HAL
open science

A Unified Theoretical Bayesian Model of Speech Communication

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

To cite this version:

Clément Moulin-Frier, Jean-Luc Schwartz, Julien Diard, Pierre Bessière. A Unified Theoretical Bayesian Model of Speech Communication. ICDHM 2010 - 1st International Conference on Applied Digital Human Modeling, Jul 2010, Miami, Floride, United States. pp.457-466. hal-01059208

HAL Id: hal-01059208

https://hal.science/hal-01059208v1

Submitted on 1 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Unified Theoretical Bayesian Model of Speech Communication

Clément Moulin-Frier[1], Jean-Luc Schwartz[1], Julien Diard[2], Pierre Bessière[3]

[1]GIPSA-Lab, Speech and Cognition Department (ex-ICP), UMR 5216,
[2]Laboratoire de Psychologie et NeuroCognition (LPNC), UMR 5105,
[3]Laboratoire d'Informatique de Grenoble (LIG-Lab), UMR 5217
[1, 2, 3]CNRS – Grenoble University.

ABSTRACT

Based on a review of models and theories in speech communication, this paper proposes an original Bayesian framework able to express each of them in a unified way. This framework allows to selectively incorporate motor processes in perception or auditory representations in production, thus implementing components of a perceptuo-motor link in speech communication processes. This provides a basis for future computational works on the joint study of perception, production and their coupling in speech communication.

Keywords: Speech Communication, Cognitive Bayesian Modeling, Sensory-Motor interaction

INTRODUCTION: MODELS AND THEORIES IN SPEECH COMMUNICATION

Speech communication involves a set of actuators for producing speech stimuli (enabling to control the orofacial system: lungs, glottis, jaw, tongue, lips, velum) and a set of sensors for perceiving them (audition of course, but also vision for lip-reading, and haptics and proprioception for sensing the state of the vocal tract). This enables the speaker to control the task in speech production that is achieving the correct gestures for uttering the adequate sounds. Hence, speech production can be conceived as a typical robotics problem, involving proximal control in reference to given distal objectives, together with learning, adaptability, or any other problem

related to cognitive robotics. But the special issue in speech communication is that the task IS communicative. The speaker is also a listener, and has probably a model of the listener incorporated in the production task itself. Production and perception are closely related in communication and probably also in the human's brain. This intimate link between production and perception in speech communication has been largely discussed by phoneticians and cognitive (neuro)psychology, but seldom addressed from a modeling perspective. This is the focus of the present paper.

A key question in speech science concerns the nature of the content of communication, with three major frameworks that are motor, auditory, and sensory-motor theories of speech communication. We shall describe here how each of these theories considers both the speech production and perception processes. Then we will propose a Bayesian formal framework able to express each of them in a unified way, and discuss the possible interpretations of this model. Finally, we will conclude on the possible functional roles for the perceptuo-motor link in speech communication.

MOTOR THEORIES

Motor theories consider the reference frames of speech communication as gestures. In the Articulatory Phonology framework (Browman and Goldstein, 1989), production is modeled as scores of overlapping gestures (Fig. 1.1), able to express the context-dependent variability of speech, without taking explicitly into account the auditory consequence of a motor event.

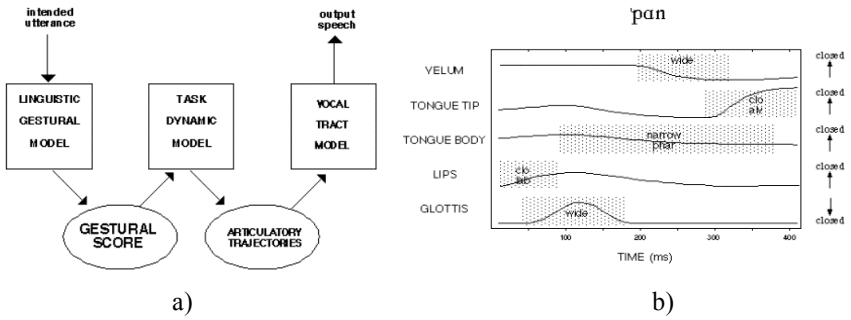

FIGURE 1.1 Articulatory Phonology (from Browman and Goldstein, 1989). a) Gestural computational model. b) Gestural score for the utterance 'pan'.

In line with the idea that the frames of speech communication are motor events, the Motor Theory of Speech Perception (Liberman and Mattingly, 1985) proposes that perceiving speech amounts to perceiving gestures. A main argument is coarticulation-driven signal variability, which makes the auditory content of a given phoneme dependent on the phonetic context (e.g. /d/ does not produce the same sound in /da/ vs. /du/, see Fig. 1.2), whereas the intended gesture is invariant. The

interest for the Motor Theory of Speech Perception was recently renewed by the discovery of mirror neurons (located in the premotor cortex of the macaque, active both when the macaque performs a transitive action or observes another individual performing the same action, Rizzolatti and Arbib, 1998).

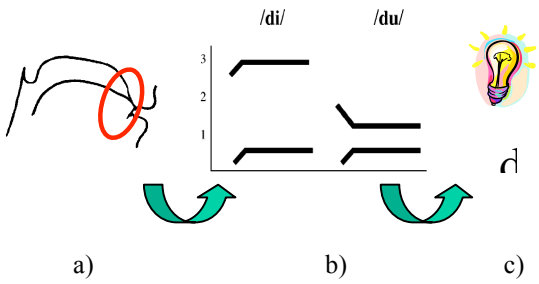

FIGURE 1.2. Illustration of the core argument for the Motor Theory of Speech Perception, regarding the phoneme /d/ in /di/ vs /du/: a) the gesture is the same, that is closing the front of the vocal tract by putting the tongue against the teeth, b) the signal is different and c) the percept is the same. Therefore, the invariant would be the gesture.

AUDITORY THEORIES

Auditory Theories consider that the reference frame for speech is auditory. In the case of speech production, the target would be a region in the auditory space (Guenther et al., 1998). The main argument is motor equivalence, showing that various articulatory configurations are used for achieving the same auditory goal, as shown in perturbation experiments: if the articulatory apparatus is constrained, e.g. by inserting a tube between the lips (Savariaux et al., 1999), speakers reorganize their motor configuration to achieve the same auditory region.

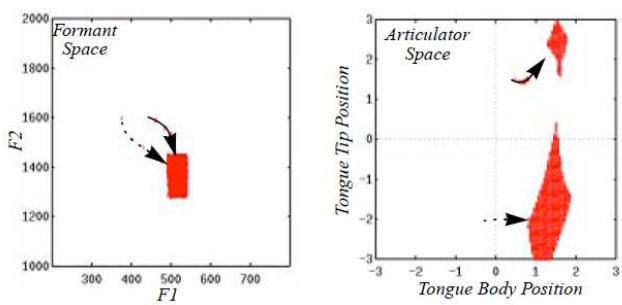

FIGURE 1.3. Motor equivalence for /r/ in English in the DIVA model (from Guenther et al., 1998). Relationship between a simple convex region corresponding to /r/ in the acoustic space (left) and the corresponding regions in the articulatory space (right). Arrows indicate model trajectories when producing /r/ starting from a /d/ configuration (solid lines) and from a /g/ configuration (dashed lines).

Speech production would exploit this adaptability, as in the case of /r/ in English, pronounced in the DIVA articulatory model (Guenther et al., 1998) with different configurations (bunched vs. retroflex) depending on the previous consonant (Fig. 1.3).

In the case of speech perception, proponents of auditory theories consider that speech perception involves auditory or multisensory representations and processing, with no reference to speech production (Diehl et al., 2004).

SENSORY-MOTOR THEORIES

Sensory-motor theories have recently emerged for both speech perception and production. They generally consider auditory frames as the core for communication, but they include the sensory-motor link inside the global architecture. They claim that in normal conditions, production involves cortical motor (frontal) areas and perception involves cortical (temporal) auditory ones, but that the perceptuo-motor link, necessary for speech acquisition, could also play a role in adverse conditions.

Regarding speech production, the DIVA model (Guenther, 2006) combines a feedforward control sub-system for on-line production, and a feedback control sub-system when the auditory consequence of a gesture is not congruent (Fig. 1.4).

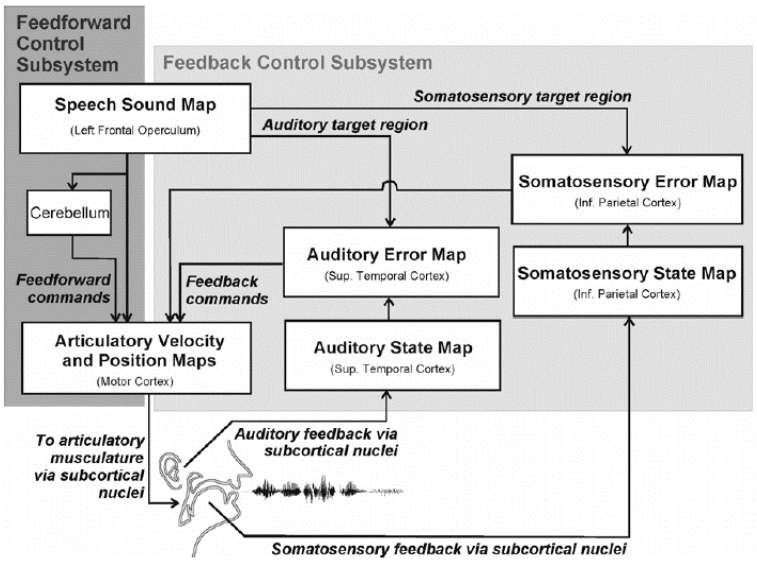

FIGURE 1.4. The DIVA model (from Guenther, 2006): a feedforward control subsystem located in the motor cortex is coupled to a parieto-temporal feedback control subsystem.

In a similar way, sensory-motor theories of speech perception argue for a core auditory (or audio-visual) system for speech perception, enhanced by motor

processes in complex conditions such as noise, through "binding" (Schwartz et al., 2010, see Fig. 1.5) or "prediction" (Skipper et al., 2007).

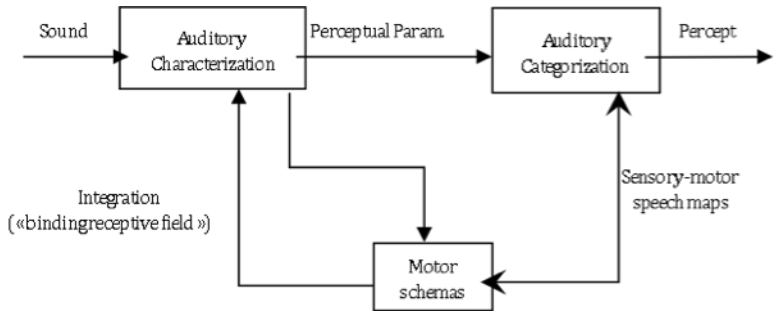

FIGURE 1.5. Perception for Action Control Theory (PACT): motor schemas are involved both in extracting relevant auditory information through binding, and in improving categorization through sensory-motor maps.

TAXONOMY

This review of speech production and perception theories and models shows that both fields share the same debates about the nature of the reference frame and the functionality of the perceptuo-motor link. Table 1.1 thus proposes an original classification of all these theories in a unified way. The next section will formalize this classification in a Bayesian framework. In other words, we aim at filling Table 1.1 with probabilistic expressions.

Table 1.1 Taxonomy of speech production and perception theories and models

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

A UNIFIED BAYESIAN FRAMEWORK

In this section, we define in probabilistic terms what are a motor subsystem, a sensory subsystem, and a sensory-motor link, and how they can be combined in a

general system for communication. We then show how selectively disabling subsystems leads to a unified expression of the six categories of Table 1.1 in probabilistic terms. We limit our analysis to an abstract model from which we extract possible interpretations. The reliability of this model in realistic simulations of language emergence has been studied in Moulin-Frier et al. (2008, 2010).

BAYESIAN MODELING

Bayesian Robot Programming (BRP, Lebeltel et al., 2004) specifies the knowledge of a sensory-motor agent as a joint probability distribution over variables of interest (typically motor, sensory and internal variables). This joint distribution is generally expressed as a product of simpler distributions, using Bayes rule and conditional independence hypotheses. Using this knowledge, a sensory-motor behavior is then defined as a conditional probability distribution computed from the joint distribution (for example: "given the values of some sensory variables, what is the probability distribution over the motor variables"), called a question to the model.

Our modeling of a general communication system is based on four variables (but we will discuss in more details the possible interpretations):

- M: the speaker's motor gesture,
- S: the listener's sensory percept,
- O_S, O_L: the object of communication (in a very general sense, hereafter the object), respectively from the speaker and the listener point of view.

We then define the three subsystems as follow:

- The motor subsystem is defined as a conditional probability distribution $P(M|O_S)$: given an object to communicate, what is the probability distribution over the speaker's motor gesture?
- The sensory subsystem is defined as a conditional probability distribution $P(O_L|S)$: given a sensory percept, what is the probability distribution over the objects that can be inferred by the listener by the listener?
- The sensory-motor subsystem is defined as a conditional probability distribution $P(S|M)$: given a motor gesture, what is the probability distribution over the sensory percepts?

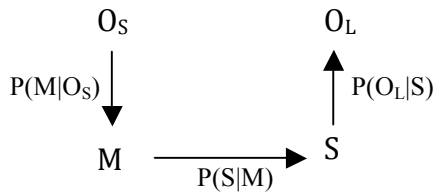

FIGURE 2.1. Model structure.

Finally, considering the *a priori* knowledge on the speaker's object $P(O_S)$ as uniform, the general communication system is a joint probability distribution:

$$P(O_S \wedge M \wedge S \wedge O_L) = P(O_S)P(M \mid O_S)\, P(S \mid M)\, P(O_L \mid S)$$

In this framework, a successful communication corresponds to equality between O_S and O_L (the object inferred by the listener must be the same as the object intended by the speaker). Therefore, each question asked to the model will be under the constraint $O_S=O_L$.

Speech production and perception are defined as the following probabilistic questions to the general model $P(O_S \wedge M \wedge S \wedge O_L)$:

- Production: $P(M|O_S=O_L)$. For a given object in the speaker's mind, and knowing that O_L is equal to O_S to evoke the same object in the listener's mind, what is the probability distribution over motor gestures?
- Perception: $P(O_S=O_L|S)$. Knowing the sensory input perceived by the listener, what is the probability distribution over objects, inferred by the listener and likely to have been in the speaker's head at the input?

In probabilistic terms, motor and sensory subsystems can be deactivated by setting the corresponding distribution as uniform (that is, without explicit knowledge about the corresponding link). Motor, auditory and sensory-motor theories of speech communication will thus be expressed as the following:

- Motor theories correspond to a deactivation of the sensory subsystem, defining $P(O_L|S)$ as a uniform distribution,
- Auditory theories correspond to a deactivation of the motor subsystem, defining $P(M|O_S)$ as a uniform distribution,
- Sensory-motor theories let both the motor and sensory subsystem active, each distribution being considered as informative.

Table 2.1 Model Taxonomy

Task / Theory	**Production** $P(M\|O_S=O_L)$	**Perception** $P(O_S=O_L\|S)$
Motor $P(O_L\|S)$=Uniform	$P(M\|O_S)$	$\sum_M P(M\|O_S)P(S\|M)$
Auditory $P(M\|O_S)$=Uniform	$\sum_S P(S\|M)P(O_L\|S)$	$P(O_L\|S)$
Sensory-motor	$P(M\|O_S)\sum_S P(S\|M)P(O_L\|S)$	$P(O_L\|S)\sum_M P(M\|O_S)P(S\|M)$

Finally, using these definitions and rules of Bayesian inference (typically Bayes and normalization rules), we can now assign a probabilistic expression to each type of theory (Table 2.1).

INTERPRETATIONS

General Model

The general model in Figure 2.1 can be interpreted from different points of view:

- As an objective model of communication, where the motor model is a model of the speaker, the sensory one a model of the listener, and the perceptuo-motor link is a model of the environment.
- As a subjective neurolinguistic model, where the motor and sensory models would correspond respectively to the motor and auditory cortices, and the perceptuo-motor link as the neural connections between them.
- As a subjective model of the Theory of Mind, where the motor model (resp. the sensory model) would be an internal representation of the speaker (resp. the listener) in the brain of the listener (resp. the speaker).

Let us focus on the computational interpretation of the behaviors defined by the probabilistic questions in Table 2.1.

Behaviors: Motor theories

In our Bayesian framework, motor theories of speech communication correspond to a deactivation of the sensory subsystem, setting $P(O_L|S)$ as uniform. Speech production thus leads to select a motor gesture M for a given object to communicate O_S according to the distribution $P(M|O_S)$, considering that the sensory subsystem does not provide any information. This is in line with Articulatory Phonology (Browman and Goldstein, 1989), which considers speech production as motor gestures scores, not influenced by the auditory consequence of those gestures.

Conversely, motor theories of speech perception are defined by the probabilistic question:

$$\sum_M P(M|O_S)P(S|M)$$

For a given auditory percept S heard by the listener, the inferred object thus corresponds to an object for which the speaker would have produced a gesture with the same auditory consequence. This is in line with the Motor Theory of speech perception (Liberman and Mattingly, 1985), which considers that perceiving speech actually amounts to perceiving the intended gestures of the speaker.

Behaviors: Auditory theories

Auditory theories of speech communication correspond in our framework to a deactivation of the motor subsystem, setting $P(M|O_S)$ as uniform. Speech production is then defined by the probabilistic question:

$$\sum_S P(S|M)P(O_L|S)$$

For a given object to communicate by the speaker, the selected motor gesture thus corresponds to a gesture for which the auditory consequence would allow the speaker to correctly infer the object using his/her sensory subsystem. This is in line with models considering that speech production targets are defined by regions in the acoustic/auditory space (Guenther et al., 1998).

Regarding speech perception, the probabilistic question is simply $P(O_L|S)$, without any information from the motor subsystem. This is in line with the claim that speech perception does not incorporate any input from speech production mechanisms (Diehl et al., 2004).

Behaviors: Sensory-motor theories

Finally, sensory-motor theories correspond in our framework to activating both the motor and the sensory subsystems, leading to distributions which are the products of those for motor and auditory theories. Speech production is thus defined by the probabilistic question:

$$P(M|O_S)\sum_S P(S|M)P(O_L|S)$$

For a given object to communicate, the selected motor gesture is then a compromise between an often-used gesture, and a gesture for which the auditory consequence would allow the speaker to correctly infer the object. This is in line with models like DIVA (Guenther, 2006) with its two components, feedforward for on-line production (the first factor) and feedback for correction (the second factor).

Regarding speech perception, the corresponding probabilistic question is:

$$P(O_L|S)\sum_M P(M|O_S)P(S|M)$$

For a given sound heard by the listener, the inferred object has both to satisfy the sensory subsystem and to correspond to an object for which the listener would have produced a motor gesture with the same auditory consequence. This is in line with the Perception for Action Control Theory (Schwartz et al., 2010), which considers the cues of speech perception as essentially auditory (the first factor), but possibly helped by access to motor knowledge (the second factor).

Conclusion and Perspectives

In this paper, we proposed a unified formal framework for speech production and perception, based on a Bayesian model able to express the major theories in the field.

In further works, our aim is to computationally study the possible functional role of the perceptuo-motor link in speech communication. Previous works (Moulin-Frier et al., 2008, 2010) already showed that this link is necessary in production for

realistic simulations of language emergence (backed by data showing that human phonological systems are optimized for perceptual distinctiveness).

Regarding speech perception, we are planning simulations showing that in simple cases like vowel categorization, the sensory subsystem is better than the motor one to infer the corresponding object (in favor of auditory theories of speech perception) but, in more complex cases like syllable categorizations, the motor subsystem can add reliable information (in favor of sensory-motor theories).

REFERENCES

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

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). "Speech Perception." *Annual Review of Psychology*, 55, 149-179.

Guenther, F.H., Hampson, M., & Johnson, D. (1998). "A Theoretical Investigation of Reference Frames for the Planning of Speech Movements." *Psychological Review*, 105, 611–633.

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

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

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

Liberman, A.M., & Mattingly, I.G. (1985). "The Motor Theory of Speech Perception Revised." *Cognition*, 21, 1–36.

Moulin-Frier, C., Schwartz, J. L., Diard, J., & Bessiere, P. (2008). *Emergence of a language through deictic games within a society of sensori-motor agents in interaction.* 8th International Seminar on Speech Production, ISSP'08, Strasbourg, France.

Moulin-Frier, C., Schwartz, J. L., Diard, J., & Bessiere, P. (2010). "Emergence of phonology through deictic games within a society of sensori-motor agents in interaction". Book chapter in *Vocalization, Communication, Imitation and Deixis*, to appear.

Rizzolatti, G., & Arbib, M. A. (1998). "Language within our grasp". *Trends in Neurosciences*, 21, 188–194.

Savariaux, C., Perrier, P., Orliaguet, J.-P., & Schwartz, J.-L. (1999). "Compensation strategies for the perturbation of French [u] using lip-tube. II. Perceptual analysis." *Journal of Acoustical Society of America*, 106, 381–393.

Schwartz, J.-L. Basirat, A., Menard, L., Sato, M., (2010). "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception." *Journal of Neurolinguistics (in press).*

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.