5th International Conference on Development and Learning
and on Epigenetic Robotics
Aug 13-16, 2015. Providence, RI USA

COSMO, a Bayesian computational model of speech communication: Assessing the role of sensory vs. motor knowledge in speech perception

Marie-Lou Barnaud*, Julien Diard†, Pierre Bessière‡ and Jean-Luc Schwartz*
*GIPSA-LAB Grenoble, France; Email: marie-loubarnaud@wanadoo.fr
†Univ. Grenoble Alpes, LPNC, F-38000 Grenoble, France CNRS, LPNC, F-38000 Grenoble, France
‡SORBONNE Universités - UPMC - ISIR, Paris, France

I. Introduction: Sensory vs. motor speech recognition

It is now widely accepted that there is a functional relationship between the speech perception and production systems in the human brain. However, the precise mechanisms and role of this relationship still remain debated.

The question of invariance and robustness in categorization are set at the center of the debate: how is stable information extracted from the variable sensory input in order to achieve speech comprehension? In this context, auditory (resp. motor, perceptuo-motor) theories propose that speech is categorized thanks to auditory [1] (resp. motor [2], perceptuo-motor [3], [4]) processes. However, experimental evidence is still scarce and does not allow to clearly distinguish between the current theories and determine whether invariance in speech perception is of an auditory or motor type [5], [6].

This is where computational models can play a crucial role. Since neurocognitive data undoubtedly show that both sensory and motor knowledge intervene in speech perception, we developed COSMO, a Bayesian model comparing sensory and motor processes in the form of probability distributions which enable both theoretical developments and quantitative simulations. A first significant result in COSMO is an indistinguishability theorem [7]: it is only by simulations of adverse conditions or partial learning that the specificity of sensory vs. motor processing can emerge and provide a basis to evaluate the role of each sub-system.

In the present work, we first present the COSMO model, and how its sensory and motor sub-systems are learned, then we describe simulations exploring the way these sub-systems differ during speech categorization. We discuss the experimental results in the light of a "narrowband vs. wideband" interpretation.

II. Method: COSMO, a Bayesian sensori-motor computational model

COSMO consists of five probabilistic variables which correspond to internal representations of an agent (see Fig 1, left): O_S for objects in the motor sub-system and O_L for the same objects in the sensory sub-system; S for the sensory input and M for the motor gestures. Lastly, to ensure that both representations of objects describe the same one there is a coherence variable C. Using some conditional hypotheses, we choose to decompose the joint distribution $P(O_S\ O_L\ M\ S\ C)$ as a product of five distributions: a prior on objects $P(O_S)$, a production system $P(M|O_S)$, a sensory-motor system $P(S|M)$, an auditory recognition system $P(O_L|S)$ and a communication validation system $P(C|O_S\ O_L)$. "Speech" in the version of COSMO presented here consists in extremely simplified "sensory" and "motor" (mono dimensional) spaces, in order to facilitate experimental manipulations and result interpretations.

Agents learn their sensory sub-system $P(S|O_L)$, their motor production system $P(M|O_S)$ and their sensory-motor system $P(S|M)$ in a supervised learning scenario, in which a master agent provides sensory signals s and their respective object o.

Learning the sensory sub-system $P(S|O_L)$ is straightforward using experimental $< s, o >$ pairs. On the other hand, learning the motor sub-system is more complicated, as it is made of both $P(M|O_S)$ and $P(S|M)$, and motor gestures are not provided to the learning agent. Instead, it infers them, in an "accomodation process": the learning agent tries to reproduce the input sensory signal s by selecting a motor gesture m. Performing m yields s', the resulting sensory output. Triplets $< m, s', o >$ are used to update the parameters of the motor sub-system. Advantageously, this learning process does not require an exhaustive exploration of the motor space M: starting from blind random exploration, it gradually focuses on the portions of the motor space that correspond to the provided acoustic inputs.

III. Experimental results: Testing the "narrowband vs. wideband" hypothesis

An initial analysis provide two results (Figures not shown here). First, the sensory sub-system is learned faster than the motor sub-system (less than 200 learning steps for the sensory one and more than 20,000 for the motor one). Second, during a perceptual categorization task (learning is stopped), the sensory sub-system discriminates objects better than its motor counterpart in normal conditions, whereas the motor sub-system is more robust in noise. By exploring the functional roles of the sensory vs. motor sub-systems during this perceptual categorization we can see that during learning, the sensory sub-system

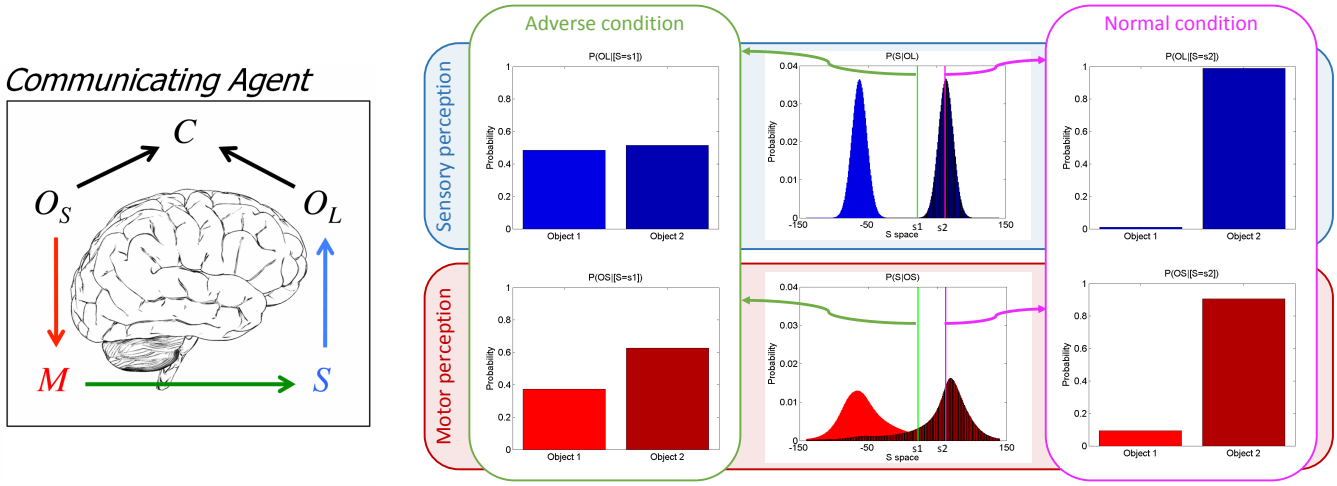

Fig. 1. **Left:** Structure of the COSMO model. **Right:** When learning is stopped, we compare the sensory (top half) and motor sub-systems (bottom half) during perceptual categorization. Middle panels show sensory prototypes associated to two objects. Left panels show perceptual categorization in adverse conditions, i.e. when the sensory input to be categorized is uncommon: in this case, the motor sub-system has better performance due to larger variance and thus, better generalization. Right panels show perceptual categorization in normal conditions, i.e. when the sensory input is common: in this case, the sensory sub-system has better performance due to its smaller variance.

is updated quickly and efficiently whereas the motor sub-system is updated slowly. Consequently, when learning is stopped before asymptotic convergence, the sensory sub-system has a lower variance than the motor sub-system (see Fig 1, right).

We interpret this as a "narrowband vs. wideband" effect. During the perceptual categorization evaluation task, we observe that the sensory sub-system provides a "narrowband" system: it is more precisely tuned to the frequently learned sensory input and hence more efficient in recognizing these inputs. Conversely, we observe that the motor sub-system has "wideband" characteristics: it is less accurate to recognize learned sensory inputs but it has better generalization properties, making it more robust to unexpected variability. These experimental results show that these two systems are complementary, which is in accordance with neurocognitive data showing the increased role of the motor system in adverse conditions [8], [9].

IV. Conclusion: Toward idiosyncrasies

The present simulations on very simple sensory-motor configurations have already been extended to more complex configurations involving synthetic plosive-vowel sequences, with basically the same pattern of results: the "auditory narrowband" vs. "motor wideband" division is confirmed with an auditory system quicker and more efficient on learned stimuli, and a motor system slower and more efficient on noisy stimuli. We are currently exploring further the learning algorithm and its ability to produce "idiosyncrasies" which are variations in learned motor and sensory strategies in the learning agent. This is based on an enriched interaction mechanism, in which the learning agent does not try to mimic sensory inputs of the master, but merely tries to be understood by the master, that is to say, producing stimuli that enable the master agent to recover the correct object.

Acknowledgment

Authors would like to thank Raphaël Laurent for helpful discussions. Research supported by a grant from the European Research Council (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s").

References

[1] R. L. Diehl, A. J. Lotto, and L. L. Holt, "Speech perception," *Annual Review of Psychology*, vol. 55, no. 1, pp. 149–179, 2004.
[2] B. Galantucci, C. A. Fowler, and M. T. Turvey, "The motor theory of speech perception reviewed," *Psychonomic bulletin & review*, vol. 13, no. 3, pp. 361–377, 2006.
[3] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The perception-for-action-control theory (pact): A perceptuo-motor theory of speech perception," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 336–354, 2012.
[4] J. I. Skipper, V. van Wassenhove, H. C. Nusbaum, and S. L. Small, "Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception," *Cerebral Cortex*, vol. 17, no. 10, pp. 2387–2399, 2007.
[5] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.
[6] A. D'Ausilio, L. Craighero, and L. Fadiga, "The contribution of the frontal lobe to the perception of speech," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 328–335, 2012.
[7] R. Laurent, "Cosmo: un modèle bayésien des interactions sensori-motrices dans la perception de la parole," Ph.D. dissertation, Grenoble University, 2014.
[8] A. D'Ausilio, I. Bufalari, P. Salmas, and L. Fadiga, "The role of the motor system in discriminating normal and degraded speech sounds," *Cortex*, vol. 48, no. 7, pp. 882–887, 2012.
[9] R. Mottonen, R. Dutton, and K. E. Watkins, "Auditory-motor processing of speech sounds," *Cerebral Cortex*, vol. 23, no. 5, pp. 1190–1197, 2012.