COSMO: a Bayesian model of sensori-motor interactions in speech perception

An overview of the PhD dissertation

Raphaël Laurent

While speech communication is a faculty that seems natural, a lot remains to be understood about the nature of the cognitive representations and processes that are involved. Central to this PhD research is the study of interactions between perception and action during production or perception of syllables. We choose Bayesian Programming as a rigorous probabilistic framework within which we provide a mathematical definition of the *COSMO* model ("Communicating Objects using Sensori-Motor Operations"), which allows to formalize both motor, auditory and perceptuo-motor theories of speech communication and to study and compare them quantitatively.

The dissertation is structured as follows. Chapter 1 provides a short introduction. Chapter 2 introduces the bibliographic background which is at the heart of all the following theoretical and computational developments. Chapter 3 describes how the *COSMO* model is built and how it provides a generic integrative framework to quantitatively study perceptuo-motor interactions in speech production and perception. We prove a theorem that states that motor and auditory theories of speech perception cannot be distinguished in some ideal learning conditions. In Chapter 4, we present an original algorithm for "learning by accomodation". *COSMO* is instantiated in an abstract theoretical framework which allows to illustrate in details how this algorithm works, how an imitation paradigm allows to learn motor skills from acoustic inputs only, and to develop idiosyncrasies as well as robustness in perception in degraded conditions. Chapter 5 aims at going towards more complex and more realistic stimuli. It describes how we model plosive-vowel syllables to generate articulatory and acoustic data showing plausible variability patterns thanks to a vocal tract model: *VLAM*. Chapter 6 extends the *COSMO* model to process syllables and uses the syllable data synthesized in Chapter 5 as inputs for the learning by accomodation algorithm. We generalize the results of Chapter 4 to syllables, and we show that the motor information allows the notion of consonantic categories to emerge. Chapter 7 summarizes our main contributions, and key aspects of our work are discussed, giving directions for future works.

This document is aimed at giving in English a concise view of the content of the 200 pages of the original dissertation which is written in French. We summarize all chapters, except Chapter 1 which is nothing but a short introduction, and Chapter 7 which concludes by discussing further some potentially interesting aspects of our work.

Chapter 2 – Theories and models to account for variability in speech

For many years, researchers have tried to characterize speech by looking for invariants corresponding to the speech units. There were originally two very contrasting views. On the one hand, motor theories defended the view that phonological invariants are motor in nature, whereas acoustic speech signals resulting from motor gestures are highly variable because of coarticulation phenomena. On the other hand, auditory theories claim that all the underlying motor information is encoded in the acoustic signal, hence speech should be characterized in an acoustic space, thanks to acoustic invariants. It is worth noting that in this debate, most of the arguments that are put forward are functional arguments: a given theory should be selected because it solves the invariance problem better. Chapter 2's Section 1 presents some of these arguments in favor of motor vs. auditory theories in the framework of speech perception and production.

More recently, with more and more evidence mainly coming from cognitive neuroscience (such as the discovery of mirror neurons to start with), it became increasingly clear that perceptual and motor representations are both involved in the perceptual processing of speech units (see Section 2). The research agenda now seems to be focused on trying to better understand how these perceptuo-motor interactions develop, how they are represented in the human brain and what modulates the respective role of perceptual and motor components. Surprisingly, the functional aspects seem to be absent from today's literature. Indeed, not much is said on what makes motor knowledge useful for perception, how it could be helpful in noise, and what kind of information – or what specific aspects of computation – are respectively processed by the motor and auditory components of the speech perception system.

To tackle these questions, we feel there is a need for computational models integrating acoustic and motor knowledge. Section 3 describes a few such models, which belong to two categories. On the one hand there are models developed in the field of Automatic Speech Recognition that try to use motor features to help capture speech variability in order to improve performances as measured by the Word Error Rate. (This at best quantifies how useful motor knowledge can be, but says little on what it is that makes it useful.) On the other hand there are a few works in the field of cognitive modeling, but none of which convincingly explains the role of motor knowledge in perception.

The ambition of this PhD dissertation is to develop an integrative computational model, and to see how it can contribute to clarify what it is exactly that motor knowledge can bring to perception. We choose the Bayesian Programming formal framework because it enables principled comparisons of different theories, and because it allows for building models of complex systems that are at once plausible, expressive, and easy to interpret.

Chapter 3 – *COSMO*: a formal and generic communicating agent model to quantitatively study perceptuo-motor interactions in spoken communication

This Chapter starts with an analysis of the spoken communication situation, which leads to proposing a conceptual model thereof (Figure 3.2). Central to this model is the idea that a shared attention mechanism (such as for instance *deixis*) allows communicating agents to assess whether the objects communicated by the speaker and those retrieved by the listener are indeed the same.

From this conceptual communication model, we build a communicating agent model based on two main ideas. First, to account for the observed co-activations of motor and auditory brain areas, our model contains an explicit link between motor (M) and sensory (S) representations. Second, to account for experimental data showing that there is a functional separation between phonological codes for production and perception, and that these codes are linked by conversion mechanisms, our model contains two separate variables O_S and O_L, linked by a variable C to ensure these representations stay coherent. Altogether, the whole communication loop is internalized into the cognitive system of each communicating agent.

This model, which we name *COSMO* (the acronym lists the model variables, and also stands for *Communicating Objetcs using Sensori-Motor Operations*), is formally defined within the framework of Bayesian Programming. Probability distributions encode subjective knowledge the agent has about the relations between its internal representations. The conceptual agent model shown in Figure 3.3 is formalized as the Bayesian network characterized by Equation 3.9.

A speech production task is formally defined within *COSMO* as computing a probability distribution of the form $P(M \mid O)$ over motor gestures M given some object O to be communicated, and a speech perception task is defined as computing a probability distribution of the form $P(O \mid S)$ over recognized objects O given a sensory input S. We define instanciations of these tasks within the framework of motor, auditory and perceptuo-motor theories by choosing a different focus in each case. Motor theories focus on $O{=}O_S$, i.e. on the speaker's perspective. Auditory theories focus on $O{=}O_L$, i.e. on the listener's perspective. Perceptuo-motor theories focus on $O{=}O_S{=}O_L$, i.e. either on the speaker's or the listener's perspective, but with the constraint that they should be coherent. Figure 3.5 shows both how Bayesian inference allows to compute production and perception tasks instanciated in the context of motor, auditory or perceptuo-motor theories, and how such computations can be interpreted. This shows that *COSMO* provides a unique integrative framework within which the different theories can be compared on a common mathematical ground.

Our approach leads at the end of this chapter to a strong theoretical result: we prove an indistinguishability theorem, according to which, given some ideal learning conditions, motor and auditory theories make identical predictions for perception tasks, and therefore cannot be distinguished empirically. What this theorem really states is that, if a perception model has enough statistical expressiveness to learn and exactly capture the realizations of a structured production model, then these two models cannot be distinguished from one another on production tasks.

Chapter 4 – Exploiting the *COSMO* model, within a simplified theoretical framework, to compare motor and auditory approaches to speech perception

Given that the ideal learning hypotheses of the indistinguishibility theorem are too restrictive to hold in practice, Chapter 4 describes the implementation of realistic learning algorithms, which allows to depart from these ideal learning conditions, and thus gives meaning to the comparison of the predictions made by the motor and auditory theories.

We implemented learning in two steps. First, the sensory system is simply learned by association of acoustic inputs with their corresponding labels in a straightfoward manner. Second, we propose an original "learning by accommodation" algorithm which allows the agent to learn and develop both motor repertoires and an internal model of the articulatori-acoustic transformation. The heart of this algorithm is a paradigm of learning by imitation, according to which the agent learns by trying to mimick the sensory inputs of the ambient acoustic environment, given its current state of knowledge. This algorithm relies on two principles: progressive refining of the internal model of the articulatory-to-acoustic mapping (with each attempt to reproduce an acoustic target, the agent acquires some knowledge of the acoustic consequence of the selected motor gesture, and this knowledge influences later choices) and progressive anchoring of the choice of adequate motor gesture (the learning agent tends to favor gestures used in the past, thus developing idiosyncrasies). Remarkably, this algorithm, which can be seen as target oriented babbling, allows to acquire motor skills from acoustic inputs only, without feedback, and without the need to explicitly bootstrap with a phase of uniform sampling or random exploration.

Whereas Chapter 3 describes an abstract and generic version of the *COSMO* model, Chapter 4 describes thouroughly how we implement it on a very simplified theoretical framework: unidimensional motor and sensory variables that are linked by a sigmoid transformation. This allows to illustrate in detail the behavior of our learning algorithms, to explore their dynamics, and to better understand the nature of the information that is learned.

We show that the auditory model, which only learns direct associations between objects and stimuli, learns/adapts faster than the motor model, which needs to build both motor repertoires and an internal model of the sensorimotor mapping. The motor model is both very precise for production tasks – the probability distributions encoding motor repertoires have low variance which provides high consistency – and a bit less precise in perception tasks – the internal model of the sensorimotor mapping is imperfectly learned. Although this imprecision makes the motor model perform a bit worse than the auditory model in normal (learning) conditions, it also brings it robustness to degradations (noise). One way to look at it is to see the auditory model as narrow-band and specialized on training conditions, and the motor model as wider-band, having explored through accomodation learning some areas of the sensorimotor space unknow to the auditory model, and able to generalize a bit better. The perceptuo-motor model performs a Bayesian sensor fusion and hence always performs better than purely motor or purely auditory perception.

Chapter 5 – Synthesizing realistic syllables within the *COSMO* framework of thanks to a vocal tract model: *VLAM*

The theoretical framework used in Chapter 4 allows to highlight interesting principles, but it is still very simplified (the motor and sensory variables are abstract and unidimensional). The remainder of the PhD dissertation aims at studying whether these results scale to more complex speech signals.

Chapter 5 starts with a bibliographic study from which we extract simplifying principles to model plosive-vowel syllables. We choose to use *VLAM* (the *Variable Linear Articulatory Model*), which is a vocal tract simulation model based on the statistical analysis of cineradiographic data from a speech corpus in French. In motor space, speech is described by a set of seven articulatory parameters, which can be interpreted in terms of phonetic and muscular commands. In sensory space, speech is described by the first three formants on a perceptual scale in Barks. From values of the articulatory parameters, *VLAM* reconstructs the vocal tract shape and computes the transfer function as well as the formants.

While we acknowledge that a number of more recent and probably better models have been proposed since, we adopt, because of its simplicity, the view proposed by Öhman with his perturbation model: plosives are viewed as local perturbations (vocal tract closing gestures) of vowel configurations within Consonant-Vowel syllables. Using this idea, we generated synthetic syllables with *VLAM* as follows. We selected prototypical articulatory configurations corresponding to prototypical formant values for the vowels /a/, /i/ and /u/ in French. Around these prototypes, we draw a set of vowel configurations from a Gaussian probability distribution on the motor space. For each vowel, we generated the plosives /b/, /d/ and /g/ with a maximal coarticulation principle: starting from the vowel articulatory configuration, the vocal tract is closed by combining the jaw with another articulator (lips for the bilabial /b/, tongue apex for the dental /d/, and tongue dorsum for the velar /g/). Within the syllable the consonant is therefore conditioned by the vowel: there is a strong anticipation of the vowel realisation when the consonant is produced.

The synthetic syllables are validated by comparing them to data from the literature. In particular, we compare formant patterns with those displayed by Sussman to propose his famous locus equations. The point is not to show that our synthetic data accurately correspond to real speech, but rather to show that we have generated syllables whose variability patterns are similar enough to the complexity of speech signals. It is one of the objectives of the following chapter to see how the motor and sensory subsystems of the *COSMO* model extended to syllables can deal with this variability.

Chapter 6 – Realistic sensori-motor learning with the *COSMO* model, and an application to syllable perception tasks

Chapter 6 describes in details the implementation of *COSMO-S* (for *COSMO*-Syllables), which is an adaptation of the *COSMO* model to the case of Plosive-Vowel syllables. *COSMO-S* is comprised of three subsystems. The auditory system associates sensory representations with the corresponding phonemic representations (i.e. the syllable labels). The sensori-motor system associates motor and sensory representations. This knowledge is stored as a forward model, but thanks to our probabilistic framework it can be used as an inverse model as well. The motor system associates motor representations with phonemic representations. The structure of the motor system reflects Öhman's perturbation hypothesis by explicitly introducing a delta variable describing the perturbation superimposed on the vowel to obtain a plosive consonant. These three subsystems are linked by coherence variables, which provide a mathematical way to implement a probabilistic switch allowing to activate or inactivate the different parts of the model during probabilistic inference.

The process for learning the model parameters is implemented in three steps inspired by a developemental chronology proposed by Kuhl. First, an auditory classifier is learned by associating sensory stimuli with object labels. Second, an internal model of the sensori-motor mapping is learned by accomodation, using the general algorithm introduced in Chapter 4. Third, motor gestures repertoires are also learned by accommodation, from sensory inputs. Selecting an appropriate motor command to reproduce an acoustic target requires to inverse the articulatori-acoustic mapping, which is many-to-one. In our Bayesian framework, this is done effortlessly by considering all motor gestures, weighted by the probability that they result in the desired output. When there are several likely candidates, a motor anchoring principle allows the learning agent to develop idiosyncrasies: performing a motor command increases the probability for it to be chosen again in the future.

Chapter 6 generalizes to syllable perception all the results obtained in the theoretical framework of Chapter 4 (although we only consider a small subset of Plosive-Vowel syllables: /ba/, /bi/, /bu/, /ga/, /gi/, /gu/, /da/, /di/, /du/). We take advantage of the *COSMO* model to tackle the question of the mergence of phonological categories. We show that, while the information present in the acoustic space seems sufficient for the notion of vocalic categories to emerge naturally, there is no obvious reason for the notion of consonant category to emerge as easily. On the other hand, we show that it is possible to learn from the auditory inputs only to associate in the motor space one (bilabial) constriction gesture to the syllables /ba/, /bi/ and /bu/, another one (dental) to /da/, /di/ and /du/, and a third one (velar) to /ga/, /gi/ and /gu/; and that this learning is facilitated by the use of *motherese* (over-articulation). This gives support to the notion that motor skill acquisition could play a role in the emergence of phonological categories.

Overall, *COSMO* provides a rich framework which allowed us to implement and quantitatively compare motor, auditory and perceptuo-motor theories of speech communication; to carry out production, perception, and imitation tasks; to propose original learning algorithms and to study their learning dynamics; to have the learning agents develop idiosyncrasies; and to study the robustness of the learned models to degraded conditions.