes $P(S^{Ag} | O_L^{Ag})$, internal model
576 $P(S^{Ag} | M^{Ag})$ and motor repertoire $P(M^{Ag} | O_S^{Ag})$, which are all implemented using Gaussian
577 probability distributions. To express the fact that the Learning Agent starts without any
578 knowledge, initial states of all these Gaussian probability distributions are characterized by
579 values of mean parameters μ at the center of their domains and initial standard deviation
580 values σ that are large relative to the domain size. This approximates to uniform probability
581 distributions.

582 To allow fair comparisons of the sensory and motor instantiations of the speech
583 perception task, the components of the sensory and motor classification systems are learned
584 independently and using the same data.

585 In the sensory recognition system based on a direct association between stimuli and
586 objects, firstly we learn sensory prototypes of the form $P(S^{Ag} | [O_L^{Ag} = o]) = Gauss(\mu_o^S, \sigma_o^S)$
587 which correspond to each object. Learning consists of computing a Gaussian probability fit
588 distribution $P(S^{Ag} | O_L^{Ag})$ from $\langle o, s \rangle$ pairs. Each time the Learning Agent receives such a
589 pair from its Master, the values of the mean μ_o^S and of the variance σ_o^S of the corresponding
590 Gaussian probability distribution are updated accordingly. An extensive study of the
591 dynamics of learning sensory prototypes and their effects on the resulting classifiers is
592 provided by Kleinschmidt & Jaeger (2015); as such dynamics are not the focus of our
593 contribution, we implement a straightforward learning procedure, where the order of learning
594 data has no effect.

595 The motor recognition system exploits the same $\langle o, s \rangle$ pairs to learn both
596 components of its pathway, namely the internal model of the motor-to-sensory transformation
597 $P(S^{Ag} | M^{Ag})$ and the motor repertoire $P(M^{Ag} | O_S^{Ag})$. We exploit a *Learning by*
598 *Accommodation* algorithm, which allows learning of the two components at the same time.
599 Importantly, this algorithm takes into account “babbling drift”, i.e. the fact, presented in the
600 Introduction, that the agent should not explore systematically and uniformly its sensory-motor
601 space but rather should focus on regions of interest provided by the Master's sounds. The
602 accommodation algorithm enables the Learning Agent to progressively focus on the Master's
603 stimuli, making learning quicker and more efficient (Barnaud, Schwartz, Diard, & Bessière,
604 2016).

605 The algorithm involves a simple imitation paradigm without any error measurements.
606 It works in the following way:

607 (i) The Learning Agent tries to mimic the sensory input s of the $\langle o, s \rangle$ pair provided
608 by the Master, by producing a motor command m given its current state of knowledge, the
609 input stimulus s and the input object o . After probabilistic inference, this amounts to

610 randomly drawing a value for m according to the following probability distribution:

$$611 \quad P(M^{Ag} | [S^{Ag} = s][O_S^{Ag} = o]) = P(M^{Ag} | [O_S^{Ag} = o]).P([S^{Ag} = s] | M^{Ag}). \quad (4)$$

612 Equation (4) shows that the choice of motor commands m is driven by two factors:

613 first, the need to match the stimulus s given by the Master, as predicted by the current state of
 614 knowledge encoded in the internal model $P(S^{Ag} | M^{Ag})$; second, the tendency to use the same
 615 motor commands that were previously associated with the object o communicated by the
 616 Master, as stored in the motor repertoire $P(M^{Ag} | O_S^{Ag})$.

617 (ii) Once selected, the motor command m is performed and has a sensory consequence
 618 s' . The Learning Agent then uses the observed correspondence of s' and m to improve the
 619 internal model by updating the parameters μ_m^S and σ_m^S of the probability distribution
 620 $P(S^{Ag} | [M^{Ag} = m])$. It also exploits the selected value m in its motor repertoire by updating
 621 the μ_o^M and σ_o^M parameters of the probability distribution $P(M^{Ag} | [O_S^{Ag} = o])$.

622 Therefore, the algorithm progressively refines the internal model of motor-to-sensory
 623 mapping, both with some endogenous random exploration due to inaccurate imitation in the
 624 first stages of the learning process and with a progressive focus on the learning stimuli that
 625 result in a better mapping around the regions of the stimuli provided by the Master. In
 626 parallel, the algorithm progressively anchors adequate motor gestures for each object, i.e.
 627 gestures producing sounds that correspond to the sensory distribution produced by the Master
 628 for the object.

629

630 **Simulation results**

631 **Learning pace: fast and focused sensory learning vs. slow and diffuse motor**
 632 **learning.**

633 ***FIGURE 5 ABOUT HERE***

634 We use the evolution of entropy $H(P(X)) = -\sum_i P(X = x_i) \log P(X = x_i)$ as a

635 numeric indicator that quantifies how much information becomes stored in the probability
636 distributions of the models. We compare learning speeds using the evolution of
637 $H\left(P(S^{Ag} | O_L^{Ag})\right)$, the entropy of the sensory model on the one hand and $H\left(P(S^{Ag} | O_S^{Ag})\right)$,
638 the entropy of the motor model on the other hand. $H\left(P(S^{Ag} | O_S^{Master})\right)$, the entropy of the
639 probability distribution over the stimuli produced by the Master Agent, which is constant
640 during learning, is used as a reference. Each of these entropy values is actually a set of
641 measurements, one for each possible object value. We therefore average them over objects
642 and, since we have two objects in these 1-D experiments, consider $\frac{1}{2}\sum_{O_L^{Ag}} H\left(P(S^{Ag} | O_L^{Ag})\right)$
643 for the sensory model, $\frac{1}{2}\sum_{O_S^{Ag}} H\left(P(S^{Ag} | O_S^{Ag})\right)$ for the motor model and
644 $\frac{1}{2}\sum_{O_S^{Master}} H\left(P(S^{Ag} | O_S^{Master})\right)$ as the Master's reference.

645 The corresponding curves are displayed in Figure 5 for the two values of nonlinearity
646 in the motor-to-sensory transformation. This Figure shows that the entropy of the sensory
647 model converges quickly to a level close to the entropy of the stimuli produced by the Master,
648 while the entropy of the motor model converges more slowly. In the linear case, the sensory
649 model is able to converge to exactly the same entropy as that of the Master Agent, whereas it
650 remains larger in the nonlinear case because the constraint that distributions $P(S^{Ag} | O_L^{Ag})$ are
651 Gaussian leads to some residual discrepancy between the models of the Master Agent and
652 Learning Agent.

653 However, whatever the nonlinearity, the motor model entropy $H\left(P(S^{Ag} | O_S^{Ag})\right)$
654 decreases more slowly than the sensory model entropy, indicating slower learning. This
655 corresponds to our prediction that the inference process is more complex in the motor model.
656 Hence the learning mechanism is slower and less efficient, since it “visits” portions of the
657 sensory-motor space that are not available to the sensory recognition system.

658 ***FIGURE 6 ABOUT HERE***

659 To support this point, let us recall that learning the motor $P(S^{Ag} | O_S^{Ag})$ model consists
660 of learning both $P(S^{Ag} | M^{Ag})$ and $P(M^{Ag} | O_S^{Ag})$. We show in Figure 6 an instance of an
661 internal $P(S^{Ag} | M^{Ag})$ model in the nonlinear case learned by the agent after 20,000 iterations
662 of learning by the accommodation algorithm. Data presented in Figure 6 first show that the
663 shape of the motor-to-sensory transformation has been adequately learned. However, it
664 appears that some regions are learned better than others: these regions, where the variance of
665 the $P(S^{Ag} | [M^{Ag} = m])$ distribution is small, correspond to those of the sensory space where
666 stimuli have been provided by the Master Agent. Other regions, far from the data of the
667 learning set, have a higher variance. However, the Learning Agent has acquired global
668 knowledge, which provides the motor learning process with what we could call a more
669 “diffuse” character.

670 In this learning process, we have implemented a mechanism that naturally leads to the
671 departure from both hypotheses H1 and H2 of the “perfect learning conditions”. Since
672 learning is intrinsically incomplete, the Learning Agent cannot fully internalize all the
673 production abilities of the Master Agent. This results in complementarity between the sensory
674 and motor models. While the sensory system can focus on the stimuli provided by the Master
675 Agent and learn them quickly and efficiently, the motor system has to learn both a sensory-
676 motor model and a motor repertoire. This more complex process is slower and less focused on
677 the learning set, because it requires exploring an intermediate motor space. However, this can
678 be useful for unlearned conditions as we will assess next.

679

680 **Evaluation of perception: the sensory model is better in clear speech, whereas the**
681 **motor model is more robust in noisy conditions.**

682 In this section we compare the models’ robustness to communication noise in an

683 evaluation experiment where the Learning Agent interacts with a Master Agent defined in the
 684 same way as previously, except that after the learning phase, we introduce a test phase where
 685 we vary the standard deviation $\sigma_m^S = \sigma^{Env}$ of the Gaussian probability distribution
 686 $P(S^{Ag} | [M^{Master} = m])$, thus encoding various levels of environmental noise.

687 The Master Agent provides $\langle o, s \rangle$ pairs of a given noise level and the agent
 688 estimates the object o' from the stimulus s using either sensory recognition, i.e. by computing
 689 the probability distribution $P(O_L^{Ag} | S^{Ag})$, or motor recognition, i.e. by computing
 690 $P(O_S^{Ag} | S^{Ag})$, as defined in Figure 2. Sensory recognition is implemented as a Gaussian
 691 classifier obtained by probabilistic inversion of the sensory prototypes $P(S^{Ag} | O_L^{Ag})$:

$$692 \quad P(O_L^{Ag} | S^{Ag}) = \frac{P(S^{Ag} | O_L^{Ag})}{\sum_{O_L^{Ag}} P(S^{Ag} | O_L^{Ag})}.$$

693 Motor recognition is implemented according to:

$$694 \quad P(O_S^{Ag} | S^{Ag}) \propto \sum_{M^{Ag}} (P(M^{Ag} | O_S^{Ag}) P(S^{Ag} | M^{Ag})).$$

695 Comparing the values of the object intended by the Master Agent, o , and that
 696 estimated by the Learning Agent, o' , we compute confusion matrices and define the
 697 recognition rate as the mean of their diagonal coefficients.

698 ***FIGURE 7 ABOUT HERE***

699 In Figure 7, we present the mean values of recognition rates for the linear and
 700 nonlinear cases. The scores are provided at three learning stages (after learning 500, 2,000 or
 701 20,000 $\langle o, s \rangle$ pairs), for a range of noise degradation, from no added noise to stimuli
 702 corrupted by high levels of noise (noise is indexed by variation of the σ^{Env} value).

703 First, we observe a large effect of nonlinearity on the sensory classifier, with a sharp
 704 decline of performance with noise in the nonlinear case. This derives from more pointed and
 705 separated probability distributions (Figure 4, top right panel). The observations that follow are
 706 independent of nonlinearity.

707 Second, since the sensory system learns rapidly, it has already converged before 500
708 learning iterations and does not evolve afterwards. It provides good recognition scores
709 without noise, with a quick degradation of performance when noise is added.

710 Third, in contrast the motor system appears to learn slowly. At the beginning of the
711 learning process (top row in Figure 7), it performs very poorly and the decrease of the
712 recognition rate as noise increases is slower than for the sensory model. When learning
713 proceeds with more iterations, the motor system performs increasingly well, the general trend
714 being that it becomes better than the sensory model in noisy conditions, though still remaining
715 poorer in the absence of noise. At the last stage of the learning process (20,000 iterations) the
716 two models give rather similar performances (we tend towards the “perfect learning
717 conditions” of the indistinguishability theorem).

718 Fourth, and finally, the perceptuo-motor model implementing a Bayesian fusion of the
719 sensory and motor recognition models according to the Equation shown in Figure 2 performs
720 better than the two isolated models under all conditions.

721 ***FIGURE 8 ABOUT HERE***

722 We now explore how the sensory system is more efficient in the absence of noise and
723 how the motor system is more efficient in its presence. This is illustrated in Figure 8, where
724 we display probability distributions for the two objects, for both motor and sensory systems.
725 Furthermore, we show the example s_{clean} stimuli for a stimulus under normal conditions, i.e.
726 without added noise, and s_{noise} for a stimulus in adverse conditions, i.e. with added noise.

727 When the “typical” s_{clean} stimulus is considered, it is close to prototypes of the motor
728 and sensory models, i.e. to the modes of corresponding probability distributions. However, the
729 sensory model, being of lower variance than the motor model, yields a less uncertain
730 probability distribution categorization than the motor process. The two models correctly
731 recognize object o_2 as the cause of the s_{clean} stimulus, but the sensory model is slightly more

732 certain of perception than the motor model is.

733 When the “noisy” stimulus (s_{noise}) is considered, it is far from prototypes of both
734 motor and sensory models. However, the motor model, being of greater variance than the
735 sensory model, generalizes better. Whereas for the sensory model, probability distributions
736 quickly fall below the ϵ threshold we defined, yielding random categorization, the motor
737 model is more robust and conserves categorization capabilities.

738

739 **Concluding Part 2: summary and predictions**

740 The present simulations let a major difference between auditory and motor learning
741 appear. Auditory learning is rapid and, by definition, perfectly focused on the acoustic stimuli
742 provided by the Master. In fact, the auditory system in COSMO is an “ideal processor” of
743 acoustic input, as in many previous Bayesian models (e.g. Norris & McQueen, 2008;
744 Kleinschmidt & Jaeger, 2015). However, the intrinsic limitation is provided by departures
745 from exactly what has been learned. This is where the motor system may become relevant.
746 Indeed, the motor system is intrinsically slower since it has a more complex inference process
747 to deal with. It is also less well tuned to the learning corpus, because of the existence of an
748 intermediate motor representation in the inference process. But it is this more complex
749 learning process that supplies the possibility of wandering around stimuli and configurations
750 that are not contained in the learning set provided by the environment. This is what makes it
751 “wider” and hence better able to process unknown stimuli.

752 It is important to stress at this stage that the auditory-narrow, motor-wide hypothesis
753 appears to be generic, i.e. intimately related to the basic COSMO structure, because of the
754 more complex structure of the motor inference process compared with the auditory one (see
755 Figure 2). We had to propose a number of technical and non-generic choices to perform
756 simulations in this part of the article. These include: (i) the motor-to-sensory transformation

757 was presumed to be nonlinear but monotonous, (ii) the Master was considered to be
758 physically similar to the Learning Agents, with the same nonlinear motor-to-sensory
759 transformation, supposing that the normalization process was solved in some way, (iii) only
760 one Master was introduced into the learning process, while an infant typically has to deal with
761 a number of Masters to learn from in the environment.

762 More realistic simulations, involving: non-monotonous motor-to-sensory
763 transformations, variations of transformations from one agent to the other, possibly in relation
764 with normalization processes between agents with different sizes and shapes of their vocal
765 tract, multiple Master Agents in the learning process, would basically result in a large increase
766 in complexity of the sensory-to-motor inference process and hence in an increase toward the
767 trend for slow and diffuse motor learning. In some sense, the 1-D simulations presented in
768 Part 2 minimize the trend towards the auditory-narrow vs. motor-wide contrast, which is
769 likely to be larger in a more realistic simulation with COSMO – as will be displayed in Part 3.
770 Therefore, it can safely be claimed that the auditory-narrow motor-wide hypothesis is an
771 intrinsic property of the COSMO structure, and probably an intrinsic characteristic of motor
772 vs. auditory decoding in a perceptuo-motor theory of speech perception.

773 This property of the model generates two predictions, in the sense that two
774 consequences follow directly from the property. These consequences were not considered
775 during modeling; they are logically entailed by the model. These predictions are in line with
776 already available data and observations pertaining to speech development and processing.

777 **Prediction 1 - auditory learning should be more rapid than motor learning**

778 A strong prediction in COSMO is that auditory learning, which typically consists of
779 learning the sensory distributions $P(S^{Ag}/O_L^{Ag})$, is a simpler process than motor learning i.e.
780 learning the motor distributions $P(M^{Ag}/O_S^{Ag})$. It is well-known that the auditory system is
781 developmentally mature before the motor one, as is reviewed in the Introduction, but this is

782 generally only related to biological constraints. Firstly, audition begins to mature before birth,
783 as is displayed by the sensitivity of newborns to language (Mehler, Jusczyk, Lambertz,
784 Halsted, Bertoni, & Amiel-Tison, 1988) or to the voice of their mother (DeCasper & Fifer,
785 1980). Secondly, critical periods seem to shape the course of development of speech
786 perception and production towards the mature stage (see a recent review in Werker & Hensch,
787 2015). Importantly, the present simulations suggest that an additional factor could be provided
788 by the complexity of the learning process. In this respect, it is of interest to mention that even
789 for vowels, language tuning in production has never been described before 10 months of age
790 (e.g. de Boysson-Bardies et al., 1989) while it occurs in speech perception as soon as 6
791 months of age (Kuhl et al., 1992), though infants are capable of producing vowel-like
792 vocalizations almost since birth and display vocal imitations as early as 4 months of age (Kuhl
793 & Meltzoff, 1996).

794 It would be of great interest in this discussion to attempt to correlate observed delays
795 in the developmental schedule with some measurement of differences in the learning period or
796 entropy reduction in a Figure such as Figure 5. However, this seems far from any reasonable
797 prediction at the present state of possible simulations.

798 **Prediction 2 - motor processing should be more important in adverse conditions**

799 This is the major prediction of the auditory-narrow motor-wide hypothesis. Indeed, it
800 is proposed as an intrinsic COSMO property that the motor system should be less efficient
801 than the auditory system in learned conditions, while the motor system gains efficiency in
802 unlearned ones, e.g. in noisy or adverse conditions. A likely consequence of this prediction is
803 that the motor system should be more involved in such adverse conditions. As reported
804 previously, this is exactly what is regularly observed for neurocognitive data, with an
805 increased BOLD (Blood-Oxygen Level Dependent) activity in fMRI (functional Magnetic
806 Resonance Imaging) data in motor regions for noisy (Binder et al., 2004; Zekveld et al.,

807 2006), or non-native stimuli (Callan et al., 2004, 2014; Wilson & Iacoboni, 2006). This is also
808 in line with evidence for motor perturbations seen in auditory perception only in noisy
809 conditions (e.g. d'Ausilio et al., 2012 vs. 2009), or for ambiguous stimuli around a phonetic
810 boundary (e.g. Möttönen & Watkins, 2009; Rogers et al., 2014).

811

812

813 Part 3 – Extracting perceptuo-motor invariance in syllabic units

814 While in the previous part the focus was on generic properties of the COSMO model,
815 we now move towards non-generic properties associated with the specific way auditory and
816 motor information is most probably distributed in a specific case of phonetic sequences made
817 of CV syllables with a C stop consonant and a V oral vowel. This part of the article will deal
818 with a problem that has long been considered crucial in the debate about auditory vs. motor
819 theories, namely the invariance problem. The question of invariance has often been raised by
820 motor theorists around the alleged lack of acoustic invariance for the plosive place of
821 articulation, considering that in this specific case motor invariance was straightforward (see
822 below). However, the case of vowels seems different and it has already been suggested that
823 invariance could be of a different nature for vowel vs. plosive place of articulation, which
824 would be auditory in one case and articulatory in the other (see Bailly, 1997; Kröger et al.,
825 2009, 2014). Therefore, COSMO appears as a perfect framework for dealing with this
826 question in a perceptuo-motor framework.

827 To address the question of auditory vs. motor invariance for vowel vs. plosive place of
828 articulation, we will need to introduce specific knowledge and hypotheses concerning CV
829 syllable production, perception and development. Furthermore, simulations will be carried out
830 on a specific model of the vocal tract, VLAM (variable linear articulatory model), enabling
831 generation of articulatory and acoustic configurations associated with CV sequences. Finally,
832 simulations will include specific simplifications about sensory and motor variables, as well as
833 about the learning process.

834 In light of this specific implementation of COSMO for syllables, which we will call
835 COSMO-S, the question we address is: in the distribution of information for plosives and
836 vowels, is there any potential evidence for differentiation of the motor and auditory systems in
837 extracting phonetic invariance from acoustic stimuli? We will firstly present a literature

838 review on the perception and production of CV syllables in relation to the place of articulation
839 cues. Then we will describe the vocal tract model VLAM, and the COSMO-S version of
840 COSMO for syllable perception and production, together with the way learning is
841 implemented in COSMO-S. Finally, we will describe simulations with COSMO-S and
842 explore what light they might shed on the question of vowel vs. plosive place of articulation
843 invariance.

844

845 **Auditory or motor cues for vowel vs. plosive place of articulation**

846 The question of the plosive place of articulation invariance has long been considered
847 as a crucial test for auditory vs. motor theories of segmental invariance. On the one hand,
848 partisans of motor theories have regularly mentioned it as a typical case, where auditory
849 invariance was out of reach while motor invariance would be directly available (Liberman et
850 al., 1967; Liberman & Mattingly, 1985). On the other hand, the classical objection to the
851 motor theory is the probable complexity of the cognitive or computational implementation of
852 the inversion process that would enable the listener to recover the proposed motor invariant
853 from the acoustic speech input: the labial gesture for bilabials, the tongue apex gesture for
854 coronals, the tongue dorsum gesture for palato-velars.

855 Partisans of auditory theories have also searched possible invariant auditory cues that
856 characterize the plosive place of articulation. The pioneer work by Delattre, Liberman, &
857 Cooper (1955) on the “acoustic locus” actually served as a precursor for both auditory and
858 motor proposals on invariance. In the framework of his Quantal Theory, Stevens proposed at
859 the end of the 1970s that there might be a local spectral invariant for the plosive place of
860 articulation, located around the position of the acoustic burst and independent of the speaker,
861 the plosive manner of articulation and the context. Bilabial spectra would be “diffuse falling”
862 (with energy all over the spectrum but more of it at low frequencies), alveolars would be

863 “diffuse raising” (idem but with more energy at high frequencies) and velars would be
864 compact (with most of the energy packed into the medium) (Blumstein & Stevens, 1979;
865 Stevens, 1980; Stevens & Blumstein, 1978). Following further proposals by Kewley-Port
866 (1983), a progressive shift was made towards dynamic cues associating spectral values with
867 the plosive and the next vowel. At the end of this process, the locus came back with the “locus
868 equations” introduced by Sussman (Sussman, Fruchter, Hilbert, & Sirosh, 1998; Sussman,
869 Hoemeke, & Ahmed, 1993; Sussman, McCaffrey, & Matthews, 1991) assuming relational
870 invariance (correlations between $F2$ values for the plosive and the next vowel) as a correlate
871 of the place of articulation. Importantly, acoustic characterization of the plosive place of
872 articulation seems basically to rely on spectral data at two instants; plosive release and vowel
873 climax.

874 Our proposal is different. In the PACT framework and in light of the perceptuo-motor
875 developmental schedule described at the beginning of this paper, we presume that in a first
876 stage, speech perception would benefit from rapidly maturing auditory processing that enables
877 infants to categorize all CV sequences available in their environments. In this first stage, the
878 motor system would not be mature and probably not even completely functionally related to
879 the speech perception system according to Kuhl et al. (2014). Hence, the infants would not
880 have at their disposal invariant cues for the plosive place of articulation. This question is
881 debated, with negative results on plosive invariance before 6 months in Bertoncini, Bijeljac-
882 Babic, Jusczyk, Kennedy, & Mehler, 1988; Eimas, 1999; vs. data suggesting the possibility to
883 discriminate /b/ from /d/ at 6 months of age, Hochmann & Papeo, 2014; and a discussion on
884 possible confounding effects in Dole, Loevenbruck, Pascalis, Schwartz, & Vilain, 2015.

885 In a second stage, after 7 months there is progressive coupling and maturation of the
886 speech motor system. Then, infants could discover that plosive-vowel sequences heard in the
887 environment are produced by specific movements of the lips for bilabials, and the tongue apex

888 or dorsum for alveolars or velars. Hence, the content of the motor repertoire would enhance
889 perceptual representations and allow invariance to emerge in a perceptuo-motor space.

890 For vowels, the situation is probably different. Indeed, auditory representations for
891 oral vowels have been described in a number of studies, and oral vowels seem properly
892 characterized in all their phonetic dimensions in a bundle of frequency parameters (e.g. ($F1-$
893 $F0$) for height, ($F2-F1$) for place of articulation and $F'2$ for rounding; all values are in Barks:
894 see Ménard, Schwartz, Boë, Kandel, & Vallée (2002). In contrast, the articulatory
895 characterization of oral vowels is less straightforward (e.g., Boë, Perrier, & Bailly, 1992) and
896 perturbation experiments suggest that invariants for vowels could be auditory rather than
897 motor (Savariaux, Perrier, & Orliaguet, 1995; Savariaux, Perrier, Orliaguet, & Schwartz,
898 1999). It is more in terms of vowel reduction that articulatory dynamics could play a role,
899 though the debate on this topic was vigorous in the 1980s and 1990s (e.g., Strange (1989) vs.
900 Nearey (1989) or Perrier, Lœvenbruck, & Payan (1996) vs. Pitermann (2000)).

901 Therefore, our hypothesis is that the auditory and motor systems could be
902 complementary in terms of the content of their representations for phonetic invariance, motor
903 or gestural cues probably being crucial for the plosive place of articulation, while auditory
904 parameters would be efficient for vowel characterization⁽⁴⁾. This is what we now propose to
905 test with COSMO. For this aim, since natural articulatory data are sparse, particularly about
906 perceptuo-motor development early in life, we will use synthetic data in the framework of the
907 articulatory model of the vocal tract, *VLAM*.

908

909 **VLAM and the generation of synthetic CV syllables**

910 *VLAM* is a realist vocal tract model (Maeda, 1990) thanks to which seven articulatory
911 parameters (*Jaw*, *Larynx*, *TongueBody*, *TongueDorsum*, *TongueApex*, *LipHeight*,
912 *LipProtrusion*) have been derived from a guided principal component analysis of

913 cineradiographic images of the vocal tract. These allow the description of the jaw and larynx
914 position, and of the tongue and lips shape. The parameters can be interpreted in terms of
915 phonetic and muscular commands (Maeda & Honda, 1994). The areas of 28 sections of the
916 vocal tract are estimated as linear combinations of these seven parameters, which then allows
917 computation of the transfer function and formants (Badin & Fant, 1984) (see Figure 9).

918 ***FIGURE 9 ABOUT HERE***

919 In short, *VLAM* is a geometric model enabling formants from articulatory parameters
920 to be computed. This model has been evaluated over the last fifteen years in terms of its
921 ability to generate vowels and plosive stimuli compatible with data from infants (Boë et al.,
922 2013), children and adults (Laurent et al., 2013; Ménard et al., 2002; Ménard, Schwartz, &
923 Aubin, 2008; Ménard, Schwartz, & Boë, 2004; Schwartz, Boë, Badin, & Sawallis, 2012b). It
924 is also the articulatory synthesizer of the DIVA (Directions Into Velocities of Articulators)
925 model of speech production (Guenther, 2006; Guenther, Hampson, & Johnson, 1998).

926 Here, *VLAM* is considered as a simplified implementation of the motor-to-auditory
927 relationship in the human vocal tract ⁽⁵⁾. It is used both to generate CV syllables thought to be
928 produced by the Master Agent and as an external simulator of the Learning Agent' vocal tract
929 so that it can learn from the perceptual consequences of the motor commands it is sending.
930 *VLAM* also incorporates a model for vocal tract scaling associated with age, thanks to which
931 the size increases with age in a nonlinear way compatible with experimental data (see Boë et
932 al., 2013; Ménard et al., 2002, 2008). However, as in Part 2, we do not consider here vocal
933 tract differences between the Learning Agent and the Master, supposing that if there were any,
934 they could be solved by appropriate normalization processes (see Ménard et al., 2002).

935 **Generation of oral vowels**

936 Vowels are defined as articulatory configurations that are not too closed, so as not to
937 generate noise in their acoustic output. This is characterized in *VLAM* by setting a constraint

938 on the constriction, which is the position of the section of the vocal tract with the smallest
939 area. The constriction area for vowels is higher than a minimum value of 0.15 cm^2 . In the
940 present set of simulations, we only considered the three extreme oral vowels /i, a, u/, which
941 provide the preferred choice in human languages (see Schwartz, Boë, Vallée & Abry, 1997).

942 Any speech sound should need all 7 VLAM parameters for its complete generation
943 and characterization. However, we have attempted to keep the number of free parameters at
944 the smallest possible value to minimize later computations. Hence, vowels are described here
945 by three VLAM articulatory parameters (*TongueBody*, *TongueDorsum* and *LipHeight*), all
946 other parameters being set to a neutral value (resting position). We define motor vowel
947 prototypes for /a i u/, using average formant values for French vowels (Meunier, 2007) as
948 targets and selecting values of the three VLAM parameters that best fit the acoustic target. For
949 each category of vowel, we generated a set of articulatory configurations according to a
950 Gaussian probability distribution centered on the prototype value.

951 **Generation of stop consonants**

952 Plosives are defined as articulatory configurations achieved just after a complete
953 closure of the vocal tract, i.e. at the time of acoustic release, which typically generates an
954 acoustic burst. In the present simulations we characterize plosives by the formants produced
955 with a constriction close to, but still slightly higher than, zero, so as to be able to compute
956 formants. We only considered the three extreme plosive places of articulation (labial, alveolar,
957 velar) that provide the preferred choice in human languages (see Schwartz et al., 2012b). The
958 unvoiced stop consonants /p, t, k/ corresponding to these places of articulation are more
959 frequent in human language than their voiced counterparts /b, d, g/, but, in the rest of this
960 paper, we keep the voiced set of consonants /b, d, g/, because voiced plosives provide the
961 clearest formant trajectories and enable a better specification of formants at the beginning of
962 the opening trajectory from the plosive to the next vowel.

963 We adopt the view proposed by Öhman (1966) that plosives are local perturbations
964 (vocal tract closing gestures) of vowel configurations within CV syllables. Therefore, we
965 synthesize plosives by closing the vocal tract from a vowel position, using *the VLAM Jaw*
966 parameter combined with one other articulator, i.e. *Jaw* and *LipHeight* for /b/, *Jaw* and
967 *TongueApex* for /d/, and *Jaw* and *TongueDorsum* for /g/. Hence, plosives are described by
968 five parameters (*Jaw*, *TongueBody*, *TongueDorsum*, *TongueApex* and *LipHeight*).
969 Furthermore, the perturbation gesture allowing a consonant to be produced from a vowel is
970 characterized by two parameters: the variation (Delta) of *Jaw* and another one from among
971 *LipHeight*, *TongueApex* or *TongueDorsum*. To obtain a consonant, both articulators should be
972 combined, so that the vocal tract constriction area reaches a value between 0.05 and 0.15 cm².
973 More specifically, the set of consonants that can be achieved from a vowel configuration of
974 the vocal tract is the set of configurations obtained by 1) going through all possible discrete
975 values of the parameter *Jaw* and 2) for each of these values selecting the value of the other
976 articulator (*LipHeight*, *TongueApex* or *TongueDorsum*) such that when the perturbation is
977 applied to the vowel the constriction area is the closest possible to 0.05 cm². The choice of
978 modeling a consonant as a perturbation added to a vowel means that consonants and vowels
979 are linked by maximal co-articulation.

980 **Representation of CV sequences**

981 Once stop consonants and vowels have been defined in terms of articulatory and
982 acoustic parameters, the question is to define an adequate representation of the trajectory from
983 C to V, characterizing the syllable in articulatory and acoustic terms. Since we showed in the
984 previous section that the data converge towards a characterization based on plosive onset and
985 vowel formants, plosive-vowel syllables are characterized as a pair of two articulatory states:
986 one for the plosive and the other for the vowel, neglecting the geometry and temporal aspects
987 of the trajectory linking these two states. Altogether, a CV sequence is associated in VLAM

988 with 8 articulatory parameters, 5 for the plosive and 3 for the vowel.

989 In the acoustic space, vowels are characterized by their first two formants
990 ($F1, F2$), which *VLAM* computes from the articulatory parameters in the open state. For
991 plosives, where $F1$ is basically the same for all configurations (around 250 Hz),
992 characterization is by $F2$ and $F3$, computed by *VLAM* in the closed state. A CV sequence is
993 associated to 4 acoustic parameters, 2 for the plosive and 2 for the vowel.

994 ***FIGURE 10 ABOUT HERE***

995 Figure 10 displays the acoustic properties of the vowels and plosives generated. The
996 representation of vowels in the ($F1, F2$) plane is classical, with /i, a, u/ at the corners of the
997 vowel triangle (Figure 10, top). The representation of plosives in the ($F2, F3$) plane is less
998 common (Figure 10, bottom). It has been extensively discussed in Schwartz et al. (2012b). We
999 observe that there is a trend toward lower ($F2, F3$) values for /b/, higher values for /d/ and
1000 medium values for /g/. This recalls the “diffuse falling” vs. “diffuse raising” vs. “compact”
1001 contrasts proposed by Stevens and Blumstein (1978), but with considerable variations of the
1002 plosive recognition depending on the vowel context.

1003 ***FIGURE 11 ABOUT HERE***

1004 We show in Figure 11 the relationship between the $F2$ values for plosives and vowels,
1005 providing a portrait that is globally coherent with the one reported by Sussman et al. (1998)
1006 for natural speech.

1007 The two-state implementation of syllabic trajectories is highly simplified in relation to
1008 natural CV sequences, and a number of more elaborate CV co-articulation models have been
1009 suggested since the pioneer one proposed by Öhman (1966). However, here we merely aimed
1010 to generate syllables whose variability patterns were similar to the complexity of real speech
1011 signals. The syllable material displayed in Figure 10 provides an adequate compromise. It
1012 corresponds to complex variations in an 8-D articulatory space and resulting in variations in a

1013 4-D acoustic space with co-articulation patterns that are globally coherent with those of
1014 natural syllables. We will next examine how the motor and auditory systems of the *COSMO*
1015 model extended to syllables can deal with this variability.

1016

1017 ***COSMO-S*, an extension of the *COSMO* model to process plosive-vowel syllables**

1018 ***FIGURE 12 ABOUT HERE***

1019 We have extended the *COSMO* model to CV syllable processing. The objects, O_S from
1020 the speaker's point of view, and O_L from a listener's perspective, refer to the syllables we
1021 consider: /ba/, /bi/, /bu/, /ga/, /gi/, /gu/, /da/, /di/, /du/. Since we model a syllable as a vowel
1022 state and a consonant state, the variable S separates into S_V and S_C , and the variable M into
1023 M_V and M_C . Apart from that, the *COSMO-S* model (see Figure 12, top) shares its global
1024 structure with *COSMO* as it is made of the same systems: (i) the auditory system associates
1025 sensory representations with the corresponding O_L syllable labels; (ii) the sensory-motor
1026 system associates motor and sensory representations; (iii) the motor system associates motor
1027 representations with O_S syllable labels.

1028 These systems are linked by λ coherence variables, which are a mathematical tool
1029 used to force duplicate variables to have the same values at all times during probabilistic
1030 inference (Bessière et al., 2013; Gilet, Diard, & Bessière, 2011). This provides a mathematical
1031 implementation of a probabilistic switch, allowing the different parts of the model to be
1032 activated or deactivated during probabilistic inference, thus permitting constraints coming
1033 from the different sub-models to be integrated into the global model. Likewise, the
1034 specification of $C = True$ in an inference task allows the combination of motor and auditory
1035 cues.

1036 The auditory system describes the knowledge the agent has of the link between O_L
1037 syllables and sensory variables: S'_V ($F1$ and $F2$ for the vowel) and S'_C ($F2$ and $F3$ for the

1038 consonant). These are implemented as 4-D Gaussian probability distributions, the mean
1039 vectors and covariance matrices of which are estimated during the learning process (see
1040 below).

1041 The sensory-motor system describes the knowledge the agent has of the motor-to-
1042 sensory mapping, i.e. of mapping between articulatory gestures M_V (vowel), M_C (consonant)
1043 and formant values S_V and S_C . Once again, mappings are described by Gaussian probability
1044 distributions, where mean vectors and covariance matrices are estimated during the learning
1045 process (see below). The term $P(M_C | M_V)$ encodes a support for consonants that can be
1046 achieved according to the perturbation hypothesis described in the section ‘Generation of stop
1047 consonants’. More specifically, for each vowel motor gesture M_V , $P(M_C | M_V)$ defines a
1048 probability distribution that is a plateau in the 5-D articulatory space for consonants. It is
1049 uniform on the possible attainments of consonants obtained by the joint use of the parameter
1050 *Jaw* and another one (either *LipHeight* for /b/, *TongueApex* for /d/ or *TongueDorsum* for /g/),
1051 and it is null everywhere else for configurations that are not consonants (because the vocal
1052 tract is not closed enough) or for configurations that cannot be reached from the M_V vowel
1053 configuration considered. The term $P(M_C | M_V)$ implements a constraint coming from the
1054 physics of the Learning Agent's vocal tract (modeled by VLAM), which does not have to be
1055 estimated in the learning stage. This constraint is implemented using conditional probability
1056 tables, assigning a constant value to each achievable consonant gesture and zero probability
1057 otherwise.

1058 The motor system describes a state of knowledge of the link between O_S syllable
1059 labels and articulatory gestures. The structure of the motor system implements a simplified
1060 co-articulation model based on Öhman’s perturbation hypothesis (Öhman, 1966). This
1061 explicitly introduces a delta variable describing the perturbation superimposed on the vowel to
1062 obtain a plosive consonant. Furthermore, we assume in *COSMO-S* that the Learning Agent

1063 would have at its disposal a set of primitive consonant gestures corresponding to the basic
 1064 places of articulation for plosives: combined jaw and lips for bilabials, combined jaw and
 1065 tongue apex for alveolars, and combined jaw and tongue dorsum for velars. The learning
 1066 process would consist of discovering these basic primitive gestures through motor
 1067 exploration, and identifying their correspondence with the CV sequences provided by the
 1068 Master Agent. Therefore, while vowels in the motor repertoire are characterized by their
 1069 articulatory configuration M'_V (*TongueBody*, *TongueDorsum* and *LipHeight* in *VLAM*),
 1070 plosives are characterized by their primitive gesture G'_C , referring to the articulator used to
 1071 make a plosive consonant in coordination with *Jaw* (*LipHeight* for /b/, *TongueDorsum* for /g/,
 1072 and *TongueApex* for /d/. G'_C is thus a categorical variable, with three possible values).

1073 Variables M'_V and G'_C are taken to be independent. Hence
 1074 $P(M'_V G'_C | O_S) = P(M'_V | O_S) \cdot P(G'_C | O_S)$. The motor configuration of the plosive in the
 1075 framework of Öhman's perturbation theory is then defined by Δ'_{MC} , the variation of the
 1076 articulators (the specific combination of *Jaw* and another specific articulator) necessary to
 1077 achieve a consonant from M'_V . The motor command for the M'_C consonant is finally obtained
 1078 by the equation $M'_C = M'_V + \Delta'_{MC}$. The term $P(\Delta'_{MC} | M'_V G'_C)$ describes how the
 1079 consonant is produced, depending on the vowel and the specific consonant gesture. This
 1080 shows explicitly that the consonant is conditioned by the vowel, which can be interpreted as
 1081 anticipation. For instance, to produce the sound /ba/, the /a/ is anticipated when /b/ is
 1082 performed, which amounts to having maximal co-articulation. $P(M'_V | O_S)$ and
 1083 $P(\Delta'_{MC} | M'_V G'_C)$ are described by Gaussian distributions, where the mean vectors and
 1084 covariance matrices are estimated during the learning process (see below). Finally,
 1085 $P(G'_C | O_S)$ is implemented with a conditional probability table (histogram), the parameters of
 1086 which are also identified during learning.

1087 The COSMO-S model is thus defined by the joint probability distribution

1088 decomposition shown in Figure 12 (bottom).

1089 Similarly to the summary of Figure 2, the Bayesian inference within the *COSMO-S*
 1090 model allows computing of conditional probability distributions. Purely motor, purely
 1091 auditory and perceptuo-motor instances of the speech perception task are implemented.
 1092 Because of the complexity of the *COSMO-S* model, we have not detailed the corresponding
 1093 Bayesian inferences here. However, they can be interpreted exactly as previously: auditory
 1094 perception is expressed as direct use of the link between auditory representations and the
 1095 corresponding object labels, motor perception as the combination of the motor repertoire with
 1096 an internal model allowing association of motor and sensory representations, and perceptuo-
 1097 motor perception as the Bayesian fusion of the auditory and motor categorization processes.
 1098 We will now describe how the Learning Agent acquires the different parts of the model.

1099

1100 **Learning in *COSMO-S***

1101 Some probability distributions of the model are not learned. Indeed, the prior $P(O_S^{Ag})$,
 1102 $P(O_L^{Ag})$ and $P(M_V^{Ag})$ are set as uniform probability distributions. The biological constraints
 1103 $P(M_C^{Ag} | M_V^{Ag})$ describe the consonants achievable from a given vowel, and are pre-computed
 1104 in *VLAM*. Finally, probability distributions over coherence variables, $P(\lambda_{SV}^{Ag} | S_V^{Ag} S_V'^{Ag})$,
 1105 $P(\lambda_{SC}^{Ag} | S_C^{Ag} S_C'^{Ag})$, $P(\lambda_{MV}^{Ag} | M_V^{Ag} M_V'^{Ag})$, $P(\lambda_{MC}^{Ag} | M_C^{Ag} \Delta_{MC}^{Ag} M_V'^{Ag})$ and $P(C^{Ag} | O_S^{Ag} O_L^{Ag})$ are
 1106 set as Dirac probability distributions, with *True* value of a probability of 1 for a given
 1107 relationship between the variables on the right hand side, respectively $S_V^{Ag} = S_V'^{Ag}$, $S_C^{Ag} =$
 1108 $S_C'^{Ag}$, $M_V^{Ag} = M_V'^{Ag}$, $M_C^{Ag} = \Delta_{MC}^{Ag} + M_V'^{Ag}$ and $O_S^{Ag} = O_L^{Ag}$.

1109 The probability distributions that the Learning Agent apprehends in *COSMO-S* are the
 1110 same as in the 1-D implementation studied in the previous section: the auditory categorization
 1111 branch $P(O_L^{Ag} | S_V^{Ag} S_C^{Ag})$, the forward model implementing the motor-to-auditory

1112 relationship $P(S_V^{Ag} S_C^{Ag} | M_V^{Ag} M_C^{Ag})$ and the motor repertoire $P(M_V^{Ag} M_C^{Ag} | O_S^{Ag})$. As
 1113 previously, we learn the auditory and motor branches independently from each other, but with
 1114 the same set of data. This allows a fair comparison between the two branches.

1115 While the forward model and the motor repertoire were learned in 1-D, a two-stage
 1116 process was implemented in *COSMO-S*. Indeed, considering the complexity of the motor-to-
 1117 auditory relationship within a 12-D space (8 motor plus 4 auditory dimensions), it appeared
 1118 easier to learn the forward model before the motor repertoire. This corresponds well to the
 1119 developmental schedule presented previously (Kuhl, 2004), which led us to proceed in three
 1120 consecutive steps:

- 1121 L1. learning the auditory categories;
- 1122 L2. learning motor-to-auditory mapping;
- 1123 L3. learning the motor repertoire.

1124 During these three learning phases, the Learning Agent interacts with a Master Agent
 1125 to obtain syllable acoustic stimuli ($F2, F3$ for the plosive, $F1, F2$ for the vowel) taken from
 1126 the data displayed in Figure 10 and, for steps L1 and L3, the corresponding syllable labels as
 1127 well. Phases L2 and L3 are independent of phase L1; hence they will be evaluated separately
 1128 in the following argument.

1129

1130 **L1: Learning the auditory categories by association.**

1131 As in our previous experiments, the auditory system, linking auditory representations
 1132 S_V^{Ag} and S_C^{Ag} and corresponding syllables O_L^{Ag} , is learned by association, through interactions
 1133 with the Master Agent. More precisely, the term $P(S_V^{Ag} S_C^{Ag} | O_L^{Ag})$ consists of 9 auditory
 1134 prototypes (one for each value of O_L^{Ag}) encoded as 4-D Gaussian probability distributions on
 1135 the formant space ($F1_V, F2_V, F2_C, F3_C$), which the agent learns in a supervised manner from
 1136 the Master Agent. This provides <formant values, syllable label> pairs. Auditory recognition

1137 $P(O_L^{Ag} | S_V^{Ag} S_C^{Ag})$ is then implemented by the Bayesian inversion of $P(S_V^{Ag} S_C^{Ag} | O_L^{Ag})$:

$$1138 \quad P(O_L^{Ag} | S_V^{Ag} S_C^{Ag}) = \frac{P(S_V^{Ag} S_C^{Ag} | O_L^{Ag})}{\sum_o P(S_V^{Ag} S_C^{Ag} | [O_L^{Ag}=o])}$$

1139

1140 **L2: Learning the motor-to-auditory mapping by accommodation.**

1141 Since we attempt to learn the sensory-motor system independently of the motor
 1142 repertoire, learning is achieved by a variant of the learning by accommodation algorithm, in
 1143 which the Learning Agent only obtains auditory input from the Master Agent, without object
 1144 labels. Given a syllable acoustic target (s_V, s_C) , and using its current state of knowledge as
 1145 given by $P(S_V^{Ag} S_C^{Ag} | M_V^{Ag} M_C^{Ag})$, the Learning Agent carries out imitation tasks, by inferring
 1146 a motor gesture (m_V, m_C) likely to reach the target. This gesture is obtained by randomly
 1147 drawing a value (m_V, m_C) according to the inversion of the current forward model:

$$1148 \quad P(M_C^{Ag} M_V^{Ag} | [S_V^{Ag} = s_V][S_C^{Ag} = s_C]) \propto$$

$$1149 \quad P(M_V^{Ag}). P([S_V^{Ag} = s_V] | M_V^{Ag}). P(M_C^{Ag} | M_V^{Ag}). P([S_C^{Ag} = s_C] | M_C^{Ag}).$$

1150 The gesture (m_V, m_C) is sent to *VLAM*, which plays the role of an external vocal tract
 1151 simulator. *VLAM* outputs the formants (s_V^*, s_C^*) corresponding to the motor command
 1152 (m_V, m_C) , and the Learning Agent updates the knowledge stored in its internal models. It
 1153 observes that the chosen motor commands produce a given set of formants. This knowledge is
 1154 stored in the probability distributions $P(S_V^{Ag} | [M_V^{Ag} = m_V])$ and $P(S_C^{Ag} | [M_C^{Ag} = m_C])$,
 1155 which are Gaussian probability distributions evolving as their parameters become updated
 1156 through the learning process.

1157 The syllable targets provided by the Master Agent to the Learning Agent are taken
 1158 from the data presented in Figure 10. Since the Learning Agent initially has no knowledge
 1159 available, it selects motor gestures randomly. New observations lead to improving the quality
 1160 of the internal model of the motor-to-auditory transformation, which in turn improves the

1161 motor inversion that relies on this internal model. This means that the computed probability
 1162 distribution $P(M_C^{Ag} M_V^{Ag} | [S_V^{Ag} = s_V][S_C^{Ag} = s_C])$ driving the choice of motor gestures and
 1163 allowing imitation of auditory inputs becomes more and more accurate. Thus, the agent
 1164 becomes better and better at reaching its targets. All along the exploration process, the
 1165 learning algorithm remains driven by the targets provided by the Master, rather than by an
 1166 exhaustive sampling of the motor space as in other systems (e.g., Bailly, 1997; Guenther,
 1167 2006).

1168

1169 **L3: Learning the motor repertoire by imitation.**

1170 The motor system is learned in a supervised way, in that syllable labels are given to
 1171 the Learning Agent along with the corresponding stimuli. But while in other research the
 1172 articulatory data are provided (Castellini et al., 2011; Canevari et al., 2013), here the Learning
 1173 Agent is only given labeled acoustic data. We use the same <formant values, syllable label>
 1174 pairs that served to learn auditory categorization in step L1, and we use the internal model of
 1175 the motor-to-auditory transformation learned in step L2 to retrieve motor information. Given
 1176 an acoustic target (s_V, s_C) and the corresponding syllable label o_S , the Learning Agent infers a
 1177 motor gesture allowing the target to be reached by inverting the motor-to-auditory mapping
 1178 and by using its present state of knowledge of the correspondence between syllables and
 1179 motor gestures. This is done by randomly drawing $(m'_V, g'_C, \delta'_{MC})$ values according to the
 1180 following probability distribution:

$$P\left(\begin{array}{c} [M_V^{Ag} = m'_V][G'_C = g'_C] \\ [\Delta'_{MC} = \delta'_{MC}] \end{array} \middle| \begin{array}{c} [S_V^{Ag} = s_V][S_C^{Ag} = s_C][O_S^{Ag} = o_S] \\ [\lambda_{MV}^{Ag} = 1][\lambda_{MC}^{Ag} = 1] \end{array}\right)$$

$$\propto \left(\begin{array}{c} P([M_V^{Ag} = m'_V]) P([S_V^{Ag} = s_V] | [M_V^{Ag} = m'_V]) \\ P([M_C^{Ag} = m'_V + \delta'_{MC}] | [M_V^{Ag} = m'_V]) P([S_C^{Ag} = s_C] | [M_C^{Ag} = m'_V + \delta'_{MC}]) \\ P([M_V^{Ag} = m'_V][G'_C = g'_C][\Delta'_{MC} = \delta'_{MC}] | [O_S^{Ag} = o_S]) \end{array} \right)$$

1181 The correspondence between the chosen motor gesture $(m'_V, g'_C, \delta'_{MC})$ and the
 1182 syllable label o_S is then used to update parameters of the following probability elements: the
 1183 Gaussian probability distribution $P(M'_V^{Ag} | O_S^{Ag})$, the histogram $P(G'_C^{Ag} | O_S^{Ag})$ and the
 1184 Gaussian probability distribution $P(\Delta'_{MC} | M'_V^{Ag} G'_C^{Ag})$.

1185

1186 **Simulation results**

1187 **Confirming the “auditory-narrowband vs. motor-wideband” portrait in *COSMO-***

1188 **S.**

1189 The aim of the simulations described in this section is to verify how the main
 1190 principles we extracted from the results of the experiments carried out in the 1-D case are
 1191 generalized to the more realistic case of syllable processing. We ran a single learning
 1192 simulation with 4,000,000 <formant values, syllable label> pairs. The first 3,000,000 were
 1193 used during the L2 learning phase, with another 1,000,000 during L3 and the full set of the
 1194 same 4,000,000 values were also used for the L1 learning phase. These were resampled from
 1195 the data presented in Figure 10 and provided to the Learning Agent by the Master Agent.
 1196 Since the L1 learning phase of the sensory model is independent from the L2 and L3 learning
 1197 phases of the perceptuo-motor and motor models and for future comparisons of the auditory
 1198 and motor models of perception to be fair, the same data is used as input to learn the
 1199 components involved in motor and auditory perception.

1200 ***FIGURE 13 ABOUT HERE***

1201 Firstly, in Figure 13 we compare the evolution through learning of the entropy
 1202 $H(P(S^{Ag} | O_L^{Ag}))$ of the auditory model, with the evolution of the entropy $H(P(S^{Ag} | O_S^{Ag}))$
 1203 of the motor model. To this end, as in Figure 5 we used the entropy $H(P(S^{Ag} | O_S^{Master}))$ of
 1204 the probability distribution over the stimuli produced by the Master Agent, a constant over the
 1205 learning process, as a reference. As in the 1-D case, we further average these entropies over

1206 objects; since we now have 9 possible objects, we consider $\frac{1}{9} \sum_{O_L^{Ag}} H \left(P(S^{Ag} | O_L^{Ag}) \right)$ for the
1207 auditory model, $\frac{1}{9} \sum_{O_S^{Ag}} H \left(P(S^{Ag} | O_S^{Ag}) \right)$ for the motor model, and
1208 $\frac{1}{9} \sum_{O_S^{Master}} H \left(P(S^{Ag} | O_S^{Master}) \right)$ as the Master's reference.

1209 We observe that, as with the 1-D model, the entropy of the auditory model converges
1210 quickly to a level close to the entropy of the stimuli produced by the Master, whereas the
1211 entropy of the motor model converges more slowly⁽⁶⁾.

1212 ***FIGURE 14 ABOUT HERE***

1213 To better display how exploration and learning proceed in *COSMO-S*, we show in
1214 Figure 14 the motor gestures m'_V selected by the Learning Agent to attempt to reproduce the
1215 auditory targets provided by the Master Agent at five stages in the learning process: before
1216 any learning took place, during the L2 learning phase, between L2 and L3, during L3, and at
1217 the end of the learning process. We observe that learning enables progressive focusing of the
1218 vowel articulatory gestures around given areas in the articulatory space, corresponding to
1219 adequate commands for the three vowels /a, i, u/, but the size of the possible motor
1220 configurations remains wide.

1221 ***FIGURE 15 ABOUT HERE***

1222 Comparatively, we display in Figure 15 the actual productions of the Learning Agent
1223 in acoustic space at the same five steps in the learning process: the size of the available
1224 auditory space is more reduced around the three vowels /a i u/ provided by the Master Agent
1225 at the end of the learning process.

1226 ***FIGURE 16 ABOUT HERE***

1227 To evaluate the global categorization ability of the auditory, motor and perceptuo-
1228 motor branches in *COSMO-S* at the end of the learning process with 4,000,000 iterations, we
1229 exploited the same methodology as with the 1-D implementation of *COSMO*. We took as

1230 input the formant values ($F2, F3$ for the consonant, $F1, F2$ for the vowel) produced by the
1231 Master Agent (and displayed in Figure 10). We added a given level of noise by adding a
1232 Gaussian perturbation to each formant value, with a given variance indexed by the noise level.
1233 We present these stimuli to the auditory, motor, or perceptuo-motor classifier defined in
1234 *COSMO-S* according to the equations derived from Figure 2. In practice, we used exact
1235 inferences for evaluation: stimuli were not sampled from the Master Agent and then decoded,
1236 rather the whole probability distribution of stimuli was used to directly compute the resulting
1237 confusion matrices for each classifier. The original object O_S^{Master} was compared with the
1238 decoded object O_L^{Agent} or O_S^{Agent} depending on the model considered. Average diagonal
1239 values of the confusion matrices provide recognition scores that are displayed in Figure 16.

1240 The pattern of results is similar to that obtained with 1-D simulations. For clean
1241 stimuli (Figure 16, noise = 0), the auditory model is more accurate than the motor one. The
1242 difference is small, considering the large difference in entropies at the end of the learning
1243 process (see Figure 13), but this is because the distributions to categorize are well separated in
1244 these simulations. The difference would be larger in less clean learning configurations. When
1245 noise is added, the motor system performance decreases less rapidly than the auditory one,
1246 and it becomes more accurate for noise levels greater than 0.5. The perceptuo-motor model
1247 capitalizes on the fusion of the two branches to provide better scores than the separate
1248 auditory and motor models, at all noise levels.

1249 **Assessing auditory vs. motor invariance for the place of articulation of vowels**
1250 **and plosives in *COSMO-S*.**

1251 We now explore our second proposal about auditory-motor complementarity,
1252 assessing how phonemic invariance could be represented in the auditory or motor branches in
1253 *COSMO-S*.

1254 The situation for vowels has already been presented in Figures 14 and 15. It is in

1255 agreement with our predictions; while the acoustic characterization of vowels is rather
1256 straightforward (see the final learning stage in Figure 15), the distribution of motor variables
1257 is more disordered (see the final learning stage in Figure 14).

1258 ***FIGURE 17 ABOUT HERE***

1259 For plosives, the acoustic configuration is more complicated than for vowels. Indeed,
1260 Figure 10 shows how intricate the formant configurations are for each plosive, due to vowel
1261 co-articulation. This is where the motor system could play a crucial role. Indeed, in Figure 17
1262 we display the evolution of the motor variable $P(G'_c{}^{Ag} | O_s^{Ag})$ distribution for the 9 objects
1263 O_s^{Ag} . Each subplot displays the evolution of the probabilities of the three possible gestures
1264 (*LipHeight*, *TongueDorsum* and *TongueApex*) for each object at each learning stage. It
1265 appears that for 8 cases out of 9, the identification of the correct gesture has been successful:
1266 *LipHeight* for /ba, bi, bu/, *TongueDorsum* for /ga, gi, gu/ and *TongueApex* for /da, di, du/.
1267 This means that the Learning Agent has selected a gesture compatible with that performed by
1268 the Master Agent, and that motor invariance is within reach through the existence of the $G'_c{}^{Ag}$
1269 parameter in the motor repertoire.

1270 However, there is one case where adequate identification has partly failed: for the
1271 object /gi/, the probability of the *TongueApex* gesture remains high, even though this is not the
1272 adequate gesture for the velar in /gi/. The reason is clear: looking at Figure 10, it can be seen
1273 that the acoustic regions for /di/ and for /gi/ are partially superimposed. This is probably due
1274 to an acoustic description of the plosives that is too simplified (e.g. lacking higher formants,
1275 burst fine characteristics or spectral dynamics, which could all play a part in improving the
1276 /di/-/gi/ contrast). However, even if there was an articulatory ambiguity, we may suppose that
1277 hyper-articulation by the Master Agent could guide the Learning Agent to solve the problem.
1278 In a further simulation, we implemented this process by having a Master Agent discarding
1279 productions before reaching an acoustic zone where both /di/ and /gi/ could be produced. With

1280 such a dataset for /gi/ production, the recovery of adequate gestures is perfect, as displayed in
1281 Figure 18.

1282 ***FIGURE 18 ABOUT HERE***

1283

1284 **Concluding Part 3: Summary and predictions**

1285 The simulations with COSMO-S enable a significant gap in complexity to be crossed
1286 and the possibility of implementing and testing the *COSMO* model in a high dimension (8
1287 articulatory + 4 acoustic parameters) could be assessed. These provide two major results.

1288 Firstly, we confirmed the auditory-narrow motor-wide portrait introduced in Part 2.
1289 Once again we obtain both quicker and more efficient learning of acoustic stimuli in the
1290 sensory compared with the motor pathway (Figures 13-15), resulting in better auditory
1291 recognition scores than motor ones without noise, but also a superiority of the motor decoding
1292 process in noisy conditions (Figure 16). In consequence, altogether the perceptuo-motor
1293 model performs better than both the auditory and the motor pathways whatever the noise level
1294 (Figure 16). This was expected, given that the auditory-narrow motor-wide hypothesis is
1295 considered to be generic and independent of the underlying specific implementation.

1296 However, it is of interest to confirm that it is displayed in a much more complex and realistic
1297 sensory-motor environment in COSMO-S compared with the 1-D simulations of Part 2.

1298 Interestingly, the motor-to-sensory transformation in COSMO-S, associated with VLAM, is
1299 no longer monotonous. Altogether, and unsurprisingly, the increase in complexity even results
1300 in a much larger difference in learning speed and efficiency between auditory and motor
1301 inference in COSMO-S compared with 1-D simulations (compare Figures 5 and 13).

1302 Secondly, the analysis of the information content of auditory vs. motor representations
1303 lets a natural complementarity appear, with plosives on one side, difficult to characterize in
1304 the auditory space, but clearly associated with a motor or gestural invariant in the motor

1305 repertoire, and vowels on the other side, for which the auditory characterization is more
1306 efficient than the motor one.

1307 The potential limitations of this study are related to the nature of the hypotheses we
1308 introduced to make the implementation of tractable simulations possible. Firstly, we had to
1309 base our work on artificial synthetic CV sequences. Indeed, *COSMO* requires a sensory-motor
1310 model able to process stimuli characterized in the motor (articulatory) and sensory (acoustic)
1311 spaces. No currently available speech production model can produce completely realistic
1312 speech stimuli. Thus we needed to restrict our simulations to synthetic stimuli generated by
1313 the model we had at our disposal, i.e. *VLAM*. However, the realism of formant data in our
1314 simulations, and the relative complexity of the material that we provided for processing in
1315 *COSMO-S*, make us confident that real speech is not out of reach of *COSMO* development.

1316 A second hypothesis in *COSMO-S* is the assumption that the Learning Agent has at its
1317 disposal a set of primitive coordination gestures, i.e. *Jaw/Lips*, *Jaw/Tongue Apex* or
1318 *Jaw/Tongue Dorsum*, corresponding respectively to labial, alveolar and velar places of
1319 articulation. This hypothesis deserves further comment. It is consistent with data on infant
1320 imitation showing that infants have at their disposal basic facial gestures that they can identify
1321 from birth on the face of their communicating partner (Meltzoff & Moore, 1977). The
1322 Articulatory Organ Hypothesis developed in the Haskins Labs at the beginning of the 2000s
1323 (see e.g., Best & McRoberts, 2003; Goldstein & Fowler, 2003) exploits precisely this kind of
1324 assumption to describe perception and control in the framework of Articulatory Phonology. It
1325 supposes that infants are able to detect in a speech signal the primary articulatory organ that
1326 produced it. Current simulations provide this hypothesis with some computational basis.

1327 These two results, “auditory-narrow, motor-wide” and “auditory-vowel, motor-
1328 plosive” provide two major sets of experimental predictions.

1329

1330 **Prediction 3 - early speech perception should be mostly auditory before the onset**
1331 **of babbling, then become progressively perceptuo-motor**

1332 The PACT proposal is that speech perception should make use of only the auditory
1333 pathway in the first months of age, then progressively capitalize on feedback from the motor
1334 system when it is mature and mainly since babbling onset around 7 months. This appears to
1335 be reinforced in the context of the auditory-narrow, motor-wide hypothesis. The learning
1336 pattern in Figure 13 strongly confirms the view that auditory perception should be mature
1337 long before motor information could be used for phonetic decoding. Considering the potential
1338 role of the motor system for perception in noisy conditions, it could be suggested that as this
1339 is still immature at the first months of age, it should not intervene specifically in adverse
1340 conditions, before some significant degree of sensory-motor development. Basically, it is after
1341 babbling onset that infants can obtain a useful amount of information on the motor inference
1342 branch.

1343 This is exactly what was described in a recent MEG (Magnetoencephalography) study
1344 on infants' brain responses to native vs. non-native stimuli at two developmental stages in the
1345 first year of age (Kuhl et al., 2014). Indeed, the data in this study showed that infants at 7
1346 months of age do not display a significantly different involvement of the motor regions
1347 (including Broca's area and the cerebellum) for native vs. non-native speech. In contrast, at
1348 11-12 months of age, i.e. after a significant amount of perceptuo-motor learning has occurred,
1349 following babbling onset at around 7 months, there is more involvement of motor regions for
1350 non-native compared with native stimuli.

1351

1352 **Prediction 4 - plosive place of articulation invariance should require motor**
1353 **knowledge**

1354 The last prediction is related to the second result obtained with COSMO-S about the

1355 “auditory-vowel, motor-plosive” hypothesis. The corresponding results in COSMO-S suggest
1356 that the identification of invariant cues for the plosive place of articulation should strongly
1357 depend on the acquisition of motor representations associated with sensory input in the motor
1358 inference process.

1359 As mentioned in the Introduction, a number of experimental data do indeed suggest
1360 that infants cannot detect the plosive place of articulation invariance before 6 months of age.
1361 A recent study by Hochmann & Papeo (2014) exploiting a novel methodology based on
1362 pupillometry provided a hint that the “b” vs. “d” contrast could be displayed independently in
1363 vowel context at 6 months. However, auditory and visual information could be at the basis of
1364 this result (Dole et al., 2015). Importantly, another recent study in our group, exploiting an
1365 inter-sensory matching procedure, provided different results compatible with the present
1366 prediction. This procedure provided no evidence for articulation plosive identification
1367 independent of vowel context at 6 months of age, but some such evidence was seen at 9
1368 months. Importantly, infants’ perceptual abilities appeared to be related to their motor abilities
1369 in babbling (Dole, Loevenbruck, Pascalis, Schwartz, & Vilain, 2016).

1370 The present prediction should not be generalized to the proposal that there would be
1371 no involvement at all of the motor system in speech perception before babbling onset. Indeed,
1372 Bruderer, Danielson, Kandhadai & Werker (2015) demonstrated that teething displays used to
1373 control infants' tongues in their mouths may interfere with the perception of non-native
1374 stimuli related to the corresponding induced tongue shapes for 6 month old subjects. The
1375 important point of our prediction is that the plosive place of articulation requires learning the
1376 sensory-to-motor correspondence in complete CV sequences that are out of reach before the
1377 onset of babbling.

1378 Finally, it is of interest that a recent analysis of fMRI responses to CV syllables using
1379 multivariate decoding shows that plosive place of articulation is indeed specifically found to

1380 be represented in regions of the brain associated with speech production, including the
1381 posterior ventral frontal cortex, the basal ganglia, and the cerebellum (Correia, Jansma &
1382 Bonte, 2015).

1383

1384

1385 **Part 4 - Three challenges for a perceptuo-motor theory of speech perception**

1386 At the end of this research, we have at our disposal the first Bayesian Perceptuo-Motor
1387 model of speech perception, COSMO, together with various implementations (from 1-D to
1388 COSMO-S). Furthermore, we have two major results about “non-perfect” learning conditions,
1389 enabling to depart from the indistinguishability theorem: the “auditory-narrow, motor-wide”
1390 and “auditory-vowel, motor-plosive” properties. This model opens a number of perspectives
1391 for future developments. We will discuss three major directions for research in the field of
1392 perceptuo-motor interactions involving potential developments in COSMO.

1393

1394 **Challenge 1: Perceptuo-motor complementarity and Perceptuo-motor fusion**

1395 The Introduction showed how publications in the field shifted from almost purely
1396 functional arguments about the auditory vs. motor controversy, somewhat lacking of
1397 experimental data, to convincing experimental neurocognitive data supporting the role of the
1398 motor system, but somewhat lacking of functionalist views about why the motor system could
1399 be useful at all. The present study attempted to provide such functionalist arguments. Future
1400 studies should attempt to provide more data about when, how and why the motor system
1401 could enhance auditory perception. Furthermore, a perceptuo-motor theory of speech
1402 perception requires a fusion process enabling efficient combination of auditory (if not visual
1403 or somatosensory) and motor information for speech decoding.

1404 Interestingly, audiovisual speech perception research has asked more or less the same
1405 questions for about the last forty years. It was shown how auditory and visual inputs could be
1406 complementary to a certain extent (e.g. Summerfield, 1987; Robert-Ribes, Schwartz,
1407 Lallouache, & Escudier, 1998). Audio-visual fusion led to many theoretical and
1408 methodological developments, proposing that it could be optimal in the Bayesian sense (see

1409 Massaro and the Fuzzy-Logical Model of Perception, 1987, 1998; in relation with Ernst &
1410 Banks, 2002), and that within fusion each sensory modality could possibly be weighted
1411 according to its reliability, depending on context, language, subjects, etc. (Schwartz, 2010).

1412 Perceptuo-motor complementarity and fusion should thus be set at a high position in
1413 the research agenda on perceptuo-motor speech perception, just as they were in past research
1414 on audiovisual speech perception.

1415

1416 **Challenge 2 – Integrating speech perception and speech production in a common**
1417 **framework**

1418 Accumulating evidence for the role of motor knowledge in speech perception may be
1419 combined with accumulating evidence for the role of perceptual representations and processes
1420 in speech motor control (see reviews in Guenther, Hampson & Johnson, 1998; Perrier, 2005).
1421 Importantly, the current perceptuo-motor model of speech perception capitalizes on a set of
1422 computational bricks traditionally involved in speech production models. Thus, sensory and
1423 motor representations are associated thanks to internal forward or inverse models (e.g.
1424 Guenther, Ghosh & Tourville, 2006; Houde & Nagarajan, 2011; Hickok, 2012; Patri, Diard &
1425 Perrier, 2015).

1426 This suggests that it could be possible to develop an integrated framework associating
1427 speech perception and speech production models within the same theoretical architecture.

1428 This is one aim of the COSMO architecture (see Moulin-Frier et al., 2012, 2015).

1429 Interestingly, the same objective has been introduced in recent phonological models (see e.g.
1430 Boersma, 2011, Boersma & Hamman, 2008).

1431

1432 **Challenge 3 – From computational architecture to neurocognitive implementation**

1433 It is widely acknowledged, at least since Marr (1982), that cognitive systems can be

1434 analyzed at different levels, three in Marr's proposal: computational, algorithmic, and
1435 representational and implementation levels. These levels are independent to a certain extent,
1436 but the computational and algorithmic architectures may shed light on the way neurocognitive
1437 implementation could be realized. Conversely, neurocognitive constraints could suggest some
1438 proposals for algorithmic considerations.

1439 At this stage, we did not elaborate in any way the possible neurocognitive means by
1440 which the various components in COSMO could be implemented in the human brain. This is
1441 not to say that such an enterprise, relating computation and implementation levels, is out of
1442 reach, as was clearly displayed by the authors of the DIVA model of speech production
1443 (Guenther et al., 2006). Considering the increasing amount of details provided by
1444 neuroscience about the neural coding of speech perception and production in the auditory and
1445 motor cortex (see e.g. Bouchard, Mesgarani, Johnson & Chang, 2013; Cheung, Hamiton,
1446 Johnson, & Chang, 2016; Formisano, De, Bonte, & Goebel, 2008; Pasley et al., 2012), it is
1447 now a challenging but intriguing and probably necessary enterprise to attempt to elaborate
1448 further the possible relationships between computational models such as COSMO and neural
1449 responses in a number of experimental tasks.

1450

1451

Conclusion

1452 This paper develops an original perspective in the debate between auditory and motor
1453 theories of speech perception. Research in the cognitive neurosciences led to the now well-
1454 accepted views that (i) motor areas are activated during speech perception, and (ii) motor
1455 knowledge seems to play a certain role in speech perceptual processing in the human brain.
1456 From these points of view, we attempted to evaluate the precise functional role of motor
1457 knowledge. In the framework of PACT, a perceptuo-motor theory of speech perception, we
1458 explored these questions in computational terms, thanks to COSMO, the first Bayesian

1459 perceptuo-motor model of speech communication.

1460 We showed here for the first time that, in conditions that are perfect in a certain sense,
1461 the information content of the auditory and motor branches of a perceptuo-motor speech
1462 processing system are exactly the same. We introduced realistic learning conditions and
1463 showed that they let a natural complementarity emerge between a “narrow-band” auditory
1464 system that is more efficient in good communication conditions, and a “wide-band” motor
1465 system that is more efficient in adverse conditions. Our simulations also suggest that
1466 invariants providing the phonetic characterization of phonological units could be perceptuo-
1467 motor rather than auditory or motor, and show how this could be achieved, with auditory cues
1468 for vowels and motor cues for the plosive place of articulation.

1469 COSMO simulations lead to a number of experimental predictions. Some of these are
1470 already being tested, with data in agreement with predictions. Others require more
1471 experimental efforts. COSMO also opens a number of perspectives in domains such as: the
1472 fusion of perceptual and motor inference in phonetic decoding; the co-development of
1473 computational models of speech production and speech perception; the possibility to apply
1474 COSMO simulations to a number of neurocognitive data on the coding and processing of
1475 speech in the human brain.

1476 This research has placed computational simulations at the heart of the debate about the
1477 role of perceptual and motor knowledge in the speech perception process. Considering the
1478 rapidly increasing amount of experimental evidence and data available about the perceptuo-
1479 motor relationship in speech communication, it seems that mathematical models can be of
1480 great help in clarifying arguments, precisising mechanisms and suggesting new predictions and
1481 experimental paradigms. Perceptuo-motor complementarity, invariance, fusion and
1482 development are crucial steps in the agenda of future research into the cognitive bases of
1483 speech communication. The first pieces in the elaboration of the *COSMO* model described

1484 and discussed in the present paper provide convincing elements for pursuing this direction.

- 1510 Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An
1511 investigation of young infants' perceptual representations of speech sounds. *Journal of*
1512 *Experimental Psychology: General*, 117, 21–33. <http://dx.doi.org/10.1037/0096-3445.117.1.21>
- 1513 Bessière, P., Laugier, C., & Siegwart, R. (2008). *Probabilistic reasoning and decision making in*
1514 *sensory-motor systems*. Springer Tracts in Advanced Robotics. Berlin: Springer-Verlag.
- 1515 Bessière, P., Mazer, E., Ahuactzin-Larios, J.-M., & Mekhnacha, K. (2013). *Bayesian Programming*.
1516 CRC Press. Boca Raton, FL.
- 1517 Best, C.C., & McRoberts, G.W. (2003). Infant Perception of Non-Native Consonant Contrasts that
1518 Adults Assimilate in Different Ways. *Language and Speech*, 46, 183–216.
1519 <http://dx.doi.org/doi:10.1177/00238309030460020701>
- 1520 Bever, T.G., & Poeppel, D. (2010). Analysis by Synthesis: A (Re-)Emerging Program of Research for
1521 Language and Vision. *Biolinguistics*, 4.2-3, 174–200. <http://www.biolinguistics.eu>
- 1522 Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., & Ward, B.D. (2004). Neural correlates of
1523 sensory and decision processes in auditory object identification. *Nature Neuroscience*, 7, 295–
1524 301. <http://dx.doi.org/doi:10.1038/nn1198>
- 1525 Blumstein, S.E., & Stevens, K.N. (1979). Acoustic invariance in speech production: Evidence from
1526 measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical*
1527 *Society of America*, 66, 1001–1017. <http://dx.doi.org/10.1121/1.383319>
- 1528 Boë, L.-J., Badin, P., Ménard, L., Captier, G., Davis, B., MacNeilage, P., Sawallis, T., & Schwartz, J.-
1529 L. (2013). Anatomy and control of the developing human vocal tract: A response to
1530 Lieberman. *Journal of Phonetics*, 41, 379–392.
- 1531 Boë, L.-J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel
1532 production: Proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*,
1533 20, 27–38.
- 1534 de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer

- 1535 model. *Acoustics Research Letters Online*, 4, 129–134. <http://dx.doi.org/10.1121/1.1613311>
- 1536 Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and
1537 evolution. In Anton Benz & Jason Mattausch (Eds.) *Bidirectional Optimality Theory*, pp. 33–
1538 72. Amsterdam: John Benjamins.
- 1539 Boersma, P., & Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint
1540 grammar. *Phonology*, 25, 217–270. <http://dx.doi.org/10.1017/S0952675708001474>
- 1541 Bouchard, K.E., Mesgarani, N., Johnson, K., & Chang E.F. (2013). Functional organization of human
1542 sensorimotor cortex for speech articulation. *Nature*, 495, 327–332.
1543 <http://dx.doi.org/10.1038/nature11911>
- 1544 de Boysson-Bardies, B. (1993). Ontogeny of language-specific syllabic production. In B. de Boysson-
1545 Bardies & S. de Schoen & P. Jusczyk & P. F. MacNeilage & J. Morton (Eds.), *Developmental*
1546 *neurocognition: Speech and face processing in the first year of life* (pp. 353-363). Dordrecht:
1547 Kluwer Academic Publishers.
- 1548 de Boysson-Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of
1549 vowel formants in babbling. *Journal of Child Language*, 16, 1-17.
- 1550 de Boysson-Bardies, B., Sagart, L., & Durant, C. (1984). Discernible differences in the babbling of
1551 infants according to target language. *Journal of Child Language*, 11, 1-15.
- 1552 Bruderer, A.G., Danielson, D.K., Kandhadai, P., & Werker, J.F. (2015). Sensorimotor influence on
1553 speech perception in infancy. *Proc Natl Acad Sc*, 112, 13531-13536.
1554 <http://dx.doi.org/10.1073/pnas.1508631112>
- 1555 Callan, D.E., Callan, A.M., & Jones J.A. (2014). Speech motor brain regions are differentially
1556 recruited during perception of native and foreign-accented phonemes for first and second
1557 language listeners. *Frontiers in Neuroscience*, 03 September 2014,
1558 <http://dx.doi.org/10.3389/fnins.2014.00275>
- 1559 Callan, D.E., Jones, J.A., Callan, A.M., & Akahane-Yamada, R. (2004). Phonetic perceptual

- 1560 identification by native- and second-language speakers differentially activates brain regions
1561 involved with acoustic phonetic processing and those involved with articulatory-
1562 auditory/orosensory internal models. *NeuroImage*, 22, 1182–1194,
1563 <http://dx.doi.org/doi:10.1016/j.neuroimage.2004.03.006>
- 1564 Canevari, C., Badino, L., D'Ausilio, A., Fadiga, L., & Metta, G. (2013). Modeling speech imitation
1565 and ecological learning of auditory-motor maps. *Frontiers in Psychology*, 4, 364,
1566 <http://dx.doi.org/10.3389/fpsyg.2013.00364>
- 1567 Castellini, C., Badino, L., Metta, G., Sandini, G., Tavella, M., Grimaldi, M., & Fadiga, L. (2011). The
1568 use of phonetic motor invariants can improve automatic phoneme discrimination. *PLoS ONE*,
1569 6, e24055. <http://dx.doi.org/10.1371/journal.pone.0024055>
- 1570 Cheung, C., Hamiton, L.S., Johnson, K., & Chang, E. F. (2016). The auditory representation of
1571 speech sounds in human motor cortex. *eLife*, 5:e12577. DOI: 10.7554/eLife.12577
- 1572 Clayards, M., Aslin, R. N., Tanenhaus, M. K., & Jacobs, R. A. (2007). Within category phonetic
1573 variability affects perceptual uncertainty. In *Proc. 16th International Congress of Phonetic*
1574 *Sciences*, Saarbrücken, Germany, pages 701–704.
- 1575 Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects
1576 optimal use of probabilistic speech cues. *Cognition*, 108, 804–809.
1577 <http://dx.doi.org/10.1016/j.cognition.2008.04.004>
- 1578 Correia, J.L., Jansma, B.M.B., & Bonte, M. (2015). Decoding articulatory features from fMRI
1579 responses in dorsal speech regions. *The Journal of Neuroscience*, 35, 15015–15025.
1580 <http://dx.doi.org/10.1523/JNEUROSCI.0977-15.2015>
- 1581 Davis, B.L., MacNeilage, P., & Matyear, C.L. (2002). Acquisition of Serial Complexity in Speech
1582 Production: A Comparison of Phonetic and Phonological Approaches to First Word
1583 Production. *Phonetica*, 59, 75–107. <http://dx.doi.org/10.1159/000066065>
- 1584 Dayan, P., & Abbott, L. (2001). *Theoretical Neuroscience*. The MIT Press, Cambridge, MA.

- 1585 DeCasper, A.J., & Fifer, W.P. (1980). Of human bonding: newborns prefer their mother's voice.
1586 *Science*, 208, 1174-1176.
- 1587 Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional Neuroimaging of
1588 Speech Perception in Infants. *Science*, 298, 2013–2015.
1589 <http://dx.doi.org/10.1126/science.1077066>
- 1590 Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., &
1591 Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of
1592 sentences in preverbal infants. *Proceedings of the National Academy of Sciences*, 103, 14240–
1593 14245, <http://dx.doi.org/10.1073/pnas.0606302103>
- 1594 Delattre, P. C., Liberman, A. M. & Cooper, F. S. (1955) Acoustic loci and transitional cues for
1595 consonants. *Journal of the Acoustical Society of America*, 27, 769–73.
1596 <http://dx.doi.org/10.1121/1.1908024>
- 1597 Deng, L. (1999). Computational models for auditory speech processing,. In K. Ponting (Ed.)
1598 *Computational Models for Speech Pattern Processing*, pp. 67-77. New York: Springer-Verlag,
1599 NATO ASI.
- 1600 Deng, L. & Ma, J. (2000). Spontaneous speech recognition using a statistical coarticulatory model for
1601 the vocal- tract-resonance dynamics. *The Journal of the Acoustical Society of America*, 108,
1602 3036–3048. <http://dx.doi.org/10.1121/1.1315288>
- 1603 Deng, L., Ramsay, G., & Sun, D. (1997). Production models as a structural basis for automatic speech
1604 recognition. *Speech Communication*, 22, 93–111. [http://dx.doi.org/10.1016/S0167-
1605 6393\(97\)00018-6](http://dx.doi.org/10.1016/S0167-6393(97)00018-6)
- 1606 Diehl, R., Lotto, A., & Holt, L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149–
1607 179. <http://dx.doi.org/10.1146/annurev.psych.55.090902.142028>
- 1608 Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological
1609 categories: Insights from Inuktitut. *Cognitive Science*, 37, 344–377.

- 1610 <http://dx.doi.org/10.1111/cogs.12008>
- 1611 Dole, M., Loevenbruck, H., Pascalis, O., Schwartz, J.-L., & Vilain, A. (2015). Perceptual abilities in
1612 relation with motor development in the first year of life. *WILD 2015 International Conference*,
1613 Stockholm.
- 1614 Dole, M., Loevenbruck, H., Pascalis, O., Schwartz, J.L, Vilain, A. (2016). Phoneme categorization
1615 depends on production abilities during the first year of life. XX ICIS 2016 - The International
1616 Congress on Infant Studies.
- 1617 Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young
1618 infants. *The Journal of the Acoustical Society of America*, *105*,1901–1911.
1619 <http://dx.doi.org/10.1121/1.426726>
- 1620 Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants.
1621 *Science*, *171*, 303–306. <http://dx.doi.org/10.1126/science.171.3968.303>
- 1622 Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically
1623 optimal fashion. *Nature*, *415*, 429–433. <http://dx.doi.org/10.1038/415429a>
- 1624 Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically
1625 modulates the excitability of tongue muscles: A TMS study. *European Journal of*
1626 *Neuroscience*, *15*, 399–402. <http://dx.doi.org/10.1046/j.0953-816x.2001.01874.x>
- 1627 Feldman, N. H., & Griffiths, T. L. (2007). A rational account of the perceptual magnet effect. In
1628 *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pp. 257–262.
- 1629 Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009a). The influence of categories on perception:
1630 Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*,
1631 *116*, 752–782. <http://dx.doi.org/10.1037/a0017196>
- 1632 Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009b). Learning phonetic categories by learning a
1633 lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp.
1634 2208–2213.

- 1635 Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing
1636 lexicon in phonetic category acquisition. *Psychological Review*, *120*, 751–778.
1637 <http://dx.doi.org/10.1037/a0034245>
- 1638 Formisano, E., De, M.F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? brain-based
1639 decoding of human voice and speech. *Science*, *322*, 970–973.
1640 <http://dx.doi.org/10.1126/science.1164318>
- 1641 Fowler, C., & Dekle, D. (1991). Listening with eye and hand: crossmodal contributions to speech
1642 perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*,
1643 816–828. <http://dx.doi.org/doi:10.1037/0096-1523.17.3.816>
- 1644 Frankel, J., Richmond, K., King, S., et Taylor, P. (2000). An automatic speech recognition system
1645 using neural networks and linear dynamic models to recover and model articulatory traces. In
1646 *Proc. Sixth International Conference on Spoken Language Processing*, Vol. 4. International
1647 Speech Communication Association.
- 1648 Galantucci, B., Fowler, C.A., & Turvey, M.T. (2006). The motor theory of speech perception
1649 reviewed. *Psychonomic Bulletin & Review*, *13*, 361–377.
1650 <http://dx.doi.org/10.3758/BF03193857>
- 1651 Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model
1652 of speech perception. *Language and Cognitive Processes*, *12*, 613–656.
- 1653 Gilet, E., Diard, J., & Bessière, P. (2011). Bayesian action–perception computational model:
1654 interaction of production and recognition of cursive letters. *PLoS ONE*, *6*, e20387.
1655 <http://dx.doi.org/10.1371/journal.pone.0020387>
- 1656 Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use.
1657 In N. O. Schiller & A. Meyer (Eds.), *Phonetics and phonology in language comprehension*
1658 *and production: Differences and similarities* (pp. 159–207). Berlin: Mouton de Gruyter.
- 1659 Grabski, K., Tremblay, P., Gracco, V.L., Girin, L., & Sato, M. (2013). A mediating role of the

- 1660 auditory dorsal pathway in selective adaptation to speech: A state-dependent transcranial
1661 magnetic stimulation study. *Brain Research*, 1515, 55–65.
1662 <http://dx.doi.org/10.1016/j.brainres.2013.03.024>
- 1663 Grafton, S.T., Arbib M.A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations
1664 in humans by positron emission tomography. 2. Observation compared with imagination.
1665 *Experimental Brain Research*, 112, 103–11. <http://dx.doi.org/10.1007/BF00227183>
- 1666 Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley & Sons,
1667 Ltd, New York, NY, USA.
- 1668 Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of*
1669 *Communication Disorders*, 39, 350–365. <http://dx.doi.org/10.1016/j.jcomdis.2006.06.013>
- 1670 Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical
1671 interactions underlying syllable production. *Brain and Language*, 96, 280-301.
1672 <http://dx.doi.org/10.1016/j.bandl.2005.06.001>
- 1673 Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames
1674 for the planning of speech movements. *Psychological Review*, 105, 611–633.
1675 <http://dx.doi.org/10.1037/0033-295X.105.4.611-633>
- 1676 Halle, M., & Stevens, K. N. (1959). Analysis by synthesis. In W. Wathen-Dunn & L. E. Woods
1677 (Eds.), *Proceedings of the seminar on speech compression and processing*. USAF Camb. Res.
1678 Ctr. 2: Paper D7.
- 1679 Hermansky, H. (1998). Should recognizers have ears? *Speech Communication*, 25, 3-27.
- 1680 Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews*
1681 *Neuroscience*. 13, 135-145. doi:10.1038/nrn3158
- 1682 Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., A. Senior, Vanhoucke, V., Nguyen,
1683 P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech
1684 recognition: The shared views of four research groups. *IEEE Signal Process. Magazine*, 29,

- 1685 82-97.
- 1686 Hochmann, J.R., & Papeo, L. (2014). The Invariance Problem in Infancy: A Pupillometry Study.
1687 *Psychological Science*, 25, 2038–2046. <http://dx.doi.org/10.1177/0956797614547918>
- 1688 Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in*
1689 *Human Neuroscience*, 5. doi:10.3389/fnhum.2011.00082
- 1690 Huang, X., & Deng, L. (2010). An Overview of Modern Speech Recognition. , in R. Herbrich & T.
1691 Graepel (Eds.) *Handbook of Natural Language Processing*, Second Edition, pp. 339-366.
1692 Chapman & Hall/CRC.
- 1693 Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C., & Rizzolatti, G. (1999).
1694 Cortical Mechanisms of Human Imitation. *Science*, 286, 2526–2528.
1695 <http://dx.doi.org/10.1126/science.286.5449.2526>.
- 1696 Imada T, Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P.K. (2006) Infant speech
1697 perception activates Broca’s area: A developmental magnetoencephalography study.
1698 *NeuroReport*, 17, 957–962.
- 1699 Ito, T., Tiede, M., & Ostry, D.J. (2009). Somatosensory function in speech perception. *Proceedings of*
1700 *the National Academy of Sciences*, 106, 1245–1248.
1701 <http://dx.doi.org/10.1073/pnas.0810063106>.
- 1702 Jacquemot, C., Dupoux, E., & Bachoud-Lévi, A.-C. (2007). Breaking the mirror: asymmetrical
1703 disconnection between the phonological input and output codes. *Cognitive Neuropsychology*,
1704 24, 3–22. <http://dx.doi.org/10.1080/02643290600683342>
- 1705 Jones, J.A., & Callan, D.E. (2003). Brain activity during audiovisual speech perception: An fMRI
1706 study of the McGurk effect. *NeuroReport*, 14, 1129–1133.
1707 <http://dx.doi.org/10.1097/00001756-200306110-00006>
- 1708 Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants.
1709 *Developmental Psychology*, 23, 648–654. <http://dx.doi.org/10.1037/0012-1649.23.5.648>

- 1710 Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop
1711 consonants. *Journal of the Acoustical Society of America*, 73, 322–335.
1712 <http://dx.doi.org/10.1121/1.388813>
- 1713 Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2009). Perception and hierarchical dynamics. *Frontiers*
1714 *in Neuroinformatics*, 3, 20. <http://dx.doi.org/10.3389/neuro.11.020.2009>
- 1715 Kingston, J., & Diehl, R.L. (1994). Phonetic knowledge. *Language*, 70, 419–54.
1716 <http://dx.doi.org/10.1353/lan.1994.0023>.
- 1717 Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration
1718 and selective adaptation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and*
1719 *Computational Linguistics*, pages 10–19. Association for Computational Linguistics.
- 1720 Kleinschmidt, D. & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar,
1721 generalizing to the similar, and adapting to the novel. *Psychological Review*, 122, 148–203.
1722 <http://dx.doi.org/10.1037/a0038695>
- 1723 Kluender KR. (1994). Speech perception as a tractable problem in cognitive science. In M.A.
1724 Gernsbacher (Ed.) *Handbook of Psycholinguistics* (pp. 173–217). San Diego, CA: Academic.
- 1725 Kröger, B.J., Kannampuzha, J., & Kaufmann, E. (2014). Associative learning and self-organization as
1726 basic principles for simulating speech acquisition, speech production, and speech perception.
1727 *EPJ Nonlinear Biomedical Physics*, 2. <http://dx.doi.org/10.1140/epjnbp15>
- 1728 Kröger, B.J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational
1729 model of speech production and perception. *Speech Communication*, 51, 793–809.
1730 <http://dx.doi.org/10.1016/j.specom.2008.08.002>
- 1731 Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar
1732 vowel categories. *Journal of the Acoustical Society of America*, 66, 1668–1679.
- 1733 Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant*
1734 *Behavior & Development*, 6, 263–285.

- 1735 Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews*
1736 *Neuroscience*, 5, 831–843. <http://dx.doi.org/10.1038/nrn1533>
- 1737 Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, T., Rivera-Gaxiola, M., & Nelson, T. (2008).
1738 Phonetic learning as a pathway to language: New data and native language magnet theory
1739 expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*,
1740 363, 979–1000. <http://dx.doi.org/10.1098/rstb.2007.2154>
- 1741 Kuhl, P. K., & Meltzoff (1996). A. Infant vocalizations in response to speech: vocal imitation and
1742 developmental change. *Journal of the Acoustical Society of America*. 100, 2425–2438.
- 1743 Kuhl, P.K., Ramírez, R.R., Bosseler, A., Lotus Lin, J.F., & Imada, T. (2014). Infants' brain responses
1744 to speech suggest Analysis by Synthesis. *Proceedings of the National Academy of Sciences*,
1745 111, 12572–12573. <http://dx.doi.org/10.1073/pnas.1410963111>
- 1746 Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. & Lindblom, B. (1992). Linguistic
1747 experience alters phonetic perception in infants by 6 months of age. *Science*, 255, 606–608.
1748 <http://dx.doi.org/10.1126/science.1736364>
- 1749 Laurent, R., Schwartz, J.-L., Bessière, P., & Diard, J. (2013). A computational model of perceptuo-
1750 motor processing in speech perception: learning to imitate and categorize synthetic CV
1751 syllables. In *Proceedings of InterSpeech* (pp. 2797–2801). Lyon, France.
- 1752 Lebeltel, O., Bessière, P., Diard, J., & Mazer, E. (2004). Bayesian robot programming. *Autonomous*
1753 *Robots*, 16, 49–79. <http://dx.doi.org/10.1023/B:AURO.0000008671.38949.43>
- 1754 Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of
1755 speech code. *Psychological Review*, 74, 431–461. <http://dx.doi.org/10.1037/h0020279>
- 1756 Liberman, A.M., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*,
1757 21, 1–36. [http://dx.doi.org/10.1016/0010-0277\(85\)90021-6](http://dx.doi.org/10.1016/0010-0277(85)90021-6)
- 1758 Liberman, A.M., & Mattingly, I.G. (1989). A specialization for speech perception. *Science*, 243, 489–
1759 494. <http://dx.doi.org/10.1126/science.2643163>

- 1760 Liberman, A.M., & Whalen, D.H. (2000). On the relation of speech to language. *Trends in Cognitive*
1761 *Sciences*, 4, 187–196. [http://dx.doi.org/10.1016/S1364-6613\(00\)01471-6](http://dx.doi.org/10.1016/S1364-6613(00)01471-6)
- 1762 Lotto, A.J. (2000). Language acquisition as complex category formation. *Phonetica*, 57, 189–96.
1763 <http://dx.doi.org/10.1159/000028472>
- 1764 Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood
1765 activation, and PARSYN. *Perception and Psychophysics*, 62, 615–625.
- 1766 MacNeilage, P.F. (1998). The frame/content theory of evolution of speech production. *Behavioral*
1767 *and Brain Sciences*, 21, 499–546. <http://dx.doi.org/10.1017/S0140525X98001265>
- 1768 Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis
1769 of vocal tract shapes using an articulatory model. In W. Hardcastle & A. Marchal (Eds.),
1770 *Speech production and speech modeling* (pp. 131–149). Kluwer Academic.
- 1771 Maeda, S. & Honda, K. (1994). From EMG to formant patterns: the implication of vowel spaces.
1772 *Phonetica*, 51, 17–29. <http://dx.doi.org/10.1159/000261955>
- 1773 Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and*
1774 *Processing of Visual Information*. W.H. Freeman and Company, New York, USA.
- 1775 Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*.
1776 Laurence Erlbaum Associates, London.
- 1777 Massaro, D. W. (1998). *Perceiving Talking Faces*. MIT, Cambridge, Ma).
- 1778 Massaro, D.W., & Oden, G.C. (1980). Evaluation and integration of acoustic features in speech
1779 perception. *Journal of the Acoustical Society of America*, 67, 996–1013.
1780 <http://dx.doi.org/10.1121/1.383941>
- 1781 McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive*
1782 *Psychology*, 18, 1–86. [http://dx.doi.org/10.1016/0010-0285\(86\)90015-0](http://dx.doi.org/10.1016/0010-0285(86)90015-0)
- 1783 McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
1784 <http://dx.doi.org/doi:10.1038/264746a0>

- 1785 McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories:
1786 insights from a computational approach. *Developmental science*, *12*, 369–378.
1787 <http://dx.doi.org/10.1111/j.1467-7687.2009.00822.x>
- 1788 Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A
1789 precursor of language acquisition in young infants. *Cognition*, *29*, 143-178.
- 1790 Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., & Iacoboni, M. (2007). The essential role of
1791 premotor cortex in speech perception. *Current Biology*, *17*, 1692–1696.
1792 <http://dx.doi.org/10.1016/j.cub.2007.08.064>
- 1793 Meltzoff, A.N., & Moore, M.K. (1977). Imitation of facial and manual gestures by human neonates.
1794 *Science*, *4312*, 75–78. <http://dx.doi.org/10.1126/science.198.4312.75>
- 1795 Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., & Vallée, N. (2002). Auditory normalization: of
1796 French vowels synthesized by an articulatory model simulating growth from birth to
1797 adulthood. *Journal of the Acoustical Society of America*, *111*, 1892–1905.
1798 <http://dx.doi.org/10.1121/1.1459467>
- 1799 Ménard, L., Schwartz, J.-L., & Aubin, J. (2008). Invariance and variability in the production of the
1800 height feature in French vowels. *Speech Communication*, *50*, 14–28.
1801 <http://dx.doi.org/10.1016/j.specom.2007.06.004>
- 1802 Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2004). The role of vocal tract morphology in speech
1803 development: Perceptual targets and sensori-motor maps for French synthesized vowels from
1804 birth to adulthood. *Journal of Speech Language and Hearing Research*, *47*, 1059–1080.
1805 [http://dx.doi.org/10.1044/1092-4388\(2004/079\)](http://dx.doi.org/10.1044/1092-4388(2004/079))
- 1806 Meunier, C. (2007). Phonétique acoustique. In P. Auzou (Ed.), *Les dysarthries* (pp. 164–173). Solal.
- 1807 Moore, R. (2007). Spoken language processing: Piecing together the puzzle. *Speech Communication*,
1808 *49*, 418–435. <http://dx.doi.org/10.1016/j.specom.2007.01.011>
- 1809 Möttönen, R., Dutton, R., & Watkins, K.E. (2013). Auditory-motor processing of speech sounds.

- 1810 *Cerebral Cortex*, 23, 1190–1197. <http://dx.doi.org/10.1093/cercor/bhs110>
- 1811 Möttönen, R., & Watkins, K.E. (2009). Motor representations of articulators contribute to categorical
1812 perception of speech sounds. *The Journal of Neuroscience*, 29, 9819–9825.
1813 <http://dx.doi.org/10.1523/JNEUROSCI.6018-08.2009>
- 1814 Moulin-Frier, C., Diard, J., Schwartz, J.-L., & Bessière, P. (2015). COSMO (“Communicating about
1815 Objects using Sensory-Motor Operations”): a Bayesian modeling framework for studying
1816 speech communication and the emergence of phonological systems. *Journal of Phonetics*, 53,
1817 5-41.
- 1818 Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., & Diard, J. (2012). Adverse conditions
1819 improve distinguishability of auditory, motor, and perceptuo-motor theories of speech
1820 perception: An exploratory Bayesian modelling study. *Language and Cognitive Processes*,
1821 27, 1240–1263. <http://dx.doi.org/10.1080/01690965.2011.645313>
- 1822 Moulin-Frier, C., Schwartz, J.-L., Diard, J., & Bessière, P. (2011). Emergence of phonology through
1823 deictic games within a society of sensori-motor agents in interaction. In A. Vilain, J.-L.
1824 Schwartz, C. Abry & J. Vauclair (Eds.) *Primate Communication and Human Language* (pp.
1825 193–220). John Benjamins Publishing Company.
- 1826 Nearey, T.M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, 18, 347–73
- 1827 Norris D., & McQueen J.M. (2008). Shortlist B: a Bayesian model of continuous speech recognition.
1828 *Psychol Rev.*, 115, 357-95. doi: 10.1037/0033-295X.115.2.357.
- 1829 Öhman, S. (1966). Coarticulation in VCV utterances: spectrographic measurements. *Journal of the*
1830 *Acoustical Society of America*, 39, 151–168. <http://dx.doi.org/10.1121/1.1909864>
- 1831 Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I., Joensuu, R., Autti, T., & Sams, M. (2005).
1832 Processing of audiovisual speech in Broca's area. *NeuroImage*, 25, 333–338.
1833 <http://dx.doi.org/10.1016/j.neuroimage.2004.12.001>.
- 1834 Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N. E. et al. (2012).

- 1835 Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251.
1836 doi:10.1371/journal.pbio.1001251
- 1837 Patri, J.F. Diard, J., & Perrier, P. (2015). Optimal speech motor control and token-to-token variability:
1838 a Bayesian modeling approach. *Biological Cybernetics*, 109, 611-626.
1839 <http://link.springer.com/article/10.1007/s00422-015-0664-4>.
- 1840 Perkell, J. & Klatt, D. H. (1986). *Invariance and variability in speech processes*. Erlbaum.
- 1841 Perrier, P. (2005). Control and representations in speech production. *ZAS Papers in Linguistics*, 40,
1842 109–132.
- 1843 Perrier, P., Løevenbruck, H., & Payan, Y. (1996). Control of tongue movements in speech: The
1844 equilibrium point hypothesis perspective. *Journal of Phonetics*, 24, 53–75.
- 1845 Pitermann, M. (2000). Effect of speaking rate and contrastive stress on formant dynamics and vowel
1846 perception. *Journal of the Acoustical Society of America*, 107, 3425–3437.
1847 <http://dx.doi.org/10.1121/1.429413>
- 1848 Polka, L., Masapollo, M., & Ménard, L. (2014). Who’s talking now? Infants’ perception of vowels
1849 with infant vocal properties. *Psychological Science*, 25, 1448–1456.
- 1850 Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006).
1851 Motor cortex maps articulatory features of speech sounds. *Proceedings of the National*
1852 *Academy of Sciences*, 103, 7865–7870. <http://dx.doi.org/10.1073/pnas.0509989103>
- 1853 Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996a). Premotor cortex and the recognition of
1854 motor actions. *Cognitive Brain Research*, 3, 131–141. [http://dx.doi.org/10.1016/0926-](http://dx.doi.org/10.1016/0926-6410(95)00038-0)
1855 6410(95)00038-0
- 1856 Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996b).
1857 Localization of grasp representations in humans by PET: 1. observation vs. execution.
1858 *Experimental Brain Research*, 111, 246–252. <http://dx.doi.org/10.1007/BF00227301>
- 1859 Robert-Ribes, J., Schwartz, J.-L., Lallouache, T. & Escudier, P. (1998). Complementarity and synergy

- 1860 in bimodal speech : auditory, visual, and audio-visual identification of French oral vowels in
1861 noise. *Journal of the Acoustical Society of America*, 103, 3677-3689.
1862 <http://dx.doi.org/10.1121/1.423069>
- 1863 Rogers, J. C., Möttönen, R., Boyles, R., & Watkins, K. E. (2014). Discrimination of speech and non-
1864 speech sounds following theta-burst stimulation of the motor cortex. *Frontiers in Psychology*,
1865 5, 754. <http://dx.doi.org/10.3389/fpsyg.2014.00754>
- 1866 Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in
1867 the theory of signal detection. *Psychonomic Bulletin & Review*, 12, 573–604.
1868 <http://dx.doi.org/10.3758/BF03196750>
- 1869 Sato, M., Grabski, K., Glenberg, A., Brisebois, A., Basirat, A., Ménard, L. & Cattaneo, L. (2011).
1870 Articulatory bias in speech categorization: evidence from use-induced motor plasticity. *Cortex*,
1871 47, 1001–1003. <http://dx.doi.org/10.1016/j.cortex.2011.03.009>
- 1872 Sato, M., Tremblay, P. & Gracco, V. (2009). A mediating role of the premotor cortex in phoneme
1873 segmentation. *Brain and Language*, 111, 1–7. <http://dx.doi.org/10.1016/j.bandl.2009.03.002>
- 1874 Savariaux, C., Perrier P., & Orliaguet, J.-P. (1995). Compensation strategies for the perturbation of
1875 the rounded vowel [u] using a lip-tube: a study of the control space in speech production.
1876 *Journal of the Acoustical Society of America*, 98, 2428–2442.
1877 <http://dx.doi.org/10.1121/1.413277>
- 1878 Savariaux, C., Perrier, P., Orliaguet, J.P. & Schwartz, J.-L. (1999). Compensation strategies
1879 for the perturbation of French [u] using a lip tube. II. Perceptual analysis. *Journal of the*
1880 *Acoustical Society of America*, 106, 381–393. <http://dx.doi.org/10.1121/1.427063>
- 1881 Scharenborg, O., Norris, D., Ten Bosch, L., & McQueen, J. M. (2005). How should a speech
1882 recognizer work? *Cognitive Science*, 29, 867-918. doi:10.1207/s15516709cog0000_37.
- 1883 Schwartz, J.L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech
1884 perception is subject-dependent. *Journal of the Acoustical Society of America*, 127, 1584-

- 1885 1594. <http://dx.doi.org/10.1121/1.3293001>
- 1886 Schwartz, J.-L., Abry, C., Boë, L.-J., & Cathiard, M. (2002). Phonology in a theory of perception-for-
1887 action-control. In J. Durand, B. Laks (Eds.) *Phonology: from Phonetics to Cognition* (pp. 255–
1888 280). Oxford: Oxford University Press.
- 1889 Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012a). The Perception-for-Action-Control
1890 Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*,
1891 25, 336–354. doi:10.1016/j.jneuroling.2009.12.004
- 1892 Schwartz, J.-L., Boë, L.-J., & Abry, C. (2007). Linking the Dispersion-Focalization Theory (DFT)
1893 and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a
1894 Perception-for-Action-Control Theory (PACT). In M.J. Solé, P. Beddor & M. Ohala (Eds.)
1895 *Experimental Approaches to Phonology* (pp. 104–124). Oxford University Press.
- 1896 Schwartz J.-L., Boë L.-J., Badin P., & Sawallis R. T. (2012b). Grounding stop place systems in the
1897 perceptuo-motor substance of speech: On the universality of the labial-coronal-velar stop
1898 series, *Journal of Phonetics*, 40, 20–36.
- 1899 Schwartz, J.L., Boë, L.J., Vallée, N., & Abry, C. (1997). Major trends in vowel system inventories.
1900 *Journal of Phonetics*, 25, 233-254.
- 1901 Shiller, D.M., Sato, M., Gracco, V.L., & Baum, S.R. (2009). Perceptual recalibration of speech
1902 sounds following speech motor learning. *Journal of the Acoustical Society of America*, 125,
1903 1103–1113. <http://dx.doi.org/10.1121/1.3058638>
- 1904 Skipper, J.I., Devlin, J.T., & Lametti, D.R. (2017). The hearing ear is always found close to the
1905 speaking tongue: Review of the role of the motor system in speech perception. *Brain and*
1906 *Language*, 164, 77-105. <http://dx.doi.org/10.1016/j.bandl.2016.10.004>
- 1907 Skipper, J.I., van Wassenhove, V., Nusbaum, H.C. & Small, S.L. (2007). Hearing lips and seeing
1908 voices: how cortical areas supporting speech production mediate audiovisual speech
1909 perception. *Cerebral Cortex*, 17, 2387–2399. <http://dx.doi.org/10.1093/cercor/bhl147>.

- 1910 Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation.
1911 In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pp. 375–380.
- 1912 Stevens, K.N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In E.
1913 David & P. Denes (Eds.), *Human communication: a unified view* (pp. 51–66). McGraw-Hill.
- 1914 Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical*
1915 *Society of America*, 68, 836–842. <http://dx.doi.org/10.1121/1.384823>
- 1916 Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17, 3-45.
- 1917 Stevens, K.N., Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants.
1918 *Journal of the Acoustical Society of America*, 64, 1358–1368.
1919 <http://dx.doi.org/10.1121/1.382102>
- 1920 Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W.
1921 Wathen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 88–102).
1922 Cambridge, MA: MIT Press.
- 1923 Strange, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of*
1924 *America*, 85, 2081–2087. <http://dx.doi.org/10.1121/1.397860>
- 1925 Sumbly, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of*
1926 *the Acoustical Society of America*, 26, 212–215. <http://dx.doi.org/doi:10.1121/1.1907309>
- 1927 Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech
1928 perception. In B Dodd and R Campbell (Eds.) *Hearing by eye: the psychology of lip-reading*
1929 (pp 3–51). Lawrence Erlbaum Associates, London.
- 1930 Sun, J., & Deng, L. (2002). An overlapping-feature based phonological model incorporating linguistic
1931 constraints: Applications to speech recognition. *Journal of the Acoustical Society of America*,
1932 111, 1086–1101.
- 1933 Sussman, H., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: the
1934 orderly output constraint. *Behavioral and Brain Sciences*, 21(02), 241–259.

- 1935 Sussman, H. M., Hoemeke, K. & Ahmed, F. (1993) A cross-linguistic investigation of locus equations
1936 as a relationally invariant descriptor for place of articulation. *Journal of the Acoustical Society*
1937 *of America*, 94, 1256–68. <http://dx.doi.org/10.1121/1.408178>
- 1938 Sussman, H. M., McCaffrey, H. A. & Matthews, S. A. (1991) An investigation of locus equations as a
1939 source of relational invariance for stop place categorization. *Journal of the Acoustical Society*
1940 *of America*, 90, 1309–25. <http://dx.doi.org/10.1121/1.401923>
- 1941 Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological*
1942 *Review*, 61, 401–409. <http://dx.doi.org/10.1037/h0058700>
- 1943 Toscano, J. C., & McMurray, B. (2008). Using the distributional statistics of speech sounds for
1944 weighting and integrating acoustic cues. In *Proceedings of the 30th Annual Conference of the*
1945 *Cognitive Science Society*, pp. 433–439.
- 1946 Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in
1947 speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34, 434–
1948 464. <http://dx.doi.org/10.1111/j.1551-6709.2009.01077.x>
- 1949 Treille, A., Cordeboeuf, C., Vilain, C., & Sato, M. (2014). Haptic and visual information speed up the
1950 neural processing of auditory speech in live dyadic interactions. *Neuropsychologia*, 57, 71–77.
1951 <http://dx.doi.org/doi:10.1016/j.neuropsychologia.2014.02.004>
- 1952 Vallabha, G. K., McClelland, J. L., Pons F., Werker, J. F., & Amano, S. (2007). Unsupervised
1953 learning of vowel categories from infant-directed speech. *Proceedings of the National*
1954 *Academy of Sciences*, 104, 13273–13278. <http://dx.doi.org/10.1073/pnas.0705369104>
- 1955 Watkins, K.E., Strafella, A.P., & Paus, T. (2003). Seeing and hearing speech excites the motor system
1956 involved in speech production. *Neuropsychologia*, 41, 989–994.
1957 [http://dx.doi.org/10.1016/S0028-3932\(02\)00316-0](http://dx.doi.org/10.1016/S0028-3932(02)00316-0)
- 1958 Werker, J.F., & Hensch, T.K. (2015). Critical Periods in Speech Perception: New Directions. *Annu.*
1959 *Rev. Psychol.* 66, 173–96.

- 1960 Werker, J.F., & Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual
1961 reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.
1962 [http://dx.doi.org/10.1016/S0163-6383\(84\)80022-3](http://dx.doi.org/10.1016/S0163-6383(84)80022-3)
- 1963 Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in
1964 productibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage*, 33,
1965 316–325. <http://dx.doi.org/10.1016/j.neuroimage>.
- 1966 Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates
1967 motor areas involved in speech production. *Nature Neuroscience*, 7, 701–702.
1968 <http://dx.doi.org/10.1038/nn1263>
- 1969 Zekveld, A.A., Heslenfeld, D.J., Festen, J.M., & Schoonhoven, R. (2006). Top–down and bottom–up
1970 processes in speech comprehension. *NeuroImage*, 32, 1826–1836.
1971 <http://dx.doi.org/10.1016/j.neuroimage.2006.04.199>
- 1972
- 1973
- 1974

1975 Footnotes

1976 ⁽¹⁾ It is well known that other sensory systems may intervene in speech perception:
1977 such as vision, through lip-reading (Sumbly & Pollack, 1954; McGurk & MacDonald, 1976;
1978 Summerfield, 1987) but also possibly somato-sensory processing (Fowler & Dekle, 1991; Ito,
1979 Tiede & Ostry, 2009; Treille et al., 2014). However, in this paper we will focus on the
1980 auditory vs. motor systems, acknowledging that these other sensory systems could be
1981 incorporated as additional sensory inputs inside a perceptuo-motor framework.

1982 ⁽²⁾ Here, we use the term “communication object” in a broad sense, conflating
1983 different levels of analysis (phonetics, phonology, syntax, semantics). In this paper, objects
1984 will only refer to phonological entities.

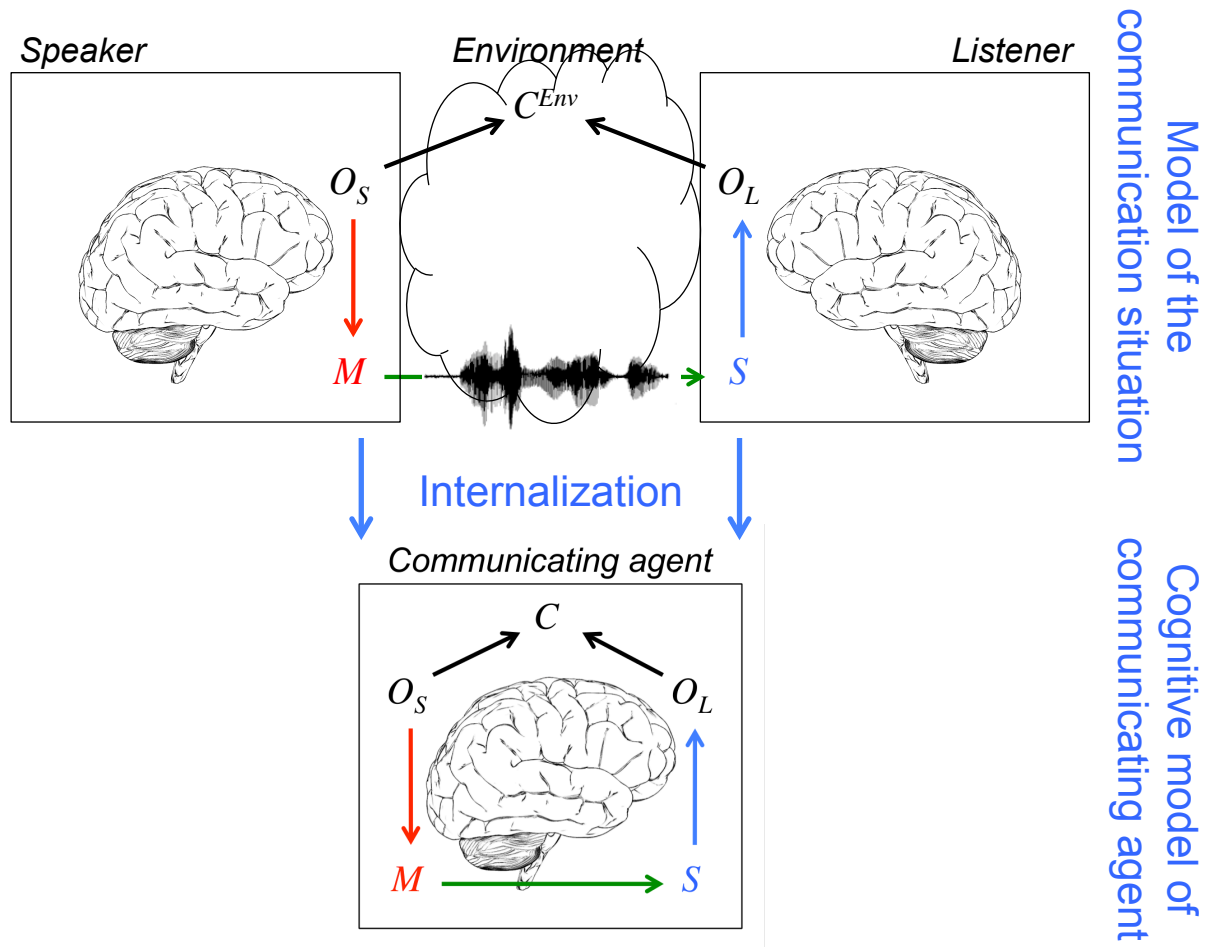
1985 ⁽³⁾ Technically, probability values below the ϵ threshold are set to ϵ during perception
1986 inference, but are set to 0 during learning. This makes learning approximate but fast, as
1987 portions of spaces with very low probabilities are dismissed altogether.

1988 ⁽⁴⁾ Notice that the visual system could intervene in this process, especially considering
1989 the natural complementarity of auditory and visual representations in the depiction of vowels
1990 and plosives (Summerfield, 1987; Robert-Ribes et al., 1998).

1991 ⁽⁵⁾ VLAM is actually an articulatory rather than a motor model of speech production.
1992 VLAM inputs are parameters controlling the shape of the tongue and lips and the position of
1993 the jaw, which are themselves the results of motor commands at a higher level (see e.g.
1994 Perrier et al., 1996; Perrier, 2005). We consider as a simplification that VLAM articulatory
1995 parameters are part of the control system and hence could provide “motor commands” at a
1996 certain level of representation in the motor pathway.

1997 ⁽⁶⁾ While we display mean entropies in this Figure, averaging over the 9 syllables,
1998 there are actually differences between entropy dynamics among the different syllables,
1999 particularly in the motor space. This is clearly seen in Figure 14, where it appears that

2000 convergence is more rapid for /u/ than for the other two vowels. The likely reason is that the
2001 available articulatory space for achieving the adequate formants is more restricted for /u/ in
2002 the available 3-D articulatory space in VLAM. Notice that slower articulatory convergence
2003 (displayed in Figure 14) can occur in spite of rapid acoustic convergence (as displayed by the
2004 rapid formant convergence for /i/ in Figure 14). It is beyond the scope of the present paper to
2005 discuss the importance and significance of these differences in convergence among vowels,
2006 plosives or syllables.
2007



2008

2009 *Figure 1.* The communication situation, which involves a speaker agent and listener agents
 2010 interacting within an environment, is internalized in communicating agents. Top, model of the
 2011 communication situation: the speaker wants to mention a linguistic object (in a broad sense,
 2012 see footnote 2) O_S . She/he produces a motor gesture M leading to the production of a sound S
 2013 propagating in the environment towards the listeners who recover linguistic objects O_L . The
 2014 success of communication is estimated by the Boolean variable C^{Env} . Bottom, all variables are
 2015 internalized to provide a cognitive model of the communicating agent.

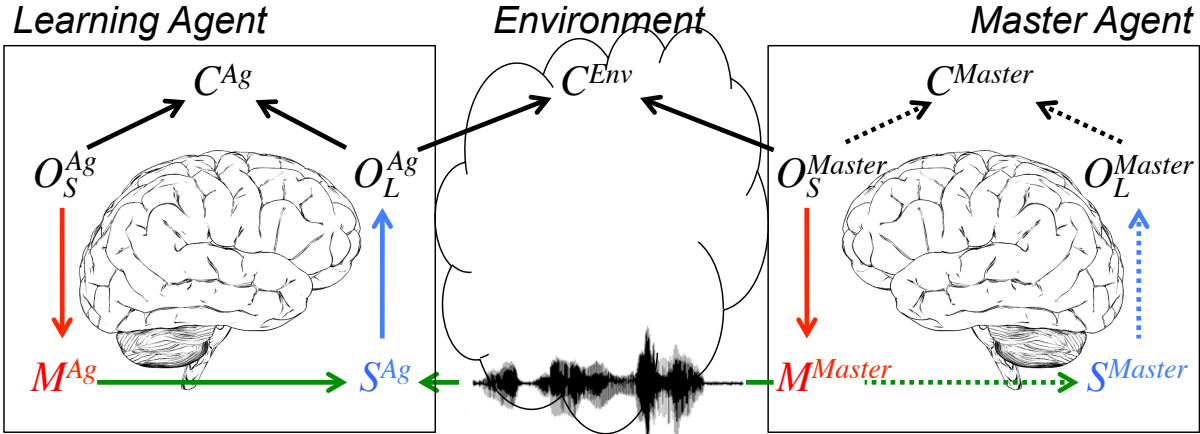
2016

	Production infer $P(M O)$	Perception infer $P(O S)$
Motor theory focus on O_S	$\underbrace{P(M O_S)}_{\text{motor repertoire}}$	$\propto \sum_M \left(\underbrace{P(M O_S)}_{\text{motor decoder}} \underbrace{P(S M)}_{\text{inverse model}} \right)$
Auditory theory focus on O_L	$\propto P(M) \sum_S \left(\underbrace{P(S M)}_{\text{direct model}} \underbrace{P(O_L S)}_{\text{sensory targets}} \right)$	$\underbrace{P(O_L S)}_{\text{sensory classifier}}$
Perceptuo-motor theory $C=True$, i.e. $O_S = O_L$	$\propto \underbrace{P(M [O_S = O_L])}_{\text{motor production}} \sum_S \left(\underbrace{P(S M) P(O_L S)}_{\text{sensory production}} \right)$	$\propto \underbrace{P([O_L = O_S S])}_{\text{sensory perception}} \sum_M \left(\underbrace{P(M O_S) P(S M)}_{\text{motor perception}} \right)$

2017
2018 *Figure 2.* Probabilistic inferences for production and perception tasks instantiated within the
2019 framework of the motor, auditory and perceptuo-motor theories. The \propto symbol denotes
2020 proportionality, i.e. to correctly obtain probability distributions, the expression shown has to
2021 be normalized. The denominations of the components of each equation refer to their possible
2022 interpretation in terms of cognitive processes:

- 2023 - sensory targets refer to the set of sensory distributions for each object (typically,
2024 “sensory” would be replaced by “auditory” in a basic auditory theory of speech
2025 perception, or possibly “audio-visual” in a modified version taking into account lip-
2026 reading: see note (1)),
- 2027 - motor repertoire refers to the set of motor distributions for each object,
- 2028 - sensory production refers to the distribution of sensory data (typically sounds) for each
2029 object,
- 2030 - motor production refers to the distribution of motor commands (typically articulatory
2031 gestures) for each object,
- 2032 - sensory classifier refers to the possibility of recovering the object from the stimulus
2033 input (typically the sound),
- 2034 - motor decoder refers to the possibility (thanks to the Bayesian summation) of
2035 recovering the object from the motor commands (or, in some variants of motor

- 2036 theories, the articulatory gesture),
- 2037 - direct model refers to the possibility of predicting sensory information from the motor
- 2038 command (typically sound from the gesture),
- 2039 - inverse model refers to the possibility of recovering the motor command from sensory
- 2040 information (typically the gesture from sound).
- 2041
- 2042



2043

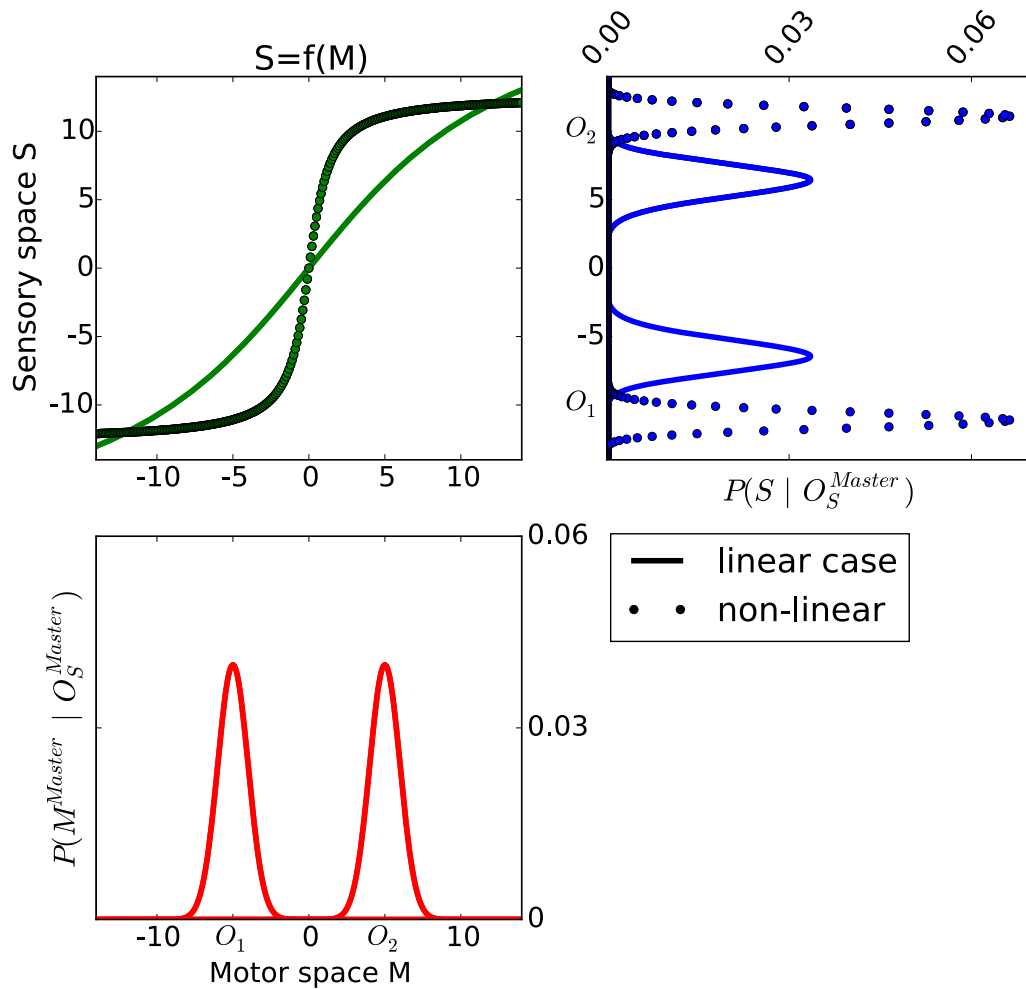
2044

Figure 3. Schema of the supervised learning scenario, where the Master Agent provides the Learning Agent with <object, stimulus> pairs.

2045

2046

2047

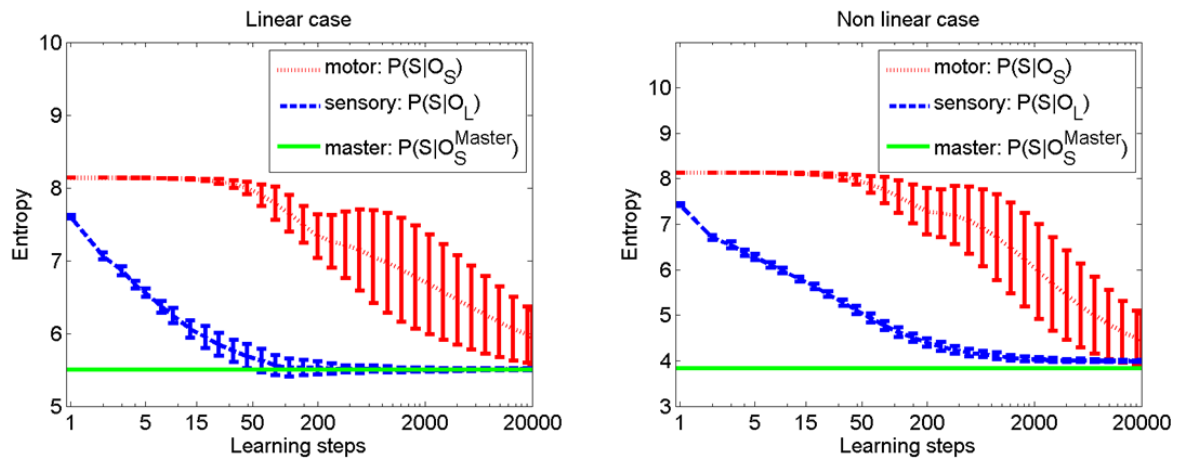


2048

2049 *Figure 4.* Summary of the stimulus production process of the Master Agent. Its motor
 2050 repertoire is shown in the lower left panel. The model f of the motor-to-sensory
 2051 transformation is shown in the upper left panel, for two values of the nonlinearity parameter
 2052 ($a = 0.01$ for the quasi-linear case and $a = 0.1$ for the nonlinear case). The probability
 2053 distributions of the resulting sensory inputs received by the Learning Agent are shown in the
 2054 upper right panel.

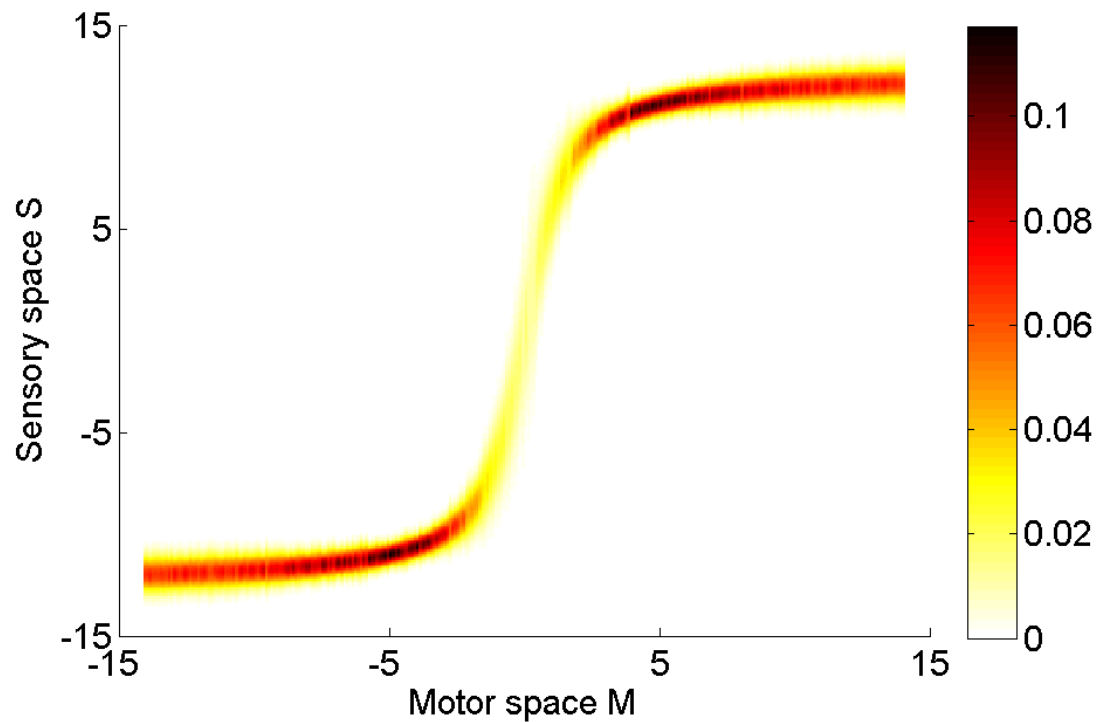
2055

2056



2057

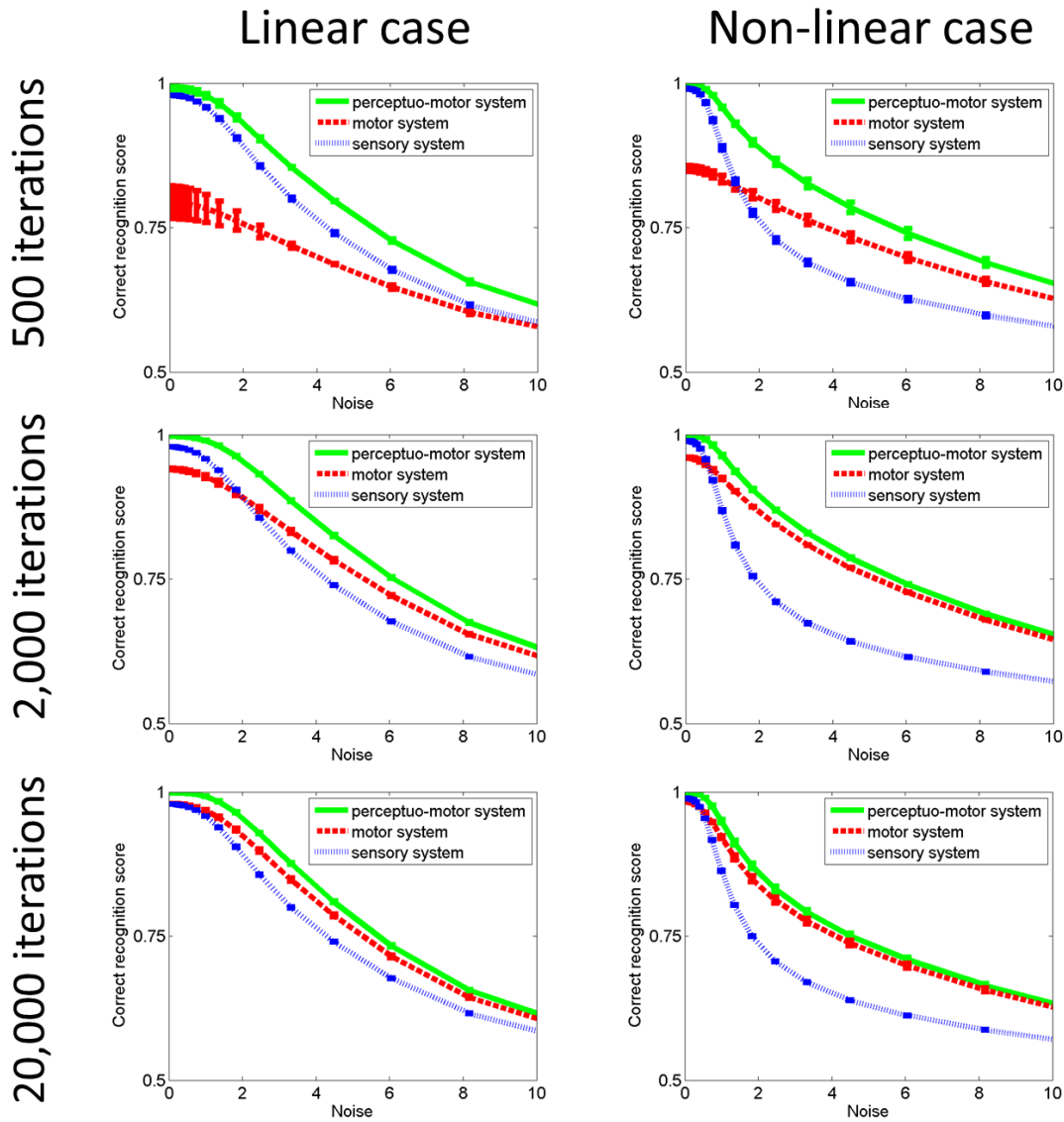
2058 *Figure 5.* Evolution of entropies of the sensory and motor models of the Learning Agents and
 2059 production system of the Master Agent, as a function of the number of iterations of the
 2060 learning algorithm, averaged over the possible object values. Left column: linear case; right
 2061 column: nonlinear case. In each case, 12 different simulations were run, corresponding to
 2062 random initializations of the learning process. The standard deviations shown are computed
 2063 over these 12 different simulations.



2064

2065 *Figure 6.* An instance of the learned internal model of the motor-to-sensory transformation,
2066 after 20,000 learning iterations, in a nonlinear setting ($\alpha = 0.1$). For each motor gesture m of
2067 the x-axis, the probability distribution over resulting sensory stimulus $P(S | [M = m])$ is read
2068 vertically, with the color code indicating probability (white to yellow to red to black color-
2069 map (light gray to black), in order of increasing probability value). Black regions (resp.
2070 yellow/light gray) therefore correspond to low-variance (resp. high variance) Gaussian
2071 probability distributions, that is to say, well-explored (resp. poorly explored) portions of the
2072 motor-to-sensory transformation.

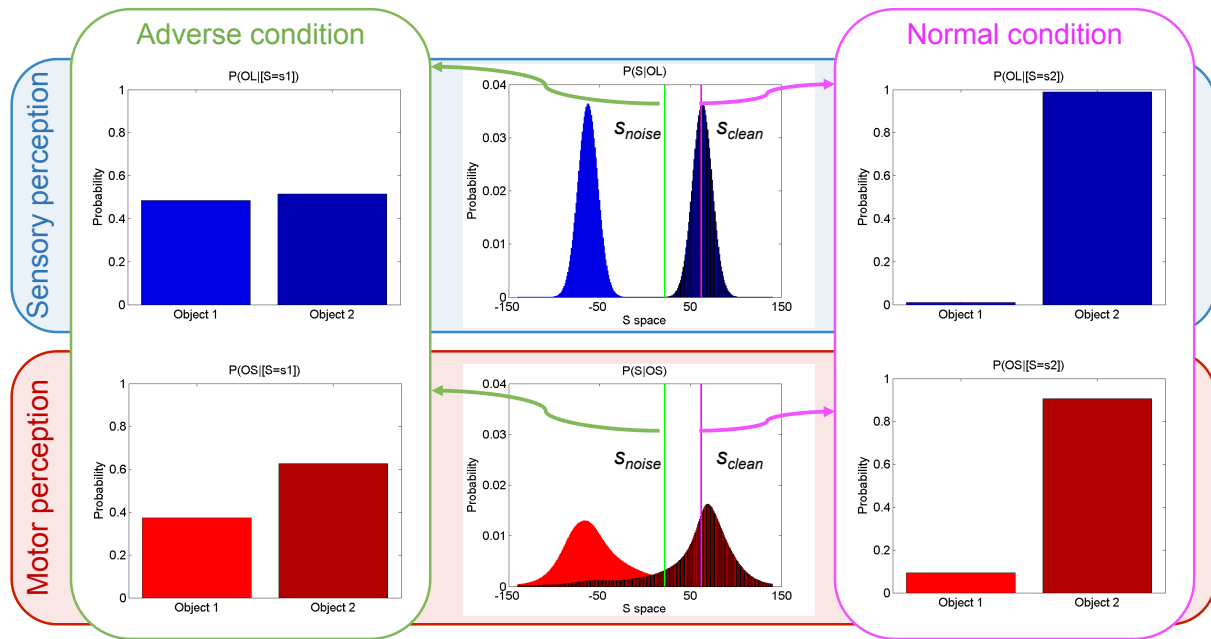
2073



2074

2075 *Figure 7.* Evolution of correct recognition scores of motor, sensory and perceptuo-motor
 2076 models of perception, as a function of environment noise. In each case, 12 different
 2077 simulations were run, corresponding to random initializations of the learning process. The
 2078 standard deviations shown are computed over these 12 different simulations.

2079



2080

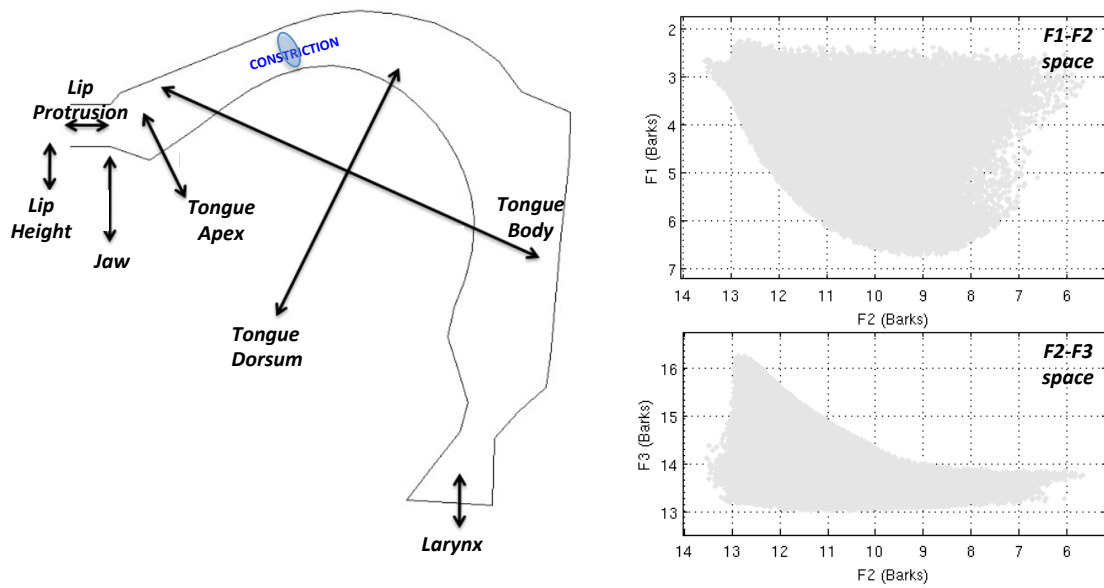
2081 *Figure 8.* Illustration (linear case, after 1,200 learning iterations) of sensory (top row) and
 2082 motor (bottom row) categorization processes on example stimuli s_{noise} in adverse conditions

2083 (left column) and s_{clean} in normal conditions (right column), as probabilistic inference from

2084 learned prototypes (center column).

2085

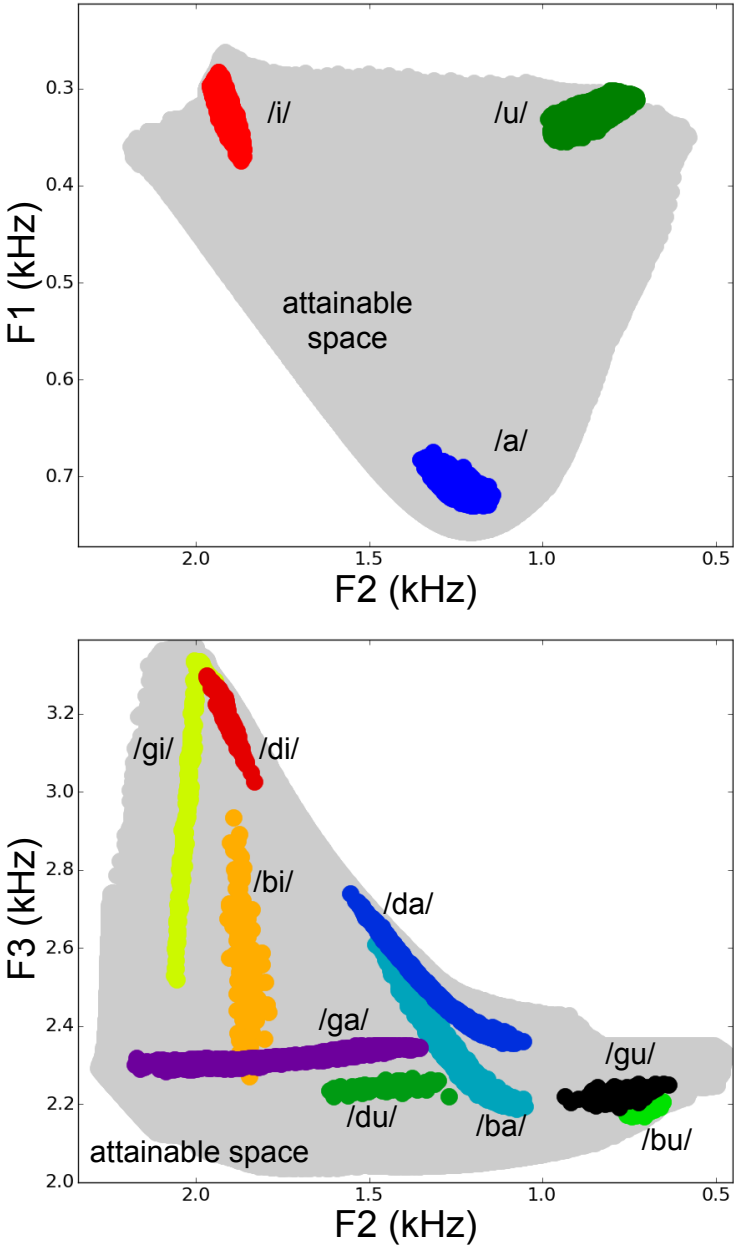
2086



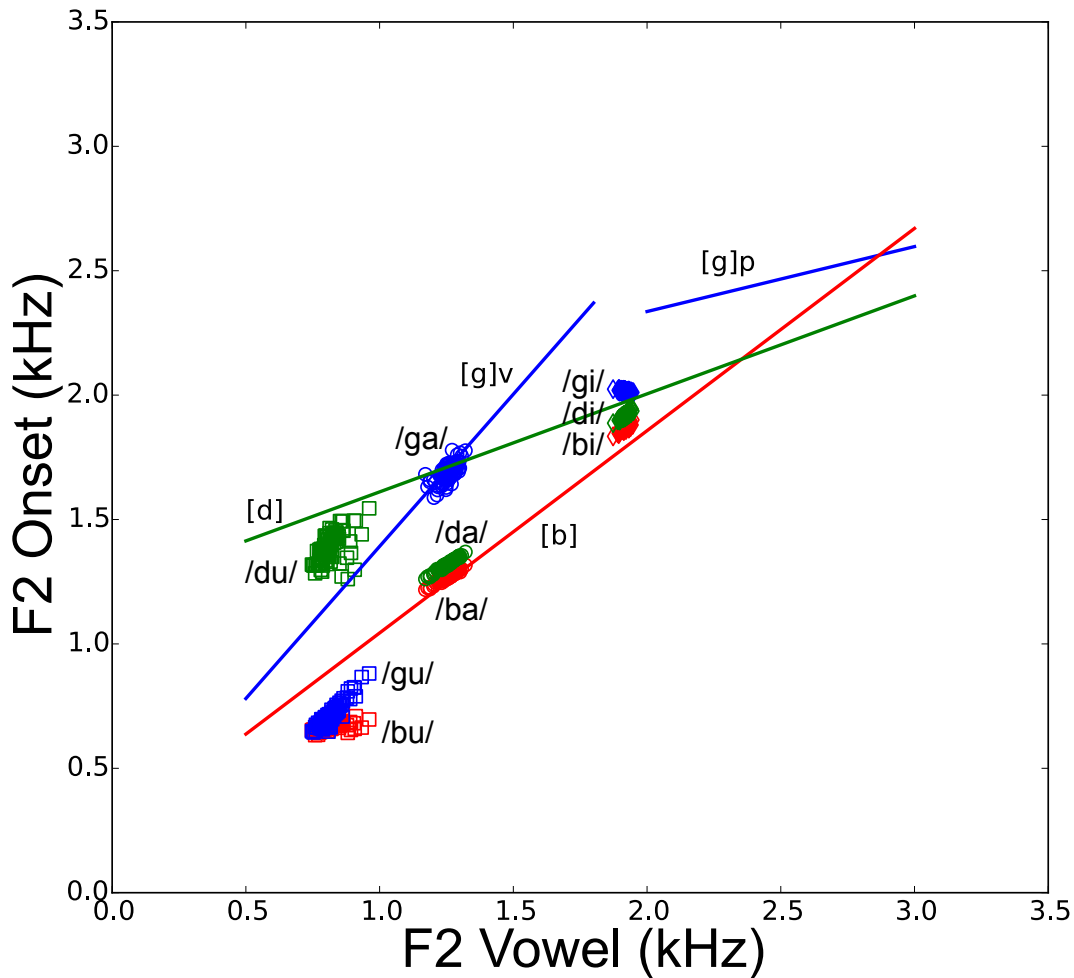
2087

2088 *Figure 9.* The vocal tract VLAM model. Left: the seven articulatory parameters (Jaw, Lip
 2089 Height and Protrusion, Tongue Body, Dorsum and Apex, and Larynx) enable the vocal tract
 2090 shape to be driven. The Constriction is defined by the position where the vocal tract area is
 2091 minimum. Vowels are constrained to have an area greater than 0.15 cm^2 . Plosives are
 2092 constrained to between 0.05 and 0.15 cm^2 . Right: plots of the regions of the acoustic space
 2093 (top: $(F1, F2)$ plane, bottom: $(F2, F3)$ plane) that result from articulatory configurations in
 2094 VLAM.

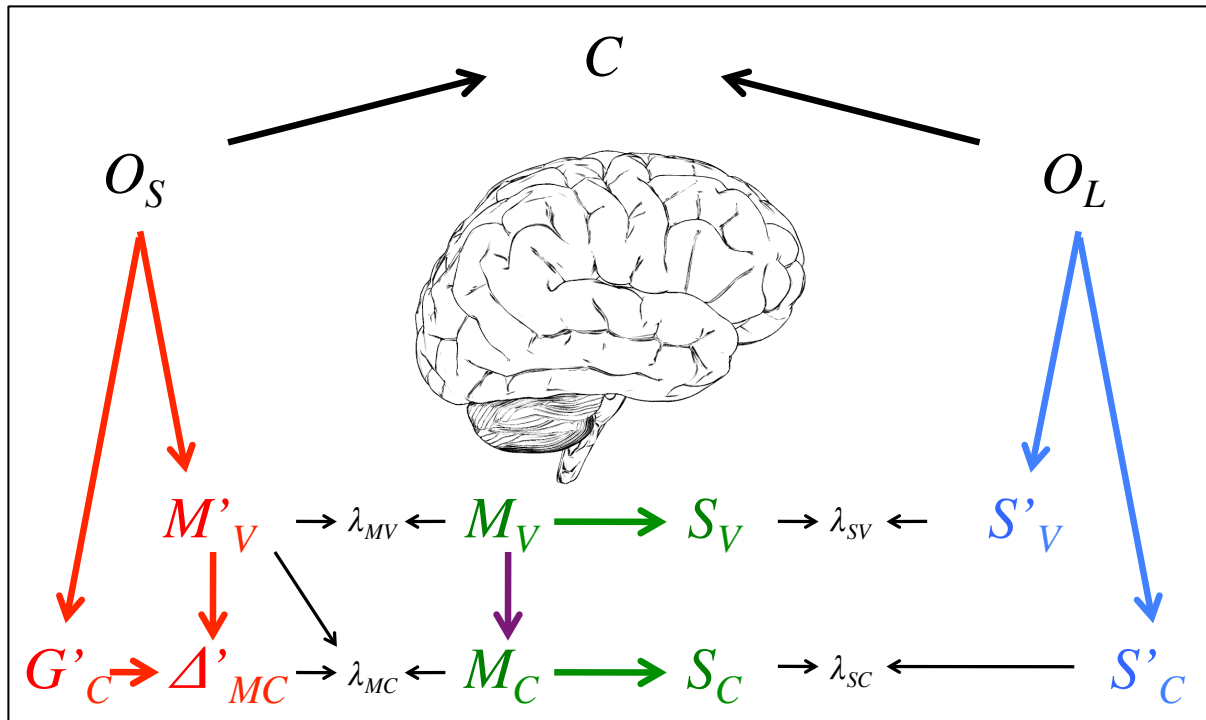
2095



2096
 2097 *Figure 10.* Synthetic syllables in acoustic space. Top: (F_2, F_1) for vowels. Bottom: (F_2, F_3)
 2098 for plosives in each syllabic context.
 2099



2100
 2101 *Figure 11.* Locus displays for VLAM simulations (color marks annotated with /ba/, /bi/, /bu/,
 2102 etc.) compared with locus equations provided by Sussman (1998) (portions of straight lines
 2103 annotated with [d], [b], etc.). For VLAM simulations, each mark is displayed at the position
 2104 corresponding to the *F2* value for the vowel on the x-axis, and the *F2* value for the consonant
 2105 on the y-axis. Sussman's locus equations are derived by pooling the same frequency
 2106 coordinates for natural utterances from 20 American English speakers (see Sussman, 1998,
 2107 Fig. 5). Sussman provides one equation for /b/, one for /d/ and two separate equations for /g/:
 2108 one when the context vowel is front and the plosive is therefore palatal, and another one when
 2109 the context vowel is back and the plosive is therefore velar.



$$\begin{aligned}
 &P(O_S G'_C M'_V \Delta'_{MC} \lambda_{MV} \lambda_{MC} M_V M_C S_V S_C \lambda_{SV} \lambda_{SC} S'_V S'_C O_L C) = \\
 &P(O_S)P(M'_V | O_S)P(G'_C | O_S)P(\Delta'_{MC} | M'_V G'_C) \\
 &P(\lambda_{MV} | M'_V M_V)P(\lambda_{MC} | M'_V \Delta'_{MC} M_C) \\
 &P(M_V)P(S_V | M_V)P(M_C | M_V)P(S_C | M_C) \\
 &P(\lambda_{SV} | S_V S'_V)P(\lambda_{SC} | S_C S'_C) \\
 &P(O_L)P(S'_V S'_C | O_L) \\
 &P(C | O_S O_L)
 \end{aligned}$$

2110

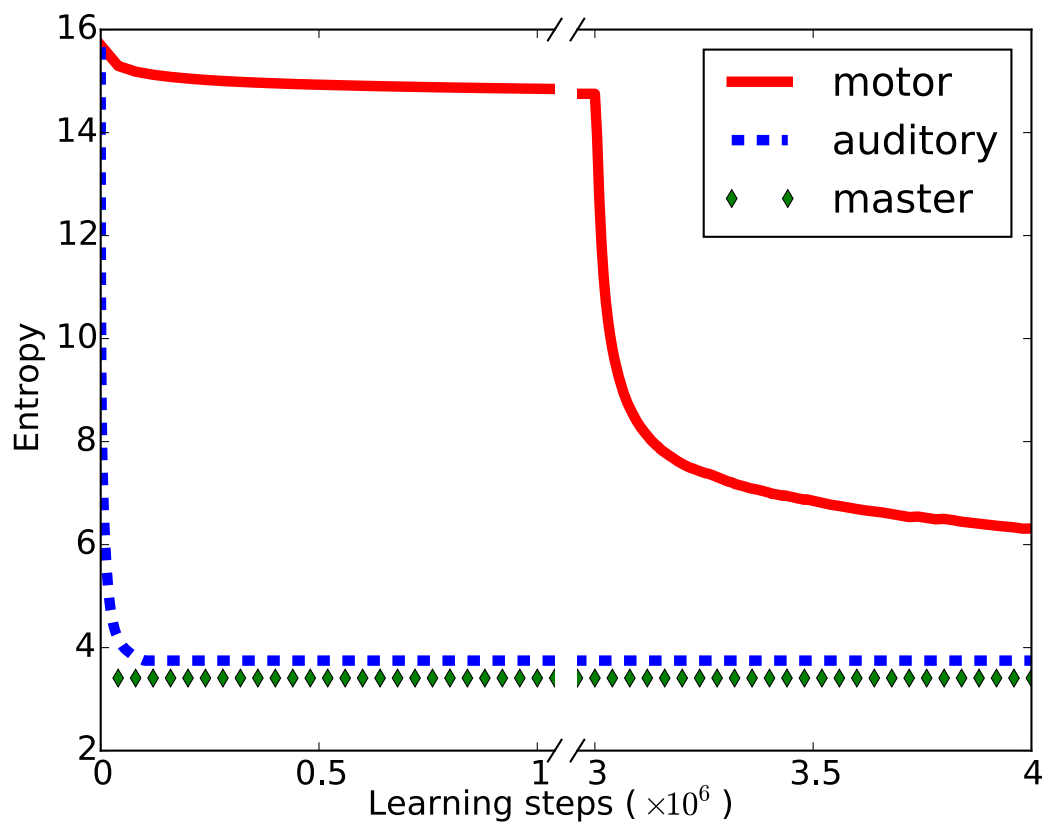
2111 *Figure 12.* The COSMO-S model for processing syllables. This is illustrated by a graphical

2112 representation (Top), and by the decomposition of its joint probability distribution as a

2113 product of probabilistic terms (Bottom). In red (left part), the motor system, in green (middle

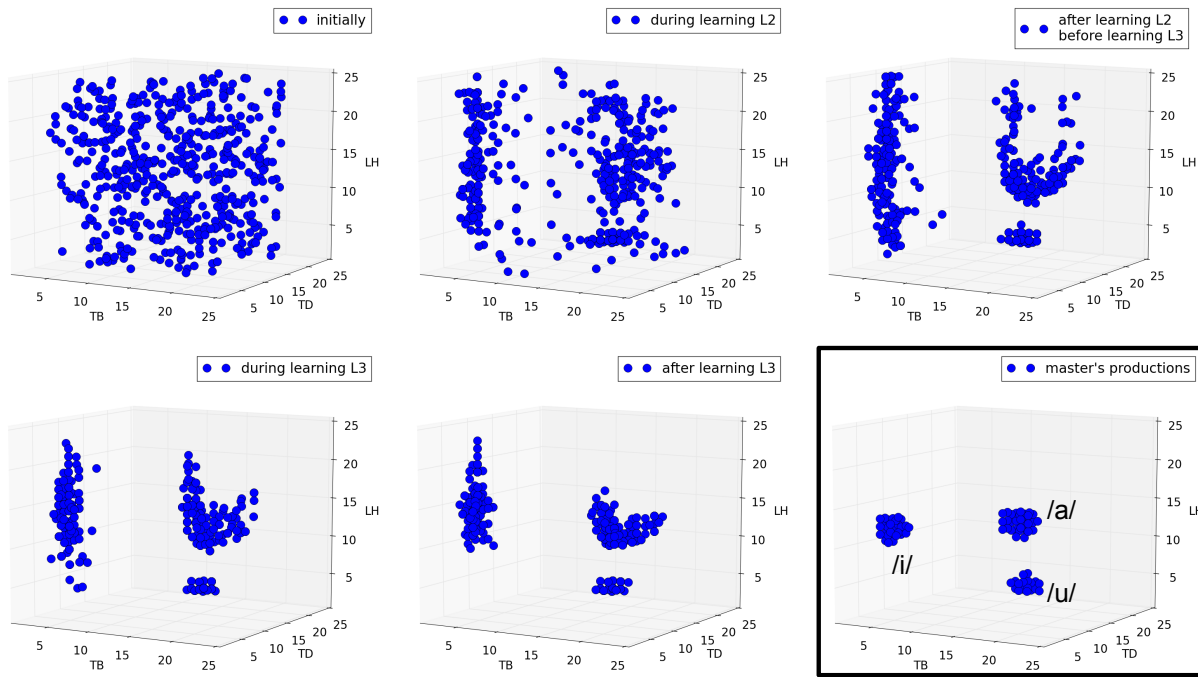
2114 part), the sensory-motor system, in blue (right part), the auditory system.

2115



2116

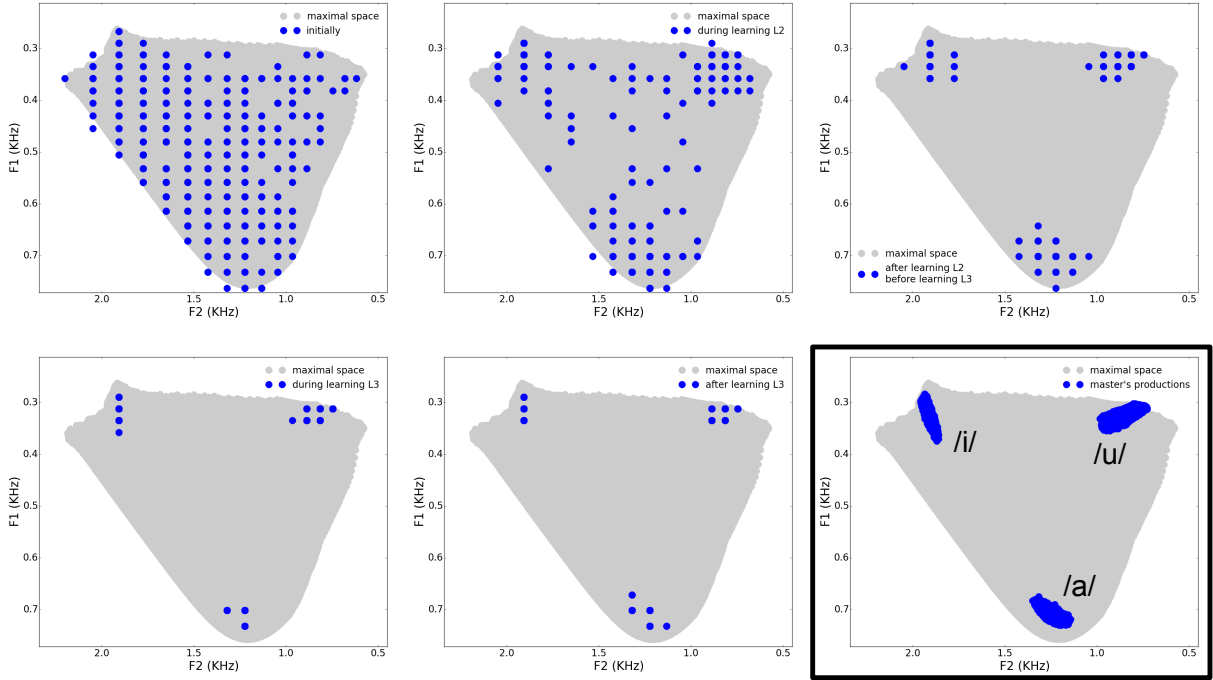
2117 *Figure 13.* Evolution of entropies of the auditory and motor models of the Learning Agents
 2118 and production system of the Master Agent, as a function of the number of iterations of the
 2119 learning algorithm, averaged over the possible object values, in the syllable experiment. For
 2120 the auditory model, learning corresponds to the sensory learning phase L1 with 4,000,000
 2121 iterations. For the motor model, learning starts with the sensory-motor learning phase L2 with
 2122 3,000,000 iterations followed by the motor learning phase L3 from 3,000,000 to 4,000,000
 2123 iterations.



2124

2125 *Figure 14.* Illustrating exploration in the motor space in COSMO-S. Each graph displays
 2126 samples from the probability distributions $P(M'_V{}^{Ag} | [S_V{}^{Ag} = s_v][O_S{}^{Ag} = o_s][\lambda_{MV}{}^{Ag} = 1])$ in the
 2127 three-dimensional space *TB* (Tongue Body), *TD* (Tongue Dorsum) and *LH* (Lip Height), with
 2128 s_v the vowel acoustic target and o_s the corresponding syllable label. Motor variables are
 2129 specified by normalized values between 0 and 25. Each panel shows 500 samples taken at 500
 2130 successive time-steps (one sample per learning iteration), during five stages of the exploration
 2131 process (see caption of each panel). The bottom right panel shows, for comparison, the motor
 2132 distribution of the Master Agent.

2133



2134

2135

Figure 15. Illustrations of the exploration in the vowel space in COSMO-S. Each graph

2136

displays the images in the acoustic ($F2, F1$) plane of the exact same motor samples as in

2137

Figure 14, via the articulatory-to-acoustic transformation. Each panel concerns the same five

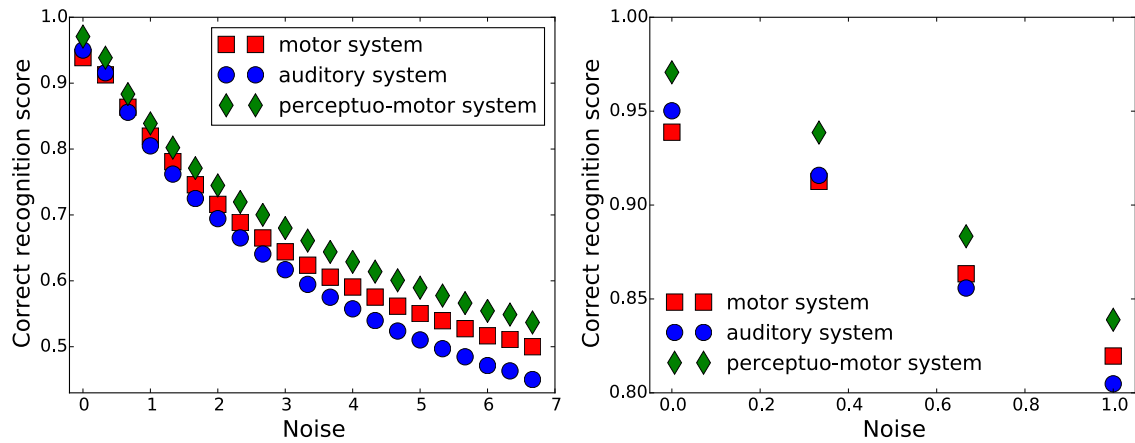
2138

stages of the exploration process as in Figure 14. For comparison, the bottom right panel

2139

shows the stimulus distribution of the Master Agent.

2140



2141

2142 *Figure 16.* Results of the classification process for syllables presented at various levels of

2143 noise. The correct recognition rates for the auditory, motor and perceptuo-motor

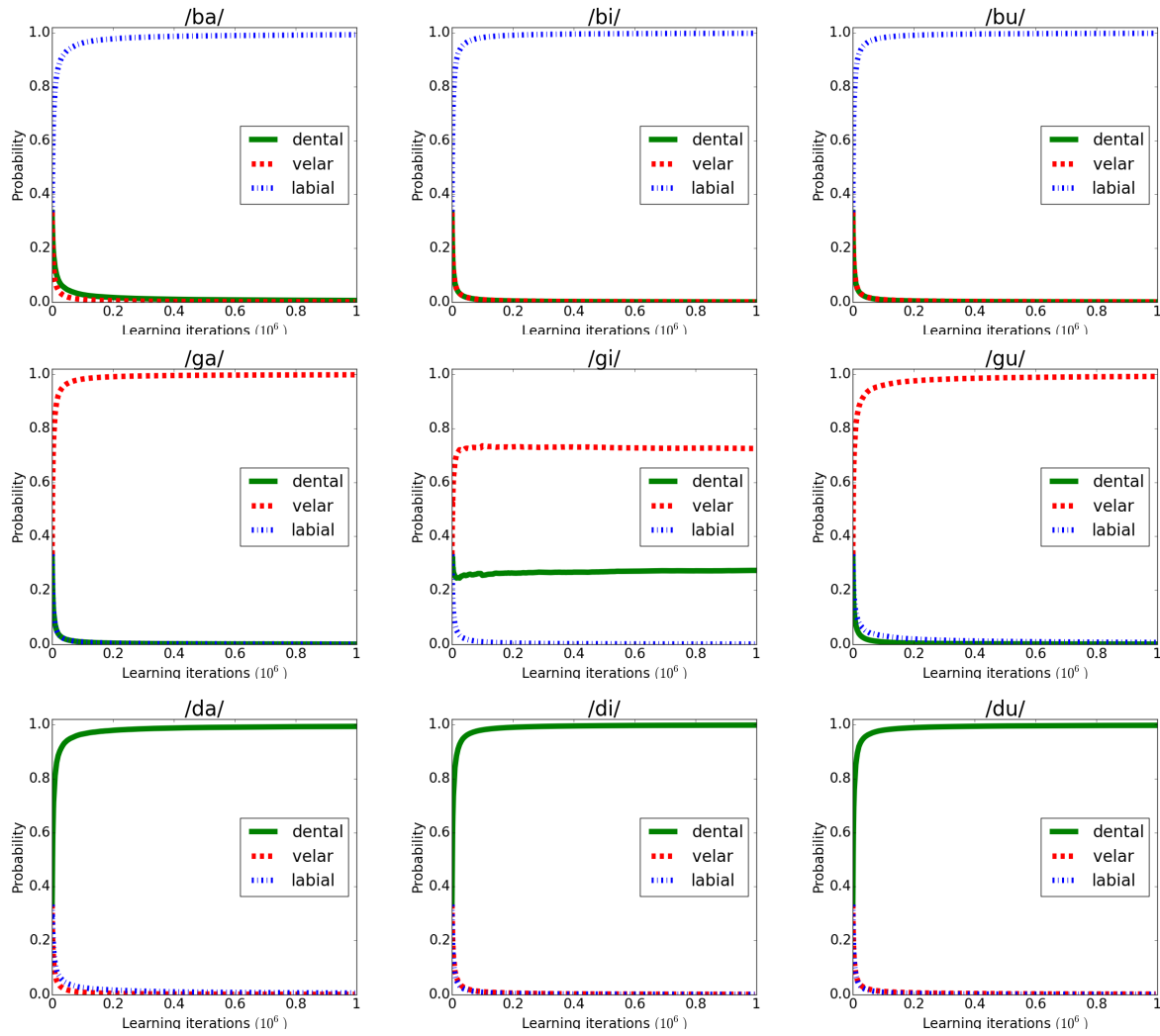
2144 implementations of the perception task in the COSMO-S model are displayed. Right plot:

2145 zoom of the left plot at low levels of noise highlights the inversion of performance between

2146 the auditory system (better under normal conditions) and the motor system (better at noisy

2147 conditions).

2148



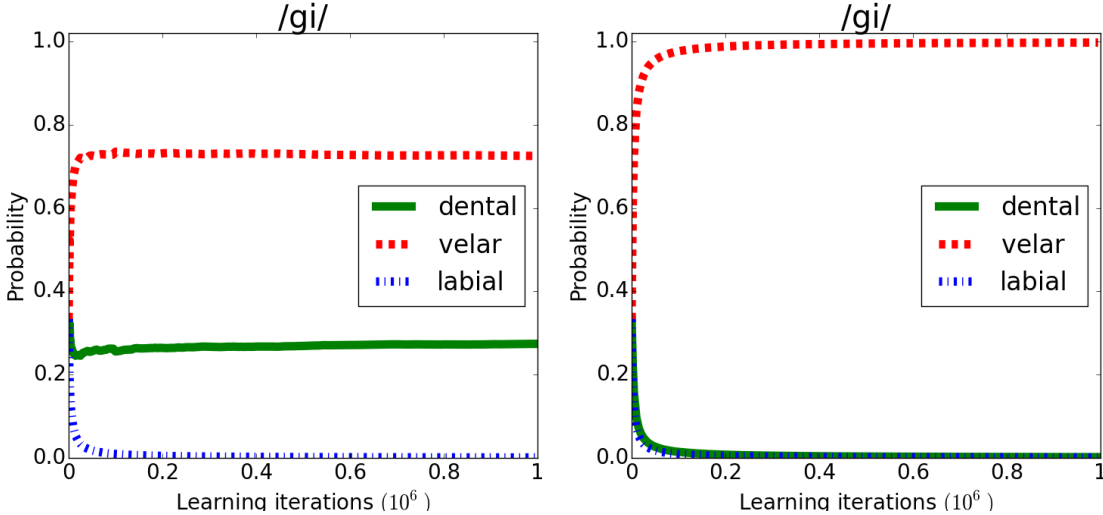
2149

2150 *Figure 17.* Learning the place of articulation for plosives. Evolution of the probabilities of the

2151 motor variable $P(G_c^{Ag} | O_S^{Ag})$ with the number of iterations in learning, for the 9 objects O_S^{Ag}

2152 (see text).

2153



2154

2155 *Figure 18.* The role of hyperarticulation in the emergence of phonological categories.

2156 Evolution of the probabilities of the motor variable $P(G'_c^{Ag} | O_S^{Ag})$ with the number of

2157 iterations in learning, for the object $O_S^{Ag} = /gi/$, comparing the cases where learning is

2158 without (left, identical to the middle panel of Figure 16) or with (right) hyper-articulation by

2159 the Master Agent (see text).

2160