Reanalyzing neurocognitive data on the role of the motor system in speech perception within COSMO, a Bayesian perceptuo-motor model of speech communication

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

Brain and Language xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Reanalyzing neurocognitive data on the role of the motor system in speech perception within COSMO, a Bayesian perceptuo-motor model of speech communication

Marie-Lou Barnaud[a,b,c,d], Pierre Bessière[e], Julien Diard[c,d], Jean-Luc Schwartz[a,b,*]

[a] Univ. Grenoble Alpes, Gipsa-lab, F-38000 Grenoble, France
[b] CNRS, Gipsa-lab, F-38000 Grenoble, France
[c] Univ. Grenoble Alpes, LPNC, F-38000 Grenoble, France
[d] CNRS, LPNC, F-38000 Grenoble, France
[e] CNRS, SORBONNE Universités UPMC, ISIR, Paris, France

ARTICLE INFO

Keywords:
Computational modeling
Neurocognitive architecture
Perceptuo-motor interactions
Adverse conditions
Motor perturbations
Repeated transcranial stimulation
Motor representations
Speech perception

ABSTRACT

While neurocognitive data provide clear evidence for the involvement of the motor system in speech perception, its precise role and the way motor information is involved in perceptual decision remain unclear. In this paper, we discuss some recent experimental results in light of COSMO, a Bayesian perceptuo-motor model of speech communication. COSMO enables us to model both speech perception and speech production with probability distributions relating phonological units with sensory and motor variables. Speech perception is conceived as a sensory-motor architecture combining an auditory and a motor decoder thanks to a Bayesian fusion process. We propose the sketch of a neuroanatomical architecture for COSMO, and we capitalize on properties of the auditory vs. motor decoders to address three neurocognitive studies of the literature. Altogether, this computational study reinforces functional arguments supporting the role of a motor decoding branch in the speech perception process.

Introduction

The neurocognitive involvement of the speech production system in speech perception is now firmly and clearly established from a number of studies covering a range of different approaches (see a recent detailed review by Skipper, Devlin, & Lametti, 2017). These include recurrent neuroimagery evidence showing that speech perception tasks elicit brain activity in cortical and sub-cortical regions associated with speech production (e.g. Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Pulvermüller et al., 2006), particularly in the case of speech in noise (Binder, Liebenthal, Possing, Medler, & Ward, 2004; Zekveld, Heslenfeld, Festen, & Schoonhoven, 2006), speech with a foreign accent (Callan, Callan, & Jones, 2014; Wilson & Iacoboni, 2006) or multisensory inputs with some kind of incongruence (Jones & Callan, 2003; Ojanen et al., 2005). In addition to such "coactivation" data between the perceptual and motor systems in speech perception tasks, various studies displayed small but significant modifications of perceptual or neurophysiological responses after direct modulation of the motor or premotor cortex by transcranial magnetic stimulation, TMS (e.g. d'Ausilio et al., 2009; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Möttönen, Dutton, & Watkins, 2013; Sato, Tremblay, & Gracco, 2009) or indirect modulation of the motor system by repeated use (Sato et al., 2011) or sensory-motor perturbations (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014; Shiller, Sato, Gracco, & Baum, 2009).

While this increasingly large number of experimental data clearly shows that the speech perception and production systems are strongly interconnected in the human brain, they do not unambiguously tell us what is the nature or functional role of these perceptuo-motor interconnections. Furthermore, the corresponding papers generally do not discuss what computational processes could explain the patterns of perceptuo-motor coactivation or motor modulation of neurocognitive data they report.

The objective of the present study is to address these questions through the use of a computational perceptuo-motor model of speech perception, COSMO. This model has been conceived as an integral model of speech communication, enabling to simulate both speech perception and speech production in a single computational architecture (Moulin-Frier, Diard, Schwartz, & Bessière, 2015; Moulin-Frier, Laurent, Bessière, Schwartz, & Diard, 2012). The core of the COSMO architecture is based on the analysis of a speech communication process

* Corresponding author at: Univ. Grenoble Alpes, Gipsa-lab, F-38000 Grenoble, France.
E-mail addresses: marie-lou.barnaud@gipsa-lab.grenoble-inp.fr (M.-L. Barnaud), jean-luc.schwartz@gipsa-lab.gre (J.-L. Schwartz).

https://doi.org/10.1016/j.bandl.2017.12.003
Received 8 March 2017; Received in revised form 17 July 2017; Accepted 2 December 2017

in which the role of the listener and the speaker are both taken into account in the modeling of speech perception and speech production. This enables us to address computationally the specific role of the motor system in speech perception.

Importantly, a mathematical model such as COSMO, developed in an algorithmic Bayesian framework (Bessière, Mazer, Ahuactzin-Larios, & Mekhnacha, 2013), allows one to tease out *knowledge* from *computations*. Indeed, neurocognitive data often mix considerations about what might be stored in a given brain area from what kind of neural activity could be elicited by a given computation in a given task. Similarly, most Bayesian models of speech perception (e.g. Feldman, Griffiths, & Morgan, 2009; Kleinschmidt & Jaeger, 2015; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007) do not properly delineate between knowledge and computations. Indeed, such models are defined at the computational level, in the sense of Marr (1982). They provide direct models of processes: in other words, one model is developed to address any one observed process. In contrast, in a Bayesian algorithmic model, the distinction between knowledge and computations is clear and operational (Diard, 2015): one model of preliminary knowledge yields, by Bayesian inference, several mathematically consistent process models. It is then possible to compare models in relation with experimental data (e.g. Gilet, Diard, & Bessière, 2011; Laurent, Barnaud, Schwartz, Bessière, & Diard, 2017).

In the following, we will present in a first section the COSMO architecture for speech perception, and how a "COSMO Agent" can learn its parameters from stimuli supplied by a "COSMO Master". The term "Agent" throughout this paper is taken to refer to any subject equipped with an adequate cognitive system for speech communication. "COSMO Agent" refers to the computational architecture that we use to describe basic computational and representational properties of the corresponding cognitive system. A "Learning Agent" has not yet learnt the pattern of correspondence between variables enabling her/him to properly achieve speech production and perception processes. On the contrary, a "Master Agent" has already mastered her/his speech communication system and is hence able to produce speech sounds associated to the adequate linguistic units. A "Master Agent" is a tutor in a broad sense (e.g. a mother or a father for a learning infant), and the Master Agent plays the role of providing stimuli to a Learning Agent for learning the adequate behavior. We will derive three computational properties that could each provide a basis for a functional role of the motor system in speech perception: "Redundancy", "Complementarity", and "Specificity". We will also propose the sketch of a neuroanatomical architecture implementing COSMO, enabling to address more precisely neurocognitive data from the literature.

From there on, we will reinterpret three sets of neurocognitive evidence for the involvement of motor processes in speech perception. In Section 2, we will show that the "Complementarity" Property could explain why the motor system would be more involved in adverse conditions (such as noisy input or foreign accent). In Section 3, we will exploit the "Redundancy" Property to discuss why a motor perturbation would result in a perceptual bias in auditory phonetic decoding. In Section 4, we will discuss why motor activations in speech production would differ from motor activations in speech perception, and how audio-motor relationships would be represented in the frontal cortex, in light of the "Specificity" Property.

Importantly, COSMO remains at this stage a computational architecture providing a proof of concept and a basis for principled reasoning about neurocognitive representations and processes, rather than fully realistic model of speech perception able to quantitatively account for all known phenomena of speech perception. Therefore, Section 5 will be the occasion of further discussions about perspectives and challenges for the future COSMO developments.

1. COSMO, a computational model of perceptuo-motor speech perception

1.1. Principles and implementation

1.1.1. Architecture

Speech can be conceived as a perceptuo-motor process enabling a speaker to change the listener's state of knowledge. We start from a very basic communication situation in which the speaker wants to designate an "object" O_S (S for "speaker") to a listener. Objects refer in COSMO to a range of possible meanings, conflating different levels of analysis (from words up to concepts and down to phonological units). In this paper, objects will only refer to phonological entities. For this designation task, the speaker produces a motor gesture M resulting in a sound S transferred by the environment to the listener, and enabling the listener to decode the object as O_L (L for "listener"). The speaker-listener interaction may be completed by an accompanying communicative action, providing a basis for reference (Moulin-Frier et al., 2015): this is achieved in COSMO by a communication Boolean variable C so that $C = True$ if and only if $O_S = O_L$.

COSMO is based on an *internalization hypothesis* according to which the whole communication chain would be internalized in the brain of each agent (Fig. 1). COSMO therefore consists in an internal loop between variables O_S, O_L, C, M and S, connecting all the internalized variables in the agent's brain, and described in a Bayesian implementation of the joint probability distribution over the 5 variables, $P(C\ O_S\ S\ M\ O_L)$. The COSMO acronym stands for both the involved communication principle, "Communicating Objects using Sensory–Motor Operations" and the set of variables involved in the model and summarized by the probability distribution $P(C\ O_S\ S\ M\ O_L)$.

COSMO agents are then entirely defined by the $P(C\ O_S\ S\ M\ O_L)$ distribution. This distribution is decomposed in the following way:

$$P(C\ O_S\ S\ M\ O_L) = P(O_S)P(M|O_S)P(S|M)P(O_L|S)P(C|O_S\ O_L), \quad (1)$$

where $P(O_S)$ is the probability to select an object for communication, $P(M|O_S)$ is the motor repertoire, $P(S|M)$ is the internal forward model, $P(O_L|S)$ is the sensory classifier and $P(C|O_S\ O_L)$ is a coherence term evaluating the success of communication (Gilet et al., 2011). The sensory classifier $P(O_L|S)$ is actually itself the result of an inference: indeed, a sub-model stores sensory prototypes $P(S|O_L)$, from which the sensory classifier is computed through Bayesian inversion (Laurent et al., 2017).

1.1.2. Learning

Learning in COSMO is performed in the course of an interaction between a Learning Agent and a Master Agent, each characterized by the COSMO distribution defined in Eq. (1) (see Fig. 1a). The $P(O_S)$ distribution is supposed uniform, considering that all objects for the speaker are equiprobable for the sake of simplicity. The coherence term $P(C|O_S\ O_L)$ is a Dirac distribution such that $P([C = True]|O_S\ O_L) = 1$ when $O_S = O_L$. The other probability distributions $P(M|O_S)$, $P(S|M)$ and $P(O_L|S)$ are unknown for the Learning Agent, i.e. they are uniform distributions at the beginning of the learning process. In contrast, in the Master Agent, these distributions encode relevant knowledge concerning how the considered phonological units are produced and how they sound. During the interaction stage, the Master Agent randomly selects an object o, and an appropriate motor command m according to its motor distribution $P(M^{Master}|O_S^{Master})$ (using superscripts to disambiguate, whenever useful, between the Master Agent's and Learning Agent's distributions). The motor command m is transformed into a sound s by the physical environment. The sound s, together with the object o, are provided to the Learning Agent to learn the parameters of

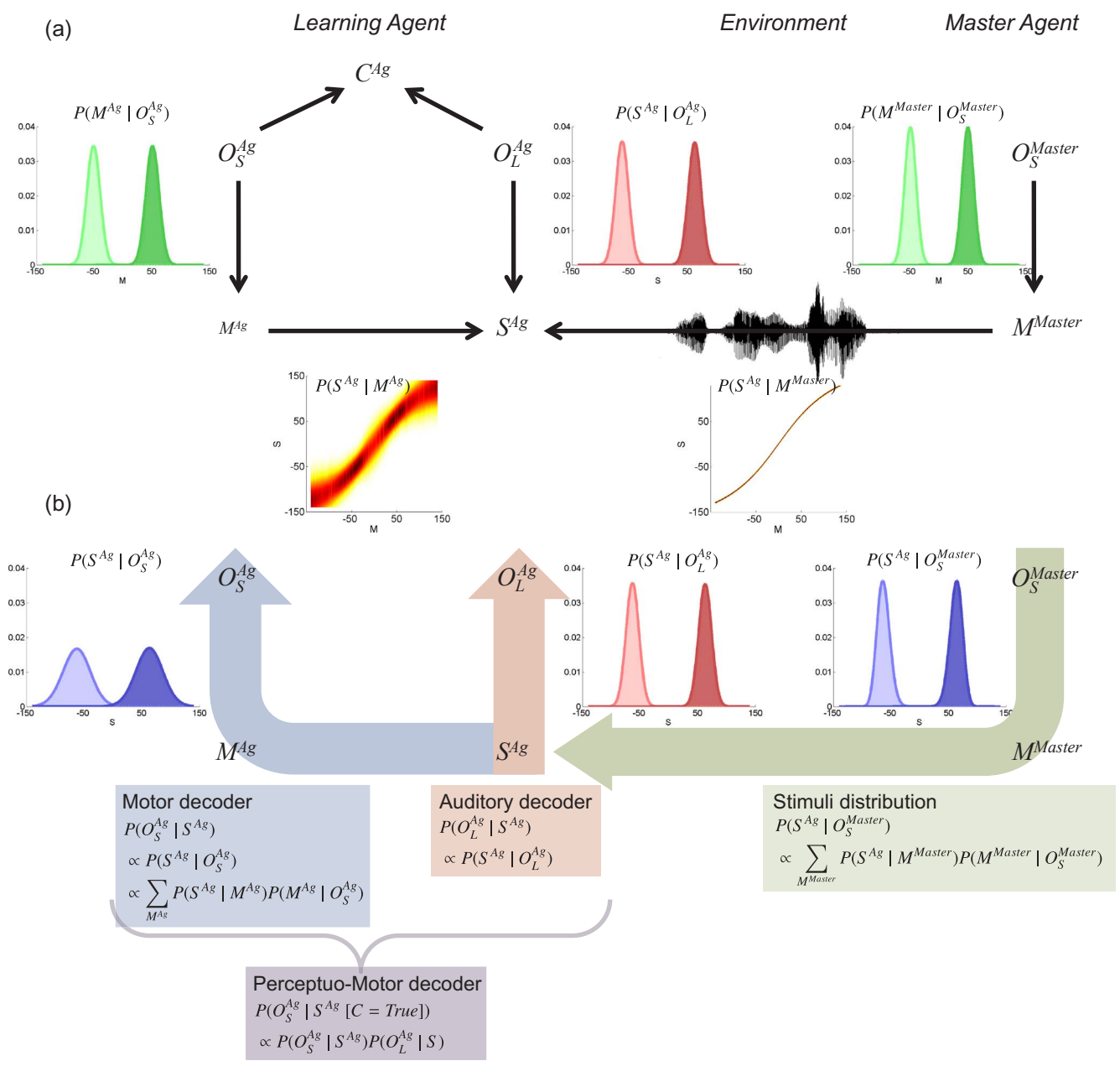

Fig. 1. The COSMO model. (a) Learning process and stored representations in the Learning Agent. The Master Agent draws motor commands according to its motor distribution $P(M^{Master}|O_S^{Master})$. They result in sounds through the physical environment distribution $P(S^{Ag}|M^{Master})$. The Learning Agent exploits (s,o) pairs provided by the Master Agent to learn its internal distributions $P(S^{Ag}|O_L^{Ag})$, $P(S^{Ag}|M^{Ag})$ and $P(M^{Ag}|O_S^{Ag})$. (b) Computations involved in a perception task. The Master Agent provides stimuli according to the distribution $P(S^{Ag}|O_S^{Master})$. Perception involves an auditory decoder – described by its inverse distribution $P(S^{Ag}|O_L^{Ag})$ – and a motor decoder – described by its inverse distribution $P(S^{Ag}|O_S^{Ag})$ – fused within a perceptuo-motor decoder according to Eq. (3).

its distributions. Parameter learning comprises three steps (Barnaud, Diard, Bessière, & Schwartz, 2015; Barnaud, Schwartz, Diard, & Bessiere, 2016).

- Setting O_L^{Ag} and O_S^{Ag}. The Learning Agent sets the values of its own variables O_L^{Ag} and O_S^{Ag} equal to the object o communicated by the Master Agent.
- Learning the sensory distribution. From the pair (s,o), the Learning Agent updates the distribution $P(S^{Ag}|O_L^{Ag})$, its repertoire of sensory prototypes.
- Learning the sensory-motor and motor distributions. The Master Agent provides no information about the motor commands associated to objects and sounds. Hence, the Learning Agent has to infer them. For this aim, the Learning Agent draws a likely motor gesture with respect to the pair (s,o) provided by the Master Agent, according to the inference: $P(M^{Ag}|[S^{Ag}=s][O_S^{Ag}=o]) \propto P(M^{Ag}|[O_S^{Ag}=o])P([S^{Ag}=s]|M^{Ag})$. Choosing a motor command according to this distribution and then producing it generates a sound s, possibly different from s. The Learning Agent updates the distribution $P(S^{Ag}|M^{Ag})$ from the pair (m,s) and the distribution $P(M^{Ag}|O_S^{Ag})$ from the pair (m,o).

Importantly, this process has the interesting property that, at the beginning of the learning stage, since the $P(S^{Ag}|M^{Ag})$ and $P(M^{Ag}|O_S^{Ag})$ distributions are uniform, the Learning Agent explores its motor space. This random "babbling" exploration is progressively replaced by an exploration process more and more focused on the Master Agent's production (Barnaud et al., 2016), in agreement with the "babbling

drift" observed in infant's sensory-motor exploration (e. g. de Boysson-Bardies, Halle, Sagart, & Durand, 1989; de Boysson-Bardies, Sagart, & Durant, 1984). Sensory-motor exploration hence combines some amount of general knowledge and some amount of information focused on the adequate sensory-motor relationships exploited by the Master.

At the end of the learning process, the Learning Agent has learned the $P(M^{Ag}|O_S^{Ag})$, $P(S^{Ag}|M^{Ag})$ and $P(S^{Ag}|O_L^{Ag})$ distributions, hopefully close but still different from those of the Master Agent. This is displayed in Fig. 1a, showing that learning enables the model to tune a set of stored distributions necessary for further communication tasks (see next section): this is the Learning Agent's stored state of knowledge.

1.1.3. Perception and production tasks

In COSMO, production tasks consist in asking the COSMO model questions of the form $P(M|O)$ that is, what motor gestures should be selected to designate a given object O? Conversely, perception tasks, which will be the focus of the present paper, consist in asking questions of the form $P(O|S)$, that is, what objects can be inferred from a given sensory input? (Notice that only auditory inputs will be considered in this paper, the case of multisensory inputs will be further discussed in Section 5.3).

COSMO enables to simulate auditory, motor or perceptuo-motor theories of speech perception by adapting the probabilistic question $P(O|S)$ at hand. The auditory and motor theories correspond respectively to selecting O_L vs. O_S in the $P(O|S)$ question. That is, in auditory theories, the pivot is the listener, and perception consists in directly relating the sensory input S (typically the sound) to the category for the listener, O_L, asking the question $P(O_L|S)$. The solution is provided by Bayesian inversion of the stored distribution $P(S|O_L)$. This computation involves no information about motor processes, as proposed by e.g. Diehl, Lotto, and Holt (2004).

In the COSMO implementation of the motor theory, the pivot is the speaker and perception consists, as posited by e.g. Liberman and Mattingly (1985), in recovering the speaker's aims from the incident sound, by computing $P(O_S|S)$. Bayesian inference in COSMO provides the following implementation of the Motor Theory of Speech Perception (Laurent et al., 2017; Moulin-Frier et al., 2015):

$$P(O_S|S) \propto \sum_M P(M|O_S)P(S|M) \tag{2}$$

This is an implementation of the Analysis-by-Synthesis process estimating a motor cause M from a sensory input S (Bever & Poeppel, 2010; Halle & Stevens, 1959). This Bayesian implementation results automatically from the model definition and Bayesian inference. It can be interpreted as performing the acoustic-to-articulatory inversion (e.g. Bailly, 1997) by a summation over all possible motor configurations of the product of factors $P(M|O_S)$ and $P(S|M)$ both stored in the Agent's state of knowledge.

Finally, a perceptuo-motor model of speech perception (such as proposed by e.g. Schwartz, Basirat, Ménard, & Sato, 2012; Schwartz, Boë, & Abry, 2007; Skipper, van Wassenhove, Nusbaum, & Small, 2007) is implemented in COSMO by computing $P(O_S|S\ [C = True])$, or, indifferently $P(O_L|S\ [C = True])$, as they both yield identical expressions:

$$\begin{aligned} P(O_S|S\ [C = True]) &= P(O_L|S\ [C = True]). \\ &\propto P(O_L|S) \sum_M P(M|O_S)P(S|M). \\ &\propto P(O_L|S)P(O_S|S). \end{aligned} \tag{3}$$

This equation means that decoding consists in attempting to find an object corresponding both to the object for the listener O_L and the object for the speaker O_S, imposing that they have the same value thanks to the constraint $C = True$. Bayesian inference shows that this is realized in COSMO by a probabilistic fusion of auditory decoding $P(O_L|S)$ and motor decoding $P(O_S|S)$.

In the framework of the perceptuo-motor theory we developed since a number of years (Schwartz, Abry, Boë, & Cathiard, 2002; Schwartz et al., 2007, 2012), we suppose that speech perception is indeed perceptuo-motor, hence modeled in COSMO by Eq. (3). Therefore, a speech perception task involves three pieces of computational processes displayed in Fig. 1b:

(1) auditory decoding by computing the distribution $P(O_L|S)$ from the Bayesian inversion of the auditory repertoire $P(S|O_L)$;
(2) motor decoding by computing the distribution $P(O_S|S)$ from the motor repertoire $P(M|O_S)$ and the sensory-motor distribution $P(S|M)$ thanks to Eq. (2);
(3) fusion of the auditory and motor routes in the final decoding stage by computing $P(O_S|S\ [C = True]) = P(O_L|S)P(O_S|S)$ according to Eq. (3).

1.2. Three properties ensuring the functional role of the motor system in speech perception

1.2.1. Redundancy

The sensor fusion process in Eq. (3) enables the decoding system to benefit from two independent inference processes, respectively $P(O_L|S)$ and $P(O_S|S)$. This results in a first functional role for the motor decoding route in COSMO, through *redundancy*. Combination of auditory and motor decoding should ensure a gain in robustness provided that the fusion process is efficient, which is the case in a Bayesian implementation like the one in COSMO (see e.g. Ernst & Banks, 2002; Massaro, 1987).

1.2.2. Complementarity

Motor decoding and auditory decoding structurally differ in COSMO. Indeed, learning the auditory distribution $P(S|O_L)$ is a direct process which requires no additional inference once the Master Agent provides pairs of objects and stimuli (o,s). The auditory decoding process then relies on a simple Bayesian inversion of this stored distribution. On the contrary, learning the motor decoding route $P(O_S|S)$ requires an additional inference on the hidden motor variable M through Eq. (2).

A consequence (see Laurent et al., 2017; see also Section 2) is that auditory decoding is perfectly fitted to the sensory distribution of the Master Agent, and hence more efficient than motor decoding to process clean stimuli. However, noisy stimuli or stimuli differing from prototypical values (e.g. produced in a different accent) do not fit well with the probability distribution learned in $P(S|O_L)$. Conversely, motor decoding appears able to process such stimuli more efficiently because its learning is more difficult and thus results in wider variance. In summary, we showed (Laurent et al., 2017) that the auditory decoder performs better than the motor decoder without noise, but that the motor decoder outperforms the auditory decoder in adverse conditions. Hence, auditory and motor decoding are *complementary*.

1.2.3. Specificity

Finally, apart from this *structural* complementarity, detailed COSMO implementation in the case of CV syllables with C a plosive and V an oral vowel showed that there might also exist an *informational* complementarity. Indeed, simulations with a "COSMO-syllables" model (Laurent, Schwartz, Bessière, & Diard, 2013; Laurent et al., 2017) showed that while vowels are better characterized in auditory than in motor terms, consonant place of articulation is poorly defined in auditory terms but well characterized in the motor domain. This is in line with the classical claims about articulatory invariance for consonant place of articulation by defenders of the Motor Theory of Speech Perception (Liberman & Mattingly, 1985). This introduces a *specificity* property associated to the motor decoding route in relation with the phonetic content of the speech input.

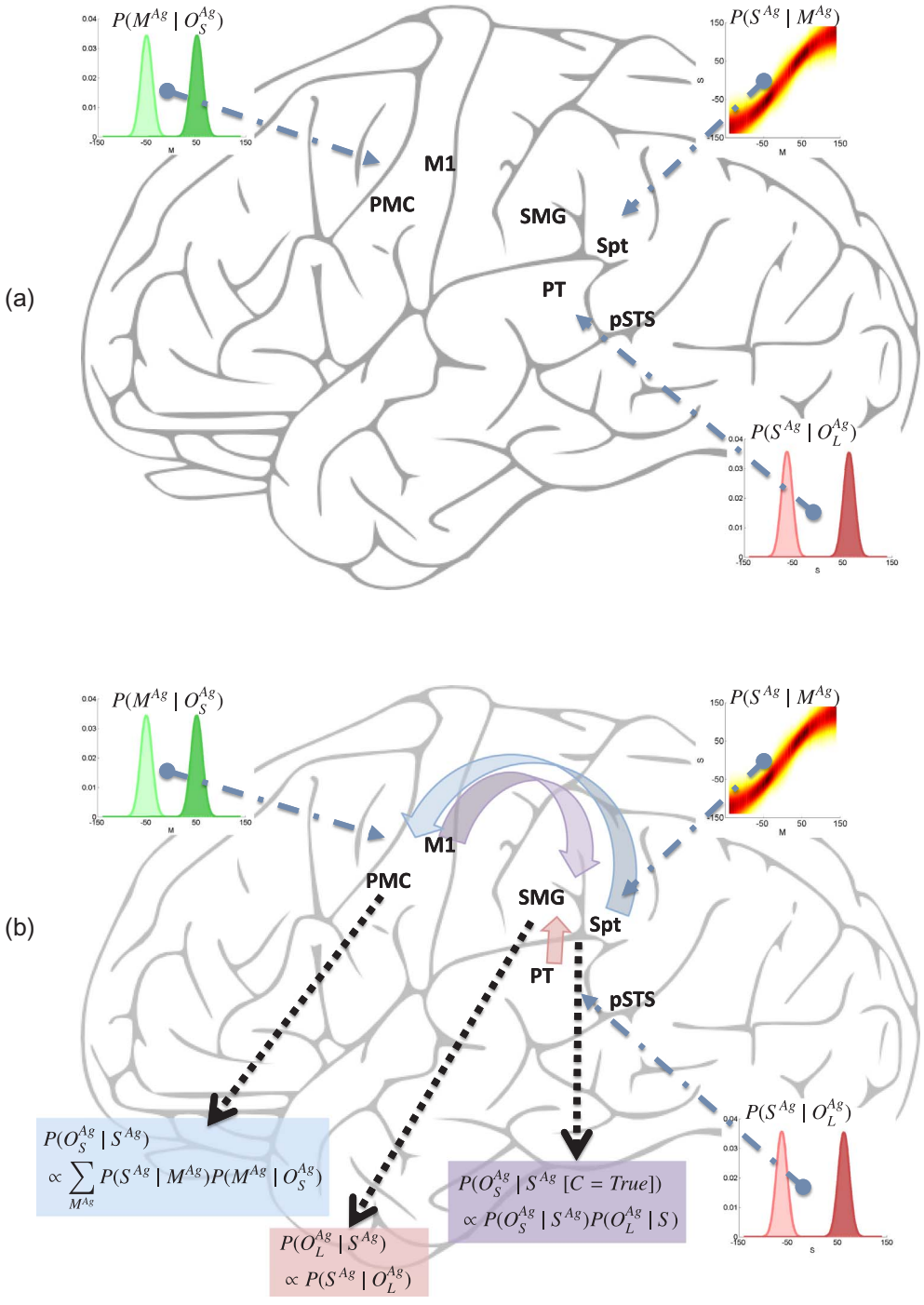

Fig. 2. Proposal for a neuroarchitecture for the COSMO model. (a) Possible localizations for the learned distributions $P(S^{Ag}|O_L^{Ag})$ in the superior posterior part of the temporal cortex (planum temporale, PT; or posterior superior temporal sulcus, pSTS), $P(S^{Ag}|M^{Ag})$ in area Spt (Sylvian parietal temporal) and $P(M^{Ag}|O_S^{Ag})$ in the motor cortex (primary motor cortex M1 or premotor cortex PMC). (b) Activity related to the computational processes for perception, with auditory decoding $P(O_L^{Ag}|S^{Ag})$ in pSTS or SMG, motor decoding $P(O_S^{Ag}|S^{Ag})$ in PMC and perceptuo-motor fusion for final decision in superior temporal (possibly area Spt) or inferior parietal (supramarginal gyrus SMG) regions.

1.3. A sketch of neuroanatomical architecture for COSMO

To be able to address neurocognitive data with COSMO, we now need to go one step further and propose some possible brain implantation of the COSMO computational architecture. This is tricky, since the COSMO architecture is a generic description of algorithmic processes, and thus is underspecified with respect to coding and implementation. For instance, all equations above refer to variables S and M, and the properties above result from their relations in the global COSMO architecture. But, in fact, variable S could also be structured, and refer to a hierarchy of processing and representations from the incoming sound, just as the variable M could refer to a hierarchy of controls and coordination before final motor-command implementation. As a result, the precise localization of COSMO sensory representations in relation to the temporal cortex and motor representations in relation to the frontal cortex could be sometimes unspecified in the remaining part of this paper (though see some more detailed discussion in Section 5.1).

Still, it is possible to adapt some inspiring models of the literature, to derive a possible sketch of neuroanatomical architecture (Fig. 2). This architecture starts from the organization of the dorsal route in the dual-stream model of the functional anatomy of language proposed by e.g. Hickok and Poeppel (2007) (see also Rauschecker & Scott, 2009). The dorsal route connects an auditory processing network in the temporal lobe (primary and secondary auditory cortex, planum temporale PT, posterior superior temporal sulcus pSTS) to an articulatory network in the frontal lobe (inferior frontal gyrus IFG, premotor cortex PMC, anterior insula, primary motor cortex M1) through a sensory-motor

interface in the inferior parietal lobule (down to the parieto-temporal boundary in the Sylvian fissure within the planum temporale, Spt). While this sensory-motor route is conceived by Hickok and Poeppel (2000, 2004, 2007) as essentially dedicated to learning processes, it is exploited by Skipper et al. (2007) for implementing an Analysis-by-Synthesis process in which a feedforward connection from posterior superior temporal areas would lead to motor goal predictions in the IFG (pars opercularis). Then, a feedback connection through premotor and motor cortices would generate a motor plan and finally, by efference copy, a sensory prediction in auditory areas, combined with the initial auditory processing to provide final decoding.

The same structure may be proposed as a neuroanatomical architecture for COSMO. The auditory repertoire $P(S|O_L)$ would be stored in the posterior superior temporal areas. The sensory-motor distribution $P(S|M)$ would be stored along the posterior-to-anterior pathway, possibly in the sensory-motor interface provided by area Spt. Finally, the motor repertoire $P(M|O_S)$ would be stored in frontal areas, probably at various levels from distributions of single articulatory variables in the primary motor cortex to motor programs associated to phonological units in the premotor cortex (Fig. 2a).

Remark that these neuroanatomical hypotheses concern distribution storage, i.e. where probability distributions would be memorized in the brain. Then, computations for speech perception would "propagate through" these representations, recruiting them, and may thus also be localized without requiring further assumptions (Fig. 2b). Firstly, auditory decoding $P(O_L|S)$ would be obtained by simple Bayesian inversion of the auditory repertoire $P(S|O_L)$ – possibly in the pSTS (Hickok & Poeppel, 2007); or at the level of the supramarginal gyrus SMG (e.g. Jacquemot & Scott, 2006; Paulesu, Frith, & Frackowiak, 1993). Secondly, motor decoding $P(O_S|S)$ – or its Bayesian inverse $P(S|O_S)$ – would be achieved by a summation and multiplication process in frontal areas, probably in the premotor cortex, according to Eq. (2). Finally, the result of motor decoding would be sent back to temporal regions, possibly through area Spt (Hickok, Okada, & Serences, 2009), for the fusion of auditory and motor decoding $P(O_S|S[C = True]) = P(O_L|S)P(O_S|S)$, yielding the final phonological decision.

Therefore, at this stage, we remain quite unspecific in terms of precise localizations, just playing with three coarse neuro-anatomical poles, respectively in temporal areas (PT, pSTS), in the inferior parietal lobule or at the temporo-parietal junction (SMG, Spt) and in frontal areas (M1, PMC). We will come back to the relationship between COSMO and neuroanatomical and neurocognitive data in more detail in the general discussion in Section 5.

2. Why should the motor system be more involved in adverse conditions (such as noisy input or foreign accent)?

As mentioned previously, it has been recurrently shown that the involvement of parieto-frontal areas associated with speech production is increased, in absolute or relative terms, compared to temporal areas associated to auditory processing, when human subjects process speech in adverse or unusual conditions. Surprisingly, not much functional interpretation is provided for this fact. It is generally considered that this could be related to motor simulation, e.g. in the framework of Analysis-by-Synthesis processes (see a review in Skipper et al., 2017). But, to our knowledge, no precise proposal has been introduced to explain why Analysis-by-Synthesis would provide a gain in accuracy in noise or atypical stimuli.

This is where the Complementarity Property introduced previously might shed some light. This property explains how motor decoding would be more robust in noise, as illustrated in Fig. 3. In this figure, as along all this study, we adopt a simplifying approach in which we search a minimal formal description of the processes at hand. The aim is to make both simulations simple and interpretation clear in terms of possible underlying neurocognitive processes. Hence, we consider a simple situation where sensory and motor variables would be one-dimensional, related by a monotonous sigmoidal relationship $S = sig(M)$. The Master Agent communicates about two objects, o^+ and o^-, both associated with a Gaussian probability distribution in the motor space, so that $P(M^{Master}|[O_S^{Master} = o^+]) = \mathcal{N}(\mu^+,\sigma)$ and $P(M^{Master}|[O_S^{Master} = o^-]) = \mathcal{N}(\mu^-,\sigma)$, with $\mathcal{N}(a,b)$ denoting a Gaussian probability distribution with mean a and standard deviation b (see insert plot in Fig. 1a). Motor values are transformed by the sigmoid into sensory values, resulting in probability distributions $P(S^{Ag}|[O_S^{Master} = o^+])$ and $P(S^{Ag}|[O_S^{Master} = o^-])$. The Learning Agent learns all its distributions from (o,s) pairs provided by the Master Agent as explained in Section 1.1.2. From these data, learning the sensory distribution $P(S^{Ag}|O_L^{Ag})$ is straightforward and efficient while learning $P(S^{Ag}|M^{Ag})$ and $P(M^{Ag}|O_S^{Ag})$ distributions involves a more complex process of motor inference (see displays of the learned distributions in the insert plots in Fig. 1a). It appears that learning of $P(S^{Ag}|M^{Ag})$ is accurate in the regions of S values provided by the Master Agent, but less so in other regions. Indeed, the babbling process, in its initial wandering phase, enables the agent to learn some basic approximation of $P(S^{Ag}|M^{Ag})$ along the (S,M) space (Barnaud et al., 2015, 2016).

As a consequence, the $P(O_S^{Ag}|S^{Ag})$ distribution, involving both the $P(S^{Ag}|M^{Ag})$ and $P(M^{Ag}|O_S^{Ag})$ distributions, is indeed wider than the $P(O_L^{Ag}|S^{Ag})$ one obtained by simple Bayesian inversion of the learned distribution $P(S^{Ag}|O_L^{Ag})$ (see the associated sensory distributions in Fig. 3a and c). Therefore, when a "prototypical" stimulus (green line in Fig. 3a–d), close to mean values μ^+ or μ^-, is sensed, it is more accurately decoded by the auditory route (see Fig. 3b vs. d, and e, left). However, less prototypical stimuli (brown line in Fig. 3a–d) fall out of the decoding abilities of the narrow auditory distribution. Indeed, all sensory and motor distributions in COSMO are truncated and set at a floor value under a given threshold. This avoids cognitively implausible numerical precision of probability distributions in neural assemblies, and ensures that probability distributions, truncated in this manner, degenerate outside of their "competence domains". On the contrary, since the motor route is wider, a larger portion of the sensory domain is above the threshold and thus it is able to process such "exotic" stimuli (see Fig. 3b vs. d and e, right). This is why auditory and motor decoding would be complementary: auditory decoding would be a *narrow-band* process (narrowly focused in the set of stimuli provided by the environment in the learning process) while motor decoding would be a *wide-band* process able to deal with unusual stimuli.

Hence, we now have at our disposal a functional mechanism possibly explaining why motor decoding should be more involved in adverse conditions. From there on, we can propose two possible explanations for the observed increase in BOLD activity in motor areas in fMRI data during speech perception in noise or adverse conditions. Firstly, there could exist some kind of automatic control of neural activity based on the compared efficiencies of the auditory and motor decoding processes. In clean or canonical conditions, auditory decoding is efficient, while motor decoding is poorer. Therefore, it can be assumed that as soon as an auditory decision is available, the transfer of neural information from temporal to frontal regions is stopped – hence frontal activity is small. On the contrary, for noisy or non-prototypical inputs, auditory decoding is less efficient, hence more neural propagation would occur towards frontal regions for motor decoding, until sufficient evidence would be acquired, enabling decision to occur in good conditions. This mechanism would produce an increase in the amount of frontal activity in the motor regions with e.g. increasing levels of noise, as can be seen in various studies such as Binder et al. (2004), Wilson and Iacoboni, (2006) or Zekveld et al. (2006).

Another possibility is that the call to motor decoding is controlled by attentional processes, according to which the weight of motor decoding in the sensory-motor fusion process – and hence the amount of neural activity in frontal regions – would be actively increased with noise to improve decoding. This is proposed by e.g. d'Ausilio, Bufalari, Salmas, and Fadiga (2012) (see also Davis & Johnsrude, 2007) who

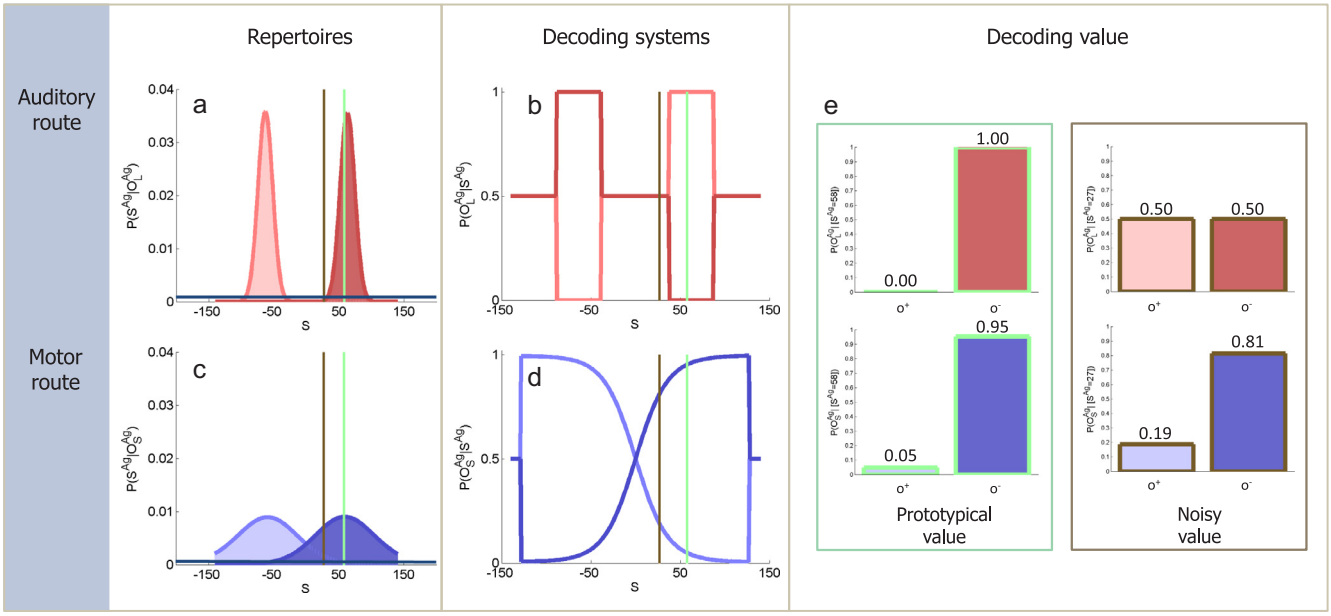

Fig. 3. (a and c) Sensory distributions respectively associated to the auditory route $P(S^{Ag}|O_L^{Ag})$ and motor route $P(S^{Ag}|O_S^{Ag})$. (b and d) Associated decoding systems $P(O_L^{Ag}|S^{Ag})$ and $P(O_S^{Ag}|S^{Ag})$. The truncation threshold makes decoding inefficient for stimuli with too small $P(S|O)$ values (see text). The brown vertical line in (a–d) corresponds to the noisy stimulus and the green vertical line corresponds to the prototypical stimulus. (e) Results of the decoding systems for a prototypical stimulus (left panel) and a noisy stimulus (right panel). The two probabilities $P(O|S)$ for (o^+) and (o^-), for the auditory route ($P(O_L^{Ag}|S^{Ag})$, top row) and the motor route ($P(O_S^{Ag}|S^{Ag})$, bottom row), are displayed for a prototypical and a noisy stimulus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

applied perturbations of frontal areas by TMS (see next section) and studied their effect on reaction times for discriminating between two categories (a labial vs. a dental plosive). Interestingly, they showed that TMS applied to the lips vs. the tongue resulted in selective acceleration of responses to the labial vs. dental plosives, but only in noisy conditions, whereas reaction times were typically identical for clean and noisy stimuli without perturbation. They conclude that this is rather in favor of an attention-driven process. In fact, the two alternative mechanisms are not exclusive, and could well play a role together; the current evidence and models do not allow us to unambiguously decide between these different hypotheses.

In conclusion, the Complementarity Property provides a possible functional explanation for the importance of the motor route for speech decoding in adverse conditions: the auditory process would be narrowly focused on the learning set and hence optimal for such learned stimuli, while the motor process would be less narrowly tuned and hence more important in noise. The corresponding neuroanatomical architecture in Fig. 2, together with the two variants introduced here above, provide a possible neuroanatomical mechanism explaining the pattern of neural activity in temporal vs. frontal areas in relation with the prototypical vs. non-prototypical nature of the sensory inputs to process.

3. Why would a motor perturbation result in a perceptual bias in auditory phonetic decoding?

We now address a second piece of neurocognitive evidence, in which perturbations of regions in frontal areas produce modifications in speech perception tasks. We will consider in the following a specific study, described in Möttönen and Watkins (2009), and show how COSMO simulates their data. We will finally discuss how these simulations could be extended or adapted towards other studies using similar experimental paradigms.

Möttönen and Watkins (2009) applied repeated Transcranial Magnetic Stimulation (rTMS) to produce a temporary disruption of the "lip region" on the left primary motor cortex. Disruption was produced by 15-min rTMS stimulation at low frequency. It was applied on two regions in the primary cortex, respectively related to the lips and to the hands, for control. The efficiency of the disruption was controlled by measuring motor evoked potentials, before or after stimulation. The perceptual task consisted in the categorization and discrimination of stimuli within synthetic continua, e.g. between /ba/ and /da/, generated by formant synthesis (Klatt, 1980). Results showed that both categorization and discrimination were impaired after stimulation in the lip area: the categorization slope was decreased and the discrimination between pairs of stimuli respectively on one and the other side of the categorization boundary was poorer (see below for explanations about the quantitative assessment of categorization and discrimination).

Results of this experiment can be simulated in COSMO, thanks to its first property of "Redundancy". Indeed, since phonetic decoding in COSMO is perceptuo-motor, the motor decoder does play a role in phonetic perception. From the cortical architecture in Fig. 2, motor decoding would involve the computation of $P(O_S|S)$ – or its Bayesian inverse $P(S|O_S)$ – in the frontal area, precisely where the perturbation is applied. According to Eq. (2), $P(O_S|S)$ involves the distributions $P(S|M)$, possibly stored in Spt in the parietal cortex, and $P(M|O_S)$, possibly stored in frontal areas. Modification of the distribution $P(M|O_S)$ by rTMS applied to the motor cortex should play a role in the final decoding output expressed by Eq. (3), hence it should modify categorization and discrimination. Let us explain this mechanism in more detail.

Firstly, we must specify what information could be stored in the $P(M|O_S)$ probability distribution. We will do this still in the simplifying process we adopted all along the paper.

We display in Fig. 4a in continuous lines what could be such a $P(M|O_S)$ distribution for two "objects" /ba/ and /da/, in a two-dimensional motor space involving lips and tongue, i.e. $M = (M_{lips}, M_{tongue})$. Of course, these motor dimensions could be rather combinations of dimensions (since both lips and tongue are in fact controlled by more than one motor parameter) or just some part of the whole lips and tongue systems (e.g. lip closing/opening or tongue tip elevation). The green ellipses respectively display $P(M|[O_S = \text{“}ba\text{”}])$ and $P(M|[O_S = \text{“}da\text{”}])$, assuming that for /ba/ what matters is a precise lip configuration, the tongue being largely unspecified, and vice versa for /da/. We quantify the difference in precision of control in each

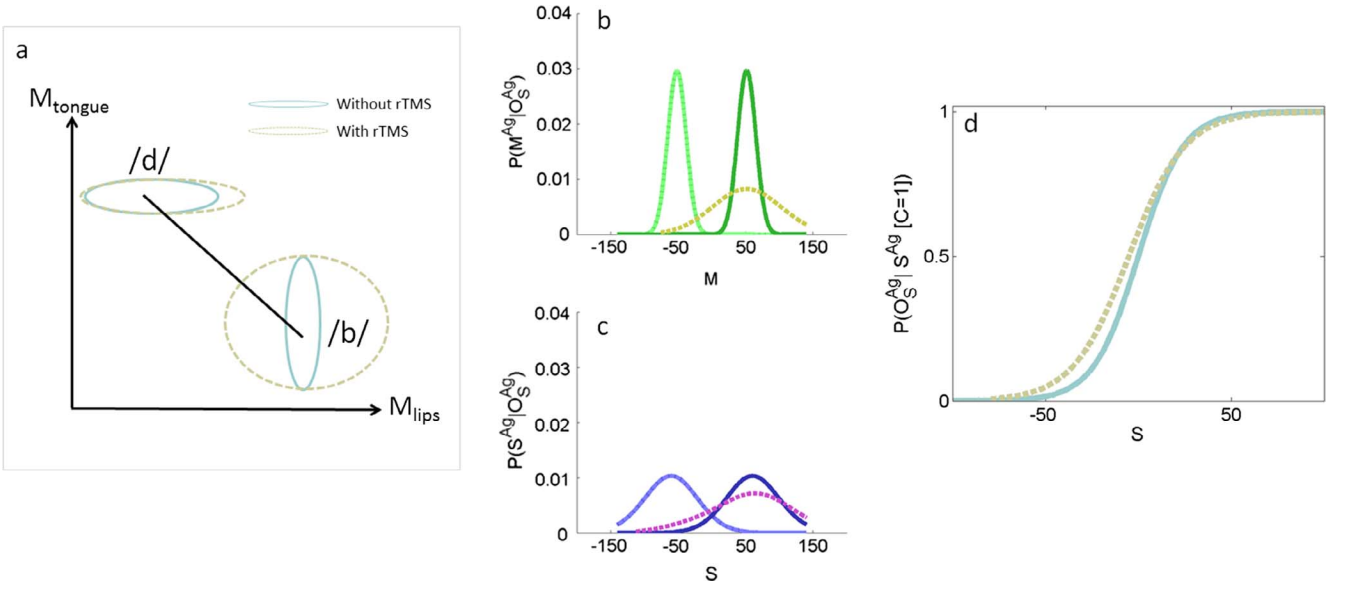

Fig. 4. (a) Schematic representation of the disruption of the lips in a two-dimensional space where one dimension corresponds to the lips and one dimension corresponds to the tongue. (b) Motor system $P(M|O_S)$ before the disruption (continuous line) and after the disruption of the second object (dotted line). (c) Sensory distribution associated to the motor route $P(S|O_S)$ before the disruption (continuous line) and after the disruption of the second object (dotted $P(M|O_S)$ line). (d) Decoding system $P(O_S|S[C = True])$ before the disruption (continuous line) and after the disruption (dotted line).

dimension by differences in variances of two-dimensional Gaussian distributions.

Then, we must specify the effect of an rTMS perturbation in the primary motor cortex M1. We suppose that applying rTMS in the lips region of M1 results in a degradation of the representation of the lips, associated to an increase of the variance of $P(M|O_S)$ for both objects along the M_{lips} dimension (dotted ellipses in Fig. 4a). It is difficult at this stage to propose how a perturbation applied to the lips region in M1 would precisely modify $P(M|O_S)$. It could occur directly, by degrading the representation of the lips parameter, hence making the $P(M_{lips}|O_S)$ distribution less accurate for all objects O_S. But it could also be envisioned that a perturbation in M1 could diffuse to the premotor cortex, as mentioned as a possibility by Möttönen and Watkins (2009), and that the variance increase would operate at this level.

So, we now have two distributions of $P(M|[O_S = \text{“}ba\text{”}])$ and $P(M|[O_S = \text{“}da\text{”}])$, respectively in normal and rTMS perturbed conditions. These distributions are bi-dimensional in the present simplified reasoning. We also consider a resulting 2D acoustic space. The principle of a /ba/-/da/ continuum as the one synthesized in Möttönen and Watkins (2009) consists in selecting a one-dimensional pathway in the multidimensional acoustic space. Indeed, /ba/-/da/ continua can be generated as specific trajectories in the (F2-F3) acoustic space of second and third formant values F2 and F3 estimated at the onset of the consonant-to-vowel trajectory, (see Serniclaes & Sprenger-Charolles, 2003). We display in Fig. 4a such a possible trajectory from /ba/ to /da/, defining a 1D continua which results in a probability distribution $P(M|O_S)$ such as the one displayed in Fig. 4b, along a parameter M summarizing the lips-to-tongue pathways. In this 1D space, we assume that distributions $P(M|O_S)$ have the same variance for the two objects /ba/ and /da/ in normal conditions. In the scenario proposed previously, the rTMS perturbation would hence result in an increase in the variance of the $P(M|O_S)$ distribution for the /ba/ category in the lips dimension, while the $P(M|O_S)$ distribution would stay essentially unchanged for /da/, as displayed in Fig. 4b by the distributions in dotted lines.

Then, the reasoning about categorization and discrimination is straightforward. Indeed, the change in $P(M|O_S)$ for /ba/ would result in an increase in the variance in the distribution $P(S|O_S)$ (see Fig. 4c). This would be reflected in the categorization behavior at the output of COSMO. Categorization is obtained here by computing $P(O_S|S[C = True])$, resulting in a product of the factors $P(O_S|S)$ and $P(O_L|S)$. $P(O_L|S)$ is unchanged by the perturbation, but $P(O_S|S)$ is modified because of the changes in $P(M|O_S)$. This results in a decrease in categorization slope (Fig. 4d), in agreement with the data in Möttönen and Watkins (2009, see their Table 1, p. 9822).

From that basis, discrimination is then computed by considering two stimuli s_1, s_2 at different positions along the 1D S dimension in Fig. 4d. Discrimination between s_1 and s_2 is supposed to be entirely due to the probability that they correspond to different classes (see e.g. Pollack & Pisoni, 1971; Repp, 1984) defined by the following equation:

$$\begin{aligned} P_{discr} = & P([O = \text{“ba”}]|[S = s_1][C = True]) \\ & P([O = \text{“da”}]|[S = s_2][C = True]) \\ + & P([O = \text{“da”}]|[S = s_1][C = True]) \\ & P([O = \text{“ba”}]|[S = s_2][C = True]). \end{aligned}$$

By taking a pair of stimuli providing a P_{discr} value equal to 0.85 in the normal condition, the value decreases to 0.75 in the simulated rTMS condition. This is compatible with the experimental values provided by Möttönen and Watkins (2009, see their Table 2, p. 9822), though, of course, values in the present simulations are just illustrative of the qualitative behavior of our model and have no pretention to be best-fit values.

Hence, COSMO, together with the underlying neurocognitive assumptions in Section 1 (for architecture) and in Section 3 (for the role of rTMS), may provide a computational interpretation of rTMS studies such as the one in Möttönen and Watkins (2009). The Redundancy Property is at the basis of the interpretation, Eq. (3) providing the core simulation principle, incorporating Bayesian fusion between the motor and auditory decoding processes. Similar reasoning could be applied to other studies by the same group on the reduction in discrimination and EEG response (Mismatch Negativity, MMN) in various kinds of motor disruption by repeated stimulation of the motor cortex (Möttönen et al., 2013; Rogers, Möttönen, Boyles, & Watkins, 2014). Conversely, perturbations by double-pulse TMS by d'Ausilio et al. (2009, 2012) rather produce a gain in categorization speed and accuracy for the category corresponding to the applied perturbation. In the scenario presented here, this would be simulated by reducing the adequate variance rather than increasing it. Indeed, the categorization slope would then become steeper with rTMS. The increase in categorization steepness would

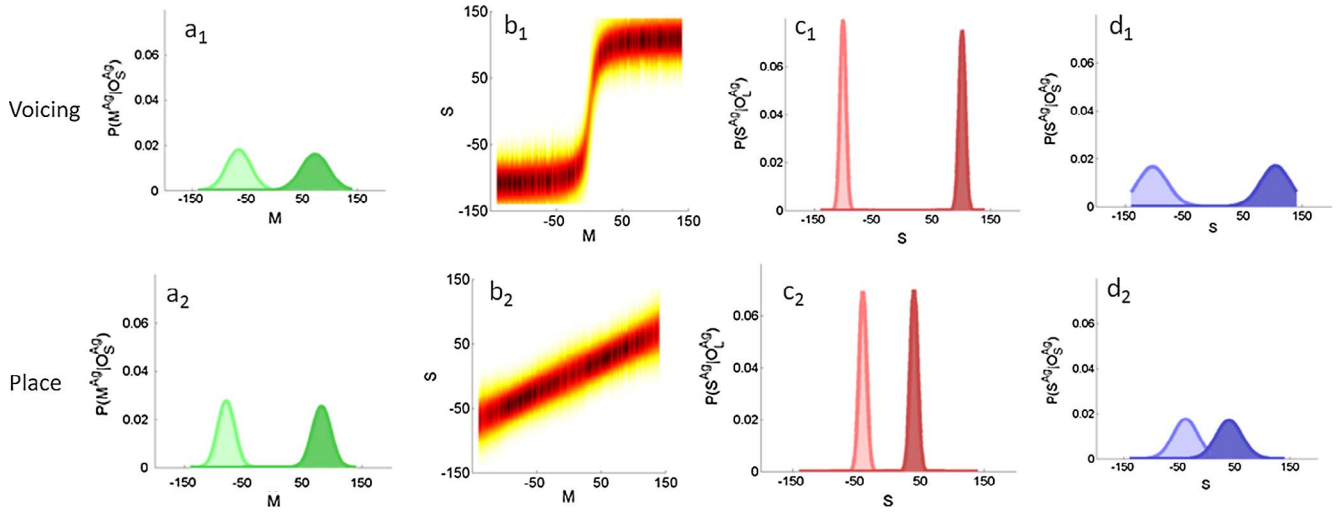

Fig. 5. Motor and sensory representations of the voicing and place dimensions in the COSMO simulations (a) Motor system $P(M|O_S)$ for the $M_{voicing}$ dimension (a_1) and the M_{place} dimension (a_2). (b) Sensory-motor system $P(S|M)$ for the voicing dimensions (b_1) and the place dimensions (b_2). (c) Sensory repertoire $P(S|O_L)$ for the voicing dimension $S_{voicing}$ (c_1) and the place dimension S_{place} (c_2).

occur in the region of the object with reduced variance, resulting in increasing categorization performance for this object, which is indeed what is observed in such experiments.

Of course, the fusion process at work in Eq. (3) remains compatible with the Complementarity Property discussed in the previous section. Indeed, if the auditory or the motor route provide a more vs. less reliable information for the decoding task, their weight in the Bayesian fusion process will be modulated accordingly (Ernst & Banks, 2002). This is the reason why motor perturbation does result in a change in phonetic decoding only when the auditory task is perturbed, either by noise (d'Ausilio et al., 2009, 2012) or by considering ambiguous synthetic stimuli around the categorization boundary (Möttönen & Watkins, 2009; Möttönen et al., 2013; Rogers et al., 2014).

In conclusion, the Redundancy Property, together with its computational implementation in Eq. (3), provides the basis for interpreting the results of various studies where a perturbation of motor areas produces a change in categorization or discrimination in an acoustic phonetic decoding task. The underlying assumption is that the motor perturbation would result in modifying the $P(M|O_S)$ distribution, leading to modification in the final decoding process $P(O_S|S[C = True])$.

4. Why would a motor representation in speech production differ from a motor representation in speech perception – and how could audio-motor relationships be represented in the frontal cortex?

The last contribution in this paper concerns the recent paper by Cheung, Hamiton, Johnson, and Chang (2016), which compares neurophysiological responses to speech production and speech perception tasks. In this study, the authors describe the activity of high-density electrode grids, providing high spatial and temporal resolution, in simple speech production and perception tasks.

Their main result concerns the way "auditory" regions (within the superior temporal gyrus STG) and "motor" regions (within the sensory-motor cortex SMC) respond in two basic conditions: aloud production vs. passive perception of CV syllables with V the vowel /a/ and C one of 8 consonants in American English /b d g p t k s ʃ/. It appears that, while the activity in the SMC region is somatotopically organized in relation with speech articulators in the production task, its organization is less clear and different in the perception task. Importantly, the authors observe that the pattern of activity in SMC in the perception task is more similar to the pattern in STG in the same task, than to the activity in SMC in the production task. They conclude that: "motor cortex does not contain articulatory representations of perceived actions in speech, but rather, represents auditory vocal information" (p. 1).

In COSMO, the patterns of activity in production in the motor cortex and in perception in the motor and in the auditory cortex actually correspond to *three* different probability distributions. As mentioned in Section 1.1.3, production tasks correspond to distributions of the type $P(M|O)$ and more precisely to distribution $P(M|O_S)$. As discussed in detail in this paper, perception tasks would involve two routes, an auditory route associated to distribution $P(O_L|S)$, or its Bayesian inverse $P(S|O_L)$, and a motor route associated to distribution $P(O_S|S)$, or its Bayesian inverse $P(S|O_S)$. We will show that the data in Cheung et al. (2016) could actually correspond to these three distributions. Possible similarities between these distributions may be reported, in agreement with neural data – but also differences that might not be noticed in that paper because of the inherent limitations of the experimental material.

Our interpretation of these data will capitalize on the third property mentioned previously, that is "Specificity", associated to the coding of phonetic information in the auditory vs. motor routes. This property applies well to the present data. To make this clear, let us enter more in detail in the phonetic content of the speech material used in this work. More precisely, we will focus on two features at the core of the study, which are place of articulation and voicing. Phonetic knowledge about the articulatory and acoustic content of these features yields a simplified but plausible scenario for understanding neuronal data. We assume that the motor control of the place contrast between e.g. /ba/ and /da/ is well represented since it involves different articulators, associated to different somatotopic positions in SMC (e.g. Bouchard, Mesgarani, Johnson, & Chang, 2013; Pulvermüller et al., 2006). On the contrary, we suppose that the voicing contrast would be less clearly represented in motor terms, considering that it mainly involves specific coordination between vocal cords and vocal tract controls, rather than a specific articulator. This is summarized – and simplified, as along all this study – in Fig. 5a, where we assume a two-dimensional motor space, i.e. $M = (M_{place}, M_{voicing})$. It would be involved and associated to four "objects" such as /ba/, /da/, /pa/and /ta/, the place dimension M_{place} displaying more contrast than the voicing dimension $M_{voicing}$, as displayed in the $P(M|O_S)$ distributions.

Then, we assume that the motor-to-sensory relationships transform the 2D motor space $M = (M_{place}, M_{voicing})$ into a 2D sensory space $S = (S_{place}, S_{voicing})$ and that the transform may be nonlinear and different from one phonetic dimension to the other. In the present case, we assume that the voicing dimension would correspond to a large enhancement in the auditory domain, as displayed in Fig. 5b_1 by a

nonlinear motor-to-sensory transform $P(S_{voicing}|M_{voicing})$. This is compatible with the Quantal Theory introduced by Stevens (1972, 1989). It is also in line with a number of perceptual data about categorical perception of Voice Onset Time contrasts in both speech and non-speech continua, suggesting possible underlying auditory discontinuities enhancing the representation of voicing in the auditory domain (e.g. Jusczyk, Pisoni, Walley, & Murray, 1980; Pisoni, 1977).

In contrast, we assume that the auditory representation of place of articulation could be difficult to characterize, because coarticulation makes the acoustic realization of plosives quite dependent of adjacent vowels. Hence, we assume that the motor-to-sensory transformation for place would decrease contrast from gestures to sounds, contrary to what was proposed for voicing. This is realized by using a linear motor-to-sensory transform with a small slope (Fig. 5b$_2$). The consequence is that the auditory pattern of the $P(S|O_L)$ distribution would be characterized by a low contrast in the S_{place} dimension, vs. a high contrast in the $S_{voicing}$ dimension (Fig. 5c).

Finally, the four syllabic objects /ba da pa ta/ are supposed to be just the combination of the two possible categories for place and voicing, in both the M and S spaces.

This pattern of distribution of motor and sensory variables for the four objects /ba da pa ta/ provides the basis for the stimuli generated by the Master Agent and that should be learned by a COSMO Learning Agent. The agent learns the three basic distributions in COSMO, $P(S|O_L)$, $P(S|M)$ and $P(M|O_S)$, as described in Section 1.1.2. This describes the probabilistic knowledge the agent has acquired for the 4 syllables. According to the neuroanatomical assumptions in Fig. 2, $P(S|O_L)$ would be stored in the STG, while $P(M|O_S)$ would be stored in the M1-PMC complex, M1 being part of the SMC.

Let us now consider the two tasks studied by Cheung et al. (2016). We begin by the speech production task. We assume that in this simple and well automatized task, production is mainly guided by the motor repertoire (a "syllabary", see Guenther, Ghosh, & Tourville, 2006) represented by the $P(M|O_S)$ distribution stored in the SMC for single articulatory dimensions. Therefore, the SMC activity in the speaking condition (Fig. 4b in Cheung et al., 2016) would be essentially related to the $P(M|O_S)$ distribution in Fig. 6a. To make the correspondence with experimental data clearer, we add in each plot of this figure a display of the d′ values between the four syllables for each distribution, computed by dividing distances between means by values of the standard deviations – equal for the four categories in each plot.

We now consider the speech perception task (Fig. 4e and f in Cheung et al., 2016). According to COSMO, it involves two decoding routes. In line with the schema in Fig. 2, the auditory route would correspond to activity in STG, mainly related to the $P(S|O_L)$ distribution, displayed in Fig. 6b. The motor decoding route would involve activity related to the $P(S|O_S)$ distribution in M1/PMC. This distribution is computed according to Eq. (2) and the result is displayed in Fig. 6c. Importantly, this distribution differs from both $P(M|O_S)$ and $P(S|O_L)$. Therefore, it is, as reported by Cheung et al. (2016), NOT equal to the cortical activity in SMC in the production task: the tasks are different, hence the computations are different, and neuronal activity, likely related to computation, also differs. But, importantly, $P(S|O_S)$ also differs from $P(S|O_L)$: they represent activity in two different – and to a certain extent complementary – pathways for phonetic decoding.

Therefore, the second part of the conclusion by Cheung et al. (2016) is, in our view, inaccurate: the motor cortex in the perception task would contain neither articulatory representations of perceived actions in speech (related to $P(M|O_S)$), nor auditory vocal information (related to $P(S|O_L)$), but rather a third information content $P(S|O_S)$ related to motor decoding. However, importantly, it appears that in the present simulations there does exist a larger resemblance between distributions $P(S|O_S)$ and $P(S|O_L)$ than between $P(S|O_S)$ and $P(M|O_S)$. This is due to the complex way distributions $P(S|O_L)$, $P(S|M)$ and $P(M|O_S)$ are combined for computing $P(S|O_S)$ in Eq. (2). More specifically, the distributions $P(S|O_S)$ and $P(S|O_L)$ are similar in terms of the positions of their mean values – which is indeed what was reported by Cheung et al. (2016). The d′ values reported in the lower plot illustrate these similarities, and provide a rough illustration of how the corresponding data could be simulated.

However, this does not imply that the information in the motor route is lost in the distribution. Of particular importance is the value of

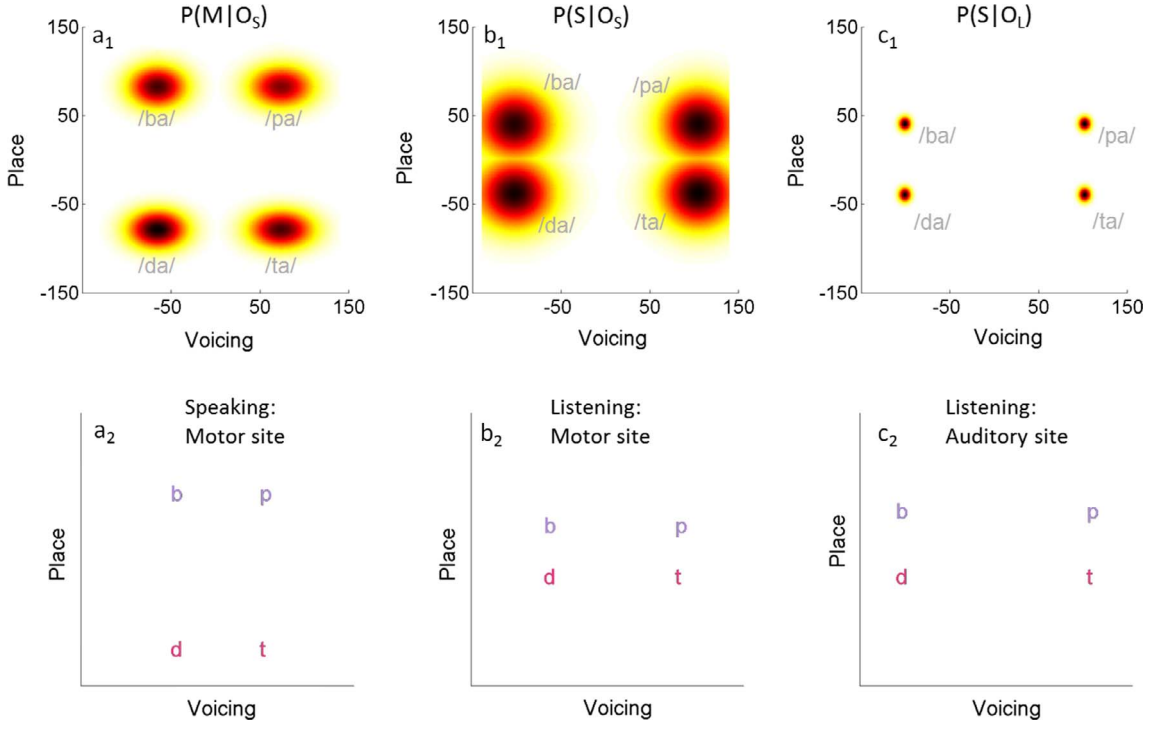

Fig. 6. Simulating neural activity in speech perception and speech production in COSMO. (a) Motor cortex activity for production: top, bottom, d′ values. (b) Motor cortex activity for perception: top, $P(S|O_S)$; bottom, d′ values. (c) Auditory cortex activity for perception: top, $P(S|O_L)$; bottom, d′ values.

the variances of these distributions, in relation with the Complementarity Property and the narrow-band characteristic of auditory decoding contrasted to the wide-band characteristic of motor decoding. Differences in variances between the auditory and the motor routes, well displayed in Fig. 6b and c, show that information in the motor decoding process is not equivalent to information in the auditory decoding process and hence that activity in the frontal areas associated to speech perception might indeed play a specific role for decoding, particularly in noise or adverse conditions.

Notice that Specificity, in the present implementation, is of course only partial. There is indeed information on both plosive place and voicing/manner in both the auditory and the motor decoding routes. Specificity is only displayed here by differences in variances of relevant distributions, which are related to the information available on each dimension in each route. This is compatible with data on the neural representation of phonetic features in STG as recently displayed by Mesgarani, Cheung, Johnson, and Chang (2014), through high-density cortical surface recordings in humans. As a matter of fact, the representation of consonants in their data is basically organized by manner of articulation, place being only a secondary feature, and much less robust in the representation. The present simulations suggest that information on place of articulation could still be available in motor areas, within the $P(S|O_S)$ distribution.

In conclusion, the Specificity Property and the Complementarity Property enable us, in COSMO, to propose an interpretation of the complex pattern of cortical representations observed by Cheung et al. (2016). Our simulations suggest a possible computational content to the observed patterns of cortical activity. Importantly, it also leads to propose some caveats to previous interpretations of such data – and to insist on the difference between representations and tasks, at the center of the Bayesian Programming Approach followed in COSMO. The assumption in the present simulations is that motor knowledge associated with the $P(M|O_S)$ distribution is indeed stored in the frontal areas and exploited for both speech perception and speech production, but that specific computations associated with each of these tasks can lead to differences in neuronal activity in the same region of interest.

5. Discussion

In this paper, we described COSMO, a computational perceptuo-motor model of speech perception, and we hypothesized, for the first time, a neuroanatomical architecture for the representations and computations associated to this model. This enabled us to propose a coherent set of explanations for three different studies corresponding to different neurocognitive paradigms involved in the exploration of perceptuo-motor interactions in speech perception. These explanations are based on three properties of the sensory-motor decoding process: Redundancy, Complementarity and Specificity. Importantly, these properties provide functional arguments supporting the role of a motor decoding route in the speech perception process.

This study has a number of limitations linked to the fact that the aim, at this stage, was to present “proof of concept” illustrations of mechanisms, rather than a mature neurocognitive/computational model of speech perception in the human brain, able to quantitatively account for observed data. Some of the limitations have already been mentioned. They concern the difficulty to relate algorithms and neuronal implementation, and the generic nature of representations and coding in COSMO, particularly concerning hierarchies of variables and processes. Other caveats could be raised concerning the need to specify relationships between linguistic objects and particular phonological and lexical/conceptual units, the lack of dynamic components at the level of both sensory-motor variables and neuronal activity, or the necessity to take into account multi-sensory stimuli. We will discuss specifically three major challenges in more detail.

5.1. Cortical architectures and computational representations

To be able to relate computational properties of COSMO to experimental findings, we introduced in Section 1.3 some coarse assumptions about possible localizations for the distributions learnt by COSMO Agents. We considered three cortical poles, respectively temporal for sensory distributions, temporo-parietal for sensory-motor distributions and frontal for motor distributions. Now that we have displayed a number of relationships between computational processes and neurocognitive data, carefully distinguishing stored distributions from Bayesian questions and inference, we may attempt to go one step further in the specification of neuroanatomical correlates of COSMO processes. However, the question of the hierarchy of computational representations inevitably arises.

In COSMO, hierarchies might be considered at the level of sensory or motor variables, or in the definition of objects. Hierarchies in the representation of sensory or motor variables refer to the fact that cognitive systems involve a number of successive steps both in the processing of sensory inputs and in the elaboration of motor commands. Cognitive sensory hierarchies, not represented yet in COSMO, would refer to successive computational/representational layers in auditory processes in the temporal cortex: from auditory analysis in the Heschl gyrus (HG) to more complex computations, either anteriorly towards the anterior STG/STS or posteriorly towards the pSTS/STG and the planum temporale, considered as a computational hub for processing complex sounds (Griffiths, Warren, & Warren, 2002) (see a review in Friederici, Brauer, & Lohmann, 2011). Neurocognitive motor hierarchies would encompass definitions of motor programs in Broca’s area pars operculum (Brodmann area 44), or in the left PMC, down to the implementation of motor commands in M1 in relation with proprioceptive information in the primary somatosensory cortex S1 or in integrative areas in the inferior parietal lobule, in addition to a number of other cortical and sub-cortical structures (Guenther & Vladusich, 2012; Guenther et al., 2006). The introduction of such cognitive sensory or motor hierarchies associated to successive layers of computation and representation is easily performed in Bayesian modeling (see e.g. Colas, Diard, & Bessière, 2010).

The second kind of hierarchy concerns objects of communication, restricted in the present paper to phonological units. If a hierarchy of linguistic objects were introduced in COSMO, it would at least include lexical units – words or their morphemic components – with relation to meaning. This opens in neuroanatomical terms the question of the ventral vs. dorsal separation in the “dual-stream model of speech processing” (Hickok & Poeppel, 2004, 2007, 2009). Indeed, it is classically considered that speech comprehension basically involves a ventral route connecting, after the Heschl gyrus, anterior regions of the superior and middle temporal gyri and further apart in the anterior temporal lobe, and then anteriorly in Broca’s area pars triangularis (Brodmann area 45) (Hickok & Poeppel, 2007). The dorsal route, already described in previous section, would be mainly used for sensory-motor learning and control in speech production.

This more precise description of the neuroanatomy of speech perception and production circuits leads to refining our proposals for COSMO distributions. Distributions associated to the S variable could refer to various successive centers such as HG and PT/pSTS, and the $P(S|O_L)$ distribution could be stored in PT or pSTS as suggested in Section 1.3. Still, anterior temporal regions could also be suggested to be involved in storing the $P(S|O_L)$ distribution, and it has indeed been suggested by e.g. Liebenthal, Binder, Spitzer, Possing, and Medler (2005) and Obleser, Zimmermann, Van Meter, and Rauschecker (2007) that while posterior temporal regions such as PT or pSTS could be involved in the coding of complex sounds independently of their phonetic nature, phonetic processing in relation with phonemic categories could rather involve the anterior part of STG/STS (see also Price, 2012).

Sensory-motor convergence is claimed by Hickok et al. (2009) (see

also Buchsbaum, Hickok, & Humphries, 2001) to occur in area Spt in the planum temporale, which is hence a good candidate for both storing the $P(S|M)$ distribution and for sensory-motor fusion. Still, proposals by Jacquemot and Scott (2006) suggest that the supramarginal gyrus is also a possible candidate for explicit phonemic access – though see Hickok (2013), for the claim that phonemic access is not part of speech comprehension. Motor distributions $P(M|O_S)$ could be stored at various levels in the motor cortex from M1 to PMC – if not Broca's pars opercularis – depending on the level of complexity of the involved speech motor task.

Importantly, a number of papers were recently published describing the connectivity between these regions in speech perception or speech production tasks, based on either diffusion tensor imaging (DTI) or dynamic causal modeling (DCM) techniques. These papers confirm the existence of bilateral links between temporal and frontal areas. Two long dorsal bundles of neural fibers are involved here: the arcuate fasciculus connects area 44 to the middle and posterior STG, while the superior longitudinal fascile connect the dorsal PMC to the middle and superior temporal gyrus through parietal regions (angular gyrus/supramarginal gyrus) (see Frey, Campbell, Pike, & Petrides, 2008; Saur et al., 2008; and reviews in Friederici & Gierhan, 2013; Friederici & Singer, 2015; Rauschecker & Scott, 2009). The bilateral connections between frontal regions (ventral inferior frontal gyrus and premotor cortex) and temporal areas in charge of speech decoding (HG, PT, STS/STG) (Lyu, Ge, Niu, Hai Tan, & Gao, 2016; Osnes, Hugdahl, & Specht, 2011) confirm the plausibility of the sensory-motor fusion process introduced in COSMO by Eq. (3) in possible relation with the sensory-motor feedforward/feedback circuit proposed in Fig. 2 (see also Skipper et al., 2007).

5.2. Neural and computational dynamics

The second question that comes in mind in the evaluation of COSMO concerns the dynamics of computational processes. This question is crucial for a neurocognitive analysis of the way cortical processes unfold over time and of the series of causal mechanisms likely to progressively elaborate the adequate response to a given speech stimulus. Still, this question is presently not considered in COSMO. Indeed, COSMO manipulates probability distributions, and no underlying temporal unfolding is attached to these manipulations. For example, Eq. (3), which involves fusion of two probability distributions respectively associated to auditory and motor decoding, does not imply any assumption about a possible sequence of computations such as the one suggested at the end of the previous section, that is a feedforward-feedback process connecting auditory and motor regions in the cortex.

There exist various ways to consider time and dynamics in COSMO. Time could be explicitly inserted as a sequence of consecutive discrete events, repeating variables O, S and M for each event. The model should then consider conditional probabilities in which variables at one discrete instant would depend on variables at previous discrete instants. This would make COSMO a dynamic Bayesian network likely to display delays, temporal loops, cycles and more generally temporal sequences, which could then be compared with neurocognitive dynamics. Another way could be to consider that computations expressed in COSMO by probability distributions – the algorithmic level in a three-level description of a cognitive model "à la Marr" (Marr, 1982) – are based on underlying neurophysiological, implementation-level processes which require time to achieve computations.

As a matter of fact, this level of neurophysiological implementation of the probabilistic computations in COSMO would be necessary to better address the relationships between COSMO principles and neurocognitive data considered in the previous sections. Indeed, it was the basis of the reasoning in Section 2 that the transfer of information from sensory to motor regions could depend on the efficiency of auditory decoding, and that neural propagation in the feedforward-feedback loop between auditory and motor areas could be controlled by the amount of evidence for decoding. It also intervenes in the reasoning in Section 4 that different probabilistic computations should result in different patterns of neural activation likely to be displayed by BOLD patterns in fMRI data. The underlying assumption here is of course that there is a relationship between probability distributions, stored or computed in a given task, and neural activation in local cortical areas.

5.3. Multisensory processing

Another perspective for COSMO concerns multisensory interactions. It is well known that speech perception involves audiovisual interactions at various levels. Visual speech improves speech perception in noise or adverse conditions (e.g. Erber, 1969; Grant & Seitz, 2000; Sumby & Pollack, 1954) but also without noise (Davis & Kim, 2004; Reisberg, McLean, & Goldfield, 1987) and even modifies phonemic decoding in silence, as classically exemplified in the McGurk effect (McGurk & MacDonald, 1976). The neuroanatomy of speech perception involves a large network with a crucial role for STS (Beauchamp, Nath, & Pasalar, 2010; Calvert, Campbell, & Brammer, 2000) and a clear involvement of the dorsal route (area Spt, intraparietal sulcus, premotor cortex, inferior frontal gyrus: see e.g. Callan et al., 2003; Miller & D'Esposito, 2005; Okada & Hickok, 2009; Skipper et al., 2007; and a review in Campbell, 2008).

The introduction of a visual component in COSMO does not raise any conceptual difficulty, as Bayesian models routinely describe multisensory fusion (e.g. Ernst & Banks, 2002), and Bayesian algorithmic architectures are easily expanded to integrate additional constraints and sensory routes in a modular fashion (Patri, Perrier, & Diard, 2016). However, this raises the question of the adequate architecture for audiovisual fusion (see reviews in Schwartz, Robert-Ribes, & Escudier, 1998; Summerfield, 1987). A visual component V could be combined to the audio component in the S variable in Fig. 1, to provide a joint sensory input to the decoding process, in what specialists of audiovisual speech perception would call "early-fusion" (the "direct identification model" in Schwartz et al., 1998); or the auditory and visual variables could be kept separate and considered as two different sensory branches enabling separate inferences on all the other COSMO variables ("late fusion"). COSMO hence provides a natural architecture for addressing the potential role of motor processes in audiovisual interactions. Importantly, the Bayesian fusion process at work in COSMO through the coherence variable C (Eq. (3)) is perfectly reminiscent of Bayesian processes proposed for audiovisual fusion (e.g. Andersen, 2015; Massaro, 1987, 1998; Schwartz, 2010).

It is also important to stress at this stage that somatosensory inputs also seem to intervene in speech perception (Gick & Derrick, 2009; Ito, Tiede, & Ostry, 2009). COSMO provides a natural framework for integrating somatosensory inputs to the speech perception process, since the sensory-motor relationships seems crucial for understanding how somatosensory stimulation essentially associated to the speech production process might intervene in speech perception.

6. Conclusion

In spite of some limitations discussed in the previous sections, and in relation with the various perspectives also introduced in these sections, the COSMO structure appears promising for interpreting neurocognitive and neurophysiological data. A crucial element regarding COSMO is that it is not merely a speech perception model but rather a model of the whole speech communication process. As such, it can be used to study both speech perception (e.g. Barnaud et al., 2015, 2016; Laurent et al., 2013, 2017; Schwartz, Barnaud, Bessière, Diard, & Moulin-Frier, 2016) and speech production (Patri, Diard, & Perrier, 2015; Patri et al., 2016). COSMO hence constitutes an integrative framework for addressing a number of questions related to perception and production, associating a developmental perspective with a characterization of online processes. This should be of interest for further analyses of the neurocognitive processes of speech communication.

Acknowledgements

The research leading to these results received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s", J.-L. Schwartz PI).

References

Andersen, T. (2015). The early maximum likelihood estimation model of audiovisual integration in speech perception. *Journal of the Acoustical Society of America, 137*, 2884–2891.

Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication, 22*, 251–267.

Barnaud, M. L., Diard, J., Bessière, P., & Schwartz, J. L. (2015). COSMO, a Bayesian computational model of speech communication: Assessing the role of sensory vs. motor knowledge in speech perception. In *The five joint IEEE international conference developmental learning and epigenetic robotics (ICDL-EPIROB 2015)* (pp. 248–249).

Barnaud, M. L., Schwartz, J. L., Diard, J., & Bessiere, P. (2016). Sensorimotor learning in a Bayesian computational model of speech communication. In *The sixth joint IEEE international conference developmental learning and epigenetic robotics (ICDL-EPIROB 2016)*.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). FMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience, 30*, 2414–2417.

Bessière, P., Mazer, E., Ahuactzin-Larios, J.-M., & Mekhnacha, K. (2013). *Bayesian programming*. Boca Raton, FL: CRC Press.

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: A (Re-)emerging program of research for language and vision. *Biolinguistics, 4*(2–3), 174–200.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience, 7*, 295–301.

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature, 495*, 327–332.

Buchsbaum, B., Hickok, G., & Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science, 25*, 663–678.

Callan, D. E., Callan, A. M., & Jones, J. A. (2014). Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners. *Frontiers in Neuroscience*. http://dx.doi.org/10.3389/fnins.2014.00275 (03 September 2014).

Callan, D. E., Jones, J. A., Munhall, K. G., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport, 14*, 2213–2217.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*, 649–657.

Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society of London B, Biological Sciences, 363*, 1001–1010.

Cheung, C., Hamiton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *eLife, 5*, e12577. http://dx.doi.org/10.7554/eLife.12577.

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

d'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex, 48*, 882–887.

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

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research, 229*(1), 132–147.

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *Quaterly Journal of Experimental Psychology, A, 57*, 1103–1121.

de Boysson-Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel ormants in babbling. *Journal of Child Language, 16*, 1–17.

de Boysson-Bardies, B., Sagart, L., & Durant, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language, 11*, 1–15.

Diard, J. (2015). *Bayesian algorithmic modeling in cognitive science. Habilitation à diriger des recherches (HDR)*. Université Grenoble Alpes.

Diehl, R., Lotto, A., & Holt, L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech Language and Hearing Research, 12*, 423–425.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433. http://dx.doi.org/10.1038/415429a.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience, 15*, 399–402.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*, 752–782.

Frey, S., Campbell, J. S., Pike, G. B., & Petrides, M. (2008). Dissociating the human language pathways with high angular resolution diffusion fiber tractography. *Journal of Neuroscience, 28*, 11435–11444.

Friederici, A. D., Brauer, J., & Lohmann, G. (2011). Maturation of the language network: From inter- to intrahemispheric connectivities. *PLoS ONE, 6*, e20726.

Friederici, A. D., & Gierhan, S. M. E. (2013). The language network. *Current Opinion in Neurobiology, 23*, 250–254.

Friederici, A. D., & Singer, W. (2015). Grounding language processing on basic neurophysiological principles. *Trends in Cognitive Sciences, 19*, 1–10.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature, 462*, 502–504.

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

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*, 1197–1208.

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neuroscience, 25*, 348–353.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*, 280–301.

Guenther, F. H., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics, 25*, 408–422.

Halle, M., & Stevens, K. N. (1959). Analysis by synthesis. In *W. Wathen-Dunn & L. E. Woods (Eds.), Proceedings of the seminar on speech compression and processing*. USAF Camb. Res. Ctr. 2: Paper D7.

Hickok, G., Okada, K., & Serences, J. T. (2009). Area Spt in the human planum temporale supports sensorimotor integration for speech processing. *Journal of Neurophysiology, 101*, 2725–2732.

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

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1), 67–99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.

Hickok, G. (2013). Do mirror neurons subserve action understanding? *Neuroscience Letters, 540*, 56–58.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 1245–1248.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences, 10*(11), 480–486.

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport, 14*, 1129–1133.

Jusczyk, A. M., Pisoni, D. B., Walley, A., & Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. *Journal of the Acoustical Society of America, 67*, 262–270.

Klatt, D. (1980). Software for cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67*, 971–995.

Kleinschmidt, D., & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review, 122*, 148–203.

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., & Ostry, D. J. (2014). Plasticity in the human speech motor system drives changes in speech perception. *Journal of Neuroscience, 34*(31), 10339–10346.

Laurent, R., Barnaud, M. L., Schwartz, J. L., Bessière, P., & Diard, J. (2017). The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. doi:http://dx.doi.org/10.1037/rev0000069.

Laurent, R., Schwartz, J.-L., Bessière, P., & Diard, J. (2013). A computational model of perceptuo-motor processing in speech perception: Learning to imitate and categorize synthetic CV syllables. *Proceedings of InterSpeech* (pp. 2797–2801). France: Lyon.

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

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15*, 1621–1631.

Lyu, B., Ge, J., Niu, Z., Hai Tan, L., & Gao, J.-H. (2016). Predictive brain mechanisms in sound-to-meaning mapping during speech processing. *The Journal of Neuroscience, 36*, 10813–10822.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. New York, USA: W.H. Freeman and Company.

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

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle, Vol. 1*. Cambridge, MA: MIT Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12*, 369–378.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology, 17*, 1692–1696. http://dx.doi.org/10.1016/j.cub.2007.08.064.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*, 1006–1010.

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience, 25*, 5884–5893.

Möttönen, R., Dutton, R., & Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cerebral Cortex, 23*, 1190–1197. http://dx.doi.org/10.1093/cercor/bhs110.

Möttönen, R., & Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *The Journal of Neuroscience, 29*, 9819–9825.

Moulin-Frier, C., Diard, J., Schwartz, J.-L., & Bessière, P. (2015). COSMO ("Communicating about Objects using Sensory-Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics, 53*, 5–41.

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., & Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor, and perceptuo-motor theories of speech perception: An exploratory Bayesian modelling study. *Language and Cognitive Processes, 27*, 1240–1263.

Obleser, J., Zimmermann, J., Van Meter, J. W., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related fMRI. *Cerebral Cortex, 17*, 2251–2257.

Ojanen, V., Möttönen, R., Pekkola, Jääskeläinen, I., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage, 25*, 333–338.

Okada, K., & Hickok, G. (2009). Two cortical mechanisms support the integration of visual and auditory speech: A hypothesis and preliminary data. *Neuroscience Letters, 452*, 219–223.

Osnes, B., Hugdahl, K., & Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *NeuroImage, 54*, 2437–2445.

Patri, J. F., Diard, J., & Perrier, P. (2015). Optimal speech motor control and token-to-token variability: A Bayesian modeling approach. *Biological Cybernetics (Modeling), 109*, 611–626.

Patri, J. F., Perrier, P., & Diard, J. (2016). Bayesian modeling in speech motor control: a principled structure for the integration of various constraints. In *17th Annual conference of the international speech communication association (Interspeech 2016), Sep 2016, San-Francisco, United States* (pp. 3588–3592).

Paulesu, E., Frith, C. D., & Frackowiak, R. S. (1993). The neural correlates of the verbal component of working memory. *Nature, 362*(6418), 342.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in steps. *Journal of the Acoustical Society of America, 61*, 1352–1361.

Pollack, I., & Pisoni, D. (1971). On the comparison between identification and discrimination tests in speech perception. *Psychonomic Science, 24*, 299–300.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heardspeech, spoken language and reading. *NeuroImage, 62*(2), 816–847.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences, 103*, 7865–7870.

Rauschecker, J., & Scott, S. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience, 12*, 718–725.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd, & R. Campbell (Eds.). *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Repp, B. H. (1984). Categorical perception: Issues, methods and findings. In N. Lass (Vol. Ed.), *Advances in basic research and practice: . Vol. 10. Speech and language* (pp. 244–335). Orlando, FL: Academic Press.

Rogers, J. C., Möttönen, R., Boyles, R., & Watkins, K. E. (2014). Discrimination of speech and non-speech sounds following theta-burst stimulation of the motor cortex. *Frontiers in Psychology, 5*, 754.

Sato, M., Grabski, K., Glenberg, A., Brisebois, A., Basirat, A., Ménard, L., & Cattaneo, L. (2011). Articulatory bias in speech categorization: Evidence from use-induced motor plasticity. *Cortex, 47*, 1001–1003. http://dx.doi.org/10.1016/j.cortex.2011.03.009.

Sato, M., Tremblay, P., & Gracco, V. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language, 111*, 1–7.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., et al. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences USA, 105*, 18035–18040.

Schwartz, J. L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *Journal of the Acoustical Society of America, 127*, 1584–1594.

Schwartz, J.-L., Abry, C., Boë, L.-J., & Cathiard, M. (2002). Phonology in a theory of perception-for-action-control. In J. Durand, & B. Laks (Eds.). *Phonology: From phonetics to cognition* (pp. 255–280). Oxford: Oxford University Press.

Schwartz, J. L., Barnaud, M. L., Bessière, P., Diard, J., & Moulin-Frier, C. (2016). Phonology in the mirror. *Physics of Life Reviews*. http://dx.doi.org/10.1016/j.plrev.2016.01.007.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics, 25*, 336–354. http://dx.doi.org/10.1016/j.jneuroling.2009.12.004.

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

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

Serniclaes, W., & Sprenger-Charolles, L. (2003). Categorical perception of speech sounds and dyslexia. Current psychology letters. *Behaviour, Brain & Cognition, 1*(10), 1–8.

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *Journal of the Acoustical Society of America, 125*, 1103–1113. http://dx.doi.org/10.1121/1.3058638.

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language, 164*, 77–105.

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

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

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. David, & P. Denes (Eds.). *Human communication: A unified view* (pp. 51–66). McGraw-Hill.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audiovisual speech perception. In Dodd, & R. Campbell (Eds.). *Hearing by eye: The psychology of lipreading* (pp. 3–51). Hillsdale, NJ: Lawrence Erlbaum Associates.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*, 13273–13278.

Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in productibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage, 33*, 316–325.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage, 32*, 1826–1836.