

Building a talking baby robot

A contribution to the study of speech acquisition and evolution

J. Serkhane⁽¹⁾, J.L. Schwartz⁽¹⁾, P. Bessière⁽²⁾

(1) ICP, Grenoble (2) Laplace-SHARP, Gravir, Grenoble

Abstract

Speech is a perceptuo-motor system. A natural computational modeling framework is provided by cognitive robotics, or more precisely speech robotics, which is also based on embodiment, multimodality, development, and interaction. This paper describes the bases of a virtual baby robot, which consists in an articulatory model that integrates the non-uniform growth of the vocal tract, a set of sensors, and a learning model. The articulatory model delivers sagittal contour, lip shape and acoustic formants from seven input parameters, which characterize the configurations of the jaw, the tongue, the lips and the larynx. To simulate the growth of the vocal tract from birth to adulthood, a process modifies the longitudinal dimension of the vocal tract shape as a function of age. The auditory system of the robot comprises a “phasic” system for event detection over time, and a “tonic” system to track formants. The model of visual perception specifies the basic lips characteristics: height, width, area and protrusion. The orosensorial channel, which provides the tactile sensation on the lips, the tongue and the palate, is elaborated as a model for the prediction of tongue-palatal contacts from articulatory commands. Learning involves Bayesian programming, in which there are two phases: (i) specification of the variables, decomposition of the joint distribution and identification of the free parameters through exploration of a learning set, and (ii) utilization which relies on questions about the joint distribution.

Two studies were performed with this system. Each of them focused on one of the two basic mechanisms, which ought to be at work in the initial periods of speech acquisition, namely vocal exploration and vocal imitation. The first study attempted to assess infants’ motor skills before and at the beginning of canonical babbling. It used the model to infer the acoustic regions, the articulatory degrees of freedom and the vocal tract shapes that are the likeliest explored by actual infants according to their vocalizations. Subsequently, the aim was to simulate data reported in the literature on early vocal imitation, in order to test whether and how the robot was able to reproduce them and to gain some insights into the actual cognitive representations that might be involved in this behavior.

Speech modeling in a robotics framework should contribute to a computational approach of sensori-motor interactions in speech communication, which seems crucial for future progress in the study of speech and language ontogeny and phylogeny.

1. Introduction

1.1. Linking perception and action in speech robotics

Speech perception and production are often studied independently one of the other. However, speech is obviously a sensori-motor system. This is the starting point of the so-called "Perception-for-Action-Control" Theory (PACT) (Schwartz *et al.*, 2002), in which we argue that *perception is the set of tools, processing and representations that enable to control action*. The PACT proposes that, as the perceptual and the motor representations are acquired together during speech development, they constrain each other in adulthood, although they belong to different domains. The main idea is that to study the perceptual and the motor representations that underlie speech in adult and that shape world's languages, a relevant strategy is to focus on how they develop in concert with each other during speech acquisition.

In this approach, a natural computational modeling framework is provided by cognitive robotics, a promoter of which is R. Brooks through the Cog project, that focuses on the notions of "[...] *embodiment and physical coupling, multimodal integration, developmental organization, and social interaction.*" (Brooks *et al.*, 1999).

Embodiment, multimodality, development and interaction are also the core of "Speech Robotics" (Abry & Badin, 1996; Laboissière, 1992), a research program in which we try to:

1. elaborate a sensory-motor virtual "robot" able to articulate and perceive speech gestures (embodiment: Boë *et al.*, 1995a; Schwartz & Boë, 2000) and to learn multisensorial-motor links (multimodality: Schwartz *et al.*, 1998) in parallel to the growth of its vocal apparatus;
2. determine what could be the exploration strategies by which this robot could evolve from vocalizing and babbling to the control of complex speech gestures (development: Abry *et al.*, 2004);
3. explore how communication principles in a society composed of such agents could shape the acoustic and articulatory structures of human languages (interaction: Berrah *et al.*, 1996).

The present project concerns a preliminary stage of this research program. It aims at giving the bases to model speech development, that is, the implementation of the virtual baby robot, which is a growing sensori-motor system able to learn and to interact (Schwartz *et al.*, 2002).

1.2. A viewpoint on speech development

The viewpoint supported is that the development of orofacial control in speech relies on two fundamental behaviors: the progressive exploration of the vocal tract sensori-motor abilities, and the imitation (overt simulation) of caretakers' language sounds. That is to say, articulatory exploration should be the way by which infants discover abilities of their vocal tracts and learn relationships between movements and percepts. At the same time, imitation ought to capitalize on the knowledge acquired by exploration to tune step by step the control of the articulatory system so as to produce the gestures and sounds of the target languages.

The first attempts to simulate speech development in robotics were based on the assumption that infants explore their entire space of articulatori-acoustic realizations then select their native language items out of all the possible ones (Bailly, 1997; Guenther, 1995). In other words, infants were supposed to start by uttering all possible speech sounds, in languages (in agreement with Jakobson, 1968). However, direct observation shows that infants do not do so (Kent & Miolo, 1995): whatever their ambient language, they only produce a certain subset of what can be performed with their phylogenetically inherited sensori-motor apparatus. Moreover, on a computational level, exhaustive exploration complicates the learning of sensori-motor links (Bessière, 2000).

Infants do not explore the whole articulatori-acoustic space in order to master their vocal tract behaviors. Further, sensori-motor developmental facts, likely to be linked with speech development, can be classified according to whether they are roughly a matter of exploration or of imitation. (a) At birth, infants are able to imitate three gestures from vision: tongue and lips protrusions, and mandible depression (Meltzoff, 2000). Although these movements, employed in adult speech, are not obviously linked with speech development, they are nonetheless available before first vocalizations. (b) At a few weeks old, infants vocalize. Moreover, they tend to direct their productions towards vowel sounds they often perceive (early vocal imitation: Kuhl & Meltzoff, 1996), and to match a vowel sound to the moving image of the face that utters it (multimodal integration: Kuhl & Meltzoff, 1992). (c) At about seven-month old, they become babblers: their mandibles move upwards and downwards in a rhythmic way, while their vocal folds vibrate. This is what has been referred to as Canonical Babbling (Koopmans-Van Beinum & Van Der Stelt, 1986; MacNeilage & Davis, 1990). (d) Later on, children begin to control, more or less successively, the number of jaw cycles, the movements of the articulators carried by each cycle independently one of each other, and finally the full shape of their “vocal resonator” (motor coordination). This enables them to master sounds and sequential patterns of their ambient languages (Vilain et al., 2000).

Section 2 depicts the sensory, the articulatory and the learning models the virtual robot is made of. At first, the aim was to specify its early motor skills: articulatory exploration was assessed from the acoustic description of vocalizations produced by actual infants both before (b phase) and at the beginning (c phase) of canonical babbling (Section 3; and see Serkhane et al., 2002). As for the imitation issue, a model of imitation was proposed and capitalized on to simulate an experiment on actual infants. The influence of parameters that tune the robot first imitation abilities were studied and lead to gain some information about the sensori-motor representation likely to underlie this behavior in infancy (Section 4; and see Serkhane et al., 2003). Section 5 gives some plans for the future of this project in relation to ontogeny and phylogeny.

2. The vocalizing baby robot

On the production level, the *Variable Linear Articulatory Model* (VLAM, Boë, 1999) provides the robot with a virtual vocal tract that integrates the non-uniform growth of human tract. As for perception, the auditory, the visual and the tactile modalities are available with a model per each modality. The relationships between the tract movements and their perceived consequences are learned (during exploration) and used (in imitation) within a Bayesian robotics formalism.

2.1. The articulatory model

The *Variable Linear Articulatory Model* (VLAM) is a version of the *Speech Maps Interactive Plant* (SMIP, Boë et al., 1995a) that integrates a model of the vocal tract growth. The core of the SMIP is Maeda’s model (Maeda, 1989) or a variant proposed by Gabioud (1994). Its elaboration consisted of a thorough statistical analysis of 519 hand-drawn midsagittal contours corresponding to a 50 frames/sec. radiographic film synchronized with a labiographic film that contained 10 sentences in French, recorded at the Strasbourg Institute of Phonetics (Bothorel et al., 1986). The midsagittal contours were analyzed with a semi-polar grid, and a guided principal component analysis found that seven parameters explained 88 % of the variance in the observed tongue contours, for the selected (adult) speaker. A linear combination of the seven parameters enables the regeneration of a

midsagittal contour of the vocal tract. The weighting values of each parameter were normalized, using the standard deviation around the mean position of the observed values as reference. The lips shape was modeled from measurements analyzed at ICP (Abry & Boë, 1986; Guiard-Marigny, 1992).

Hence the articulatory model delivers sagittal contour and lips shape from the seven input parameters (hereafter P_i , $i=1..7$), which may be interpreted in terms of phonetic commands, and correspond respectively more or less to the jaw (J), the tongue body (TB), dorsum (TD) and tip (TT), the lip protrusion (LP) and height (LH), and the larynx height (Lx) (Fig. 1). The area function of the vocal tract is estimated from the midsagittal dimensions with a set of coefficients derived from tomographic studies. The formants and the transfer function are calculated from the area function, and a sound can be generated from formant frequencies and bandwidths.

From this basis, it was possible to implement a growth model that enables to replace the adult “robot” by a “baby” one. Systematic measurements of the vocal tract from birth to adulthood do not exist at present. However, it was possible to take advantage of cranio-facial measures established at different ages by Goldstein (1980). These data were closely fitted by (double) sigmoidal curves, which characterize the general skeletal and muscular growth. To give account of the vocal tract growth, the articulatory VLAM model (*Variable Linear Articulatory Model*), developed by Maeda (cf. Boë & Maeda, 1998) describes the evolution of the horizontal and vertical dimensions from a newborn to a female or to a male adult. As proposed by Goldstein, the growth process was introduced by modifying the longitudinal dimension of the vocal tract according to two scaling factors: one for the anterior part of the vocal tract and the other for the pharynx, interpolating the zone in-between. So, the non-uniform growth of the vocal tract can be simulated year-by-year and month-by-month. Similarly, typical F_0 values were adjusted to follow the growth data presented by Mackenzie Beck (1997). A more detailed presentation of the model, together with the assessment of its agreement with both morphological and acoustical data on infants and children, can be found in Ménard et al. (2002, 2004).

2.2. The sensory models

2.2.1. Auditory model

The tracking of speech gestures must involve a way to capture and characterize the basic components of the speaker's vocal actions, namely timing and targets (Schwartz et al., 1992). A series of influential works realized in the Pavlov Institute of Leningrad in the 70s led Chistovich to propose a basic architecture for the auditory processing of speech sounds. It consists of one system specialized in temporal processing and detection of acoustic events, and the other continuously delivering various analyses about the spectral content of the input (Chistovich, 1976, 1980). The neurophysiological bases for these processing are already available in primary neurons in the auditory nerve, or secondary neurons in the cochlear nucleus (which is the first auditory processing center in the central nervous system). This provides the basis of the auditory system of the robot (Fig. 2).

The system specialized in event detection is based on so-called “phasic” units in the central nervous system, namely “on” and “off” units responding only to quick increases and decreases of the neural excitation in a given spectral region. We developed a physiologically plausible module for the detection of articulatory-acoustic events such as voicing onset / offset, bursts, vocalic onset / offset (Piquemal et al., 1996; Wu et al., 1996) in the cochlear nucleus. These events, which allow the labeling of every major discontinuity in the speech signal, are crucial for the control of timing in speech production (Abry et al., 1985, 1990).

The system specialized in spectral processing needs so-called “tonic” units responding continuously to a given stimulus, and then enabling precise statistics and computations about

the variations of excitation depending on their characteristic frequency. Though the debate on the role of formants in the auditory processing of speech is far from being closed (e.g. Bladon, 1982; Pols, 1975), it seems that basic neurophysiological ingredients are available for formant detection in the auditory nerve, through spatio-temporal statistics (Delgutte, 1984); and higher in the auditory chain, as early as in the cochlear nucleus, through lateral inhibition mechanisms for contrast reinforcement. Hence formants are the basic spectral parameters characterizing speech sounds in our system.

2.2.2. *Visual model*

In the multisensorial framework, the robot needs eyes as much as ears. Indeed, it is quite well known that speech is not only heard but also seen (e.g. Dodd & Campbell, 1987; Campbell et al., 1998). Speechreading enables to partly follow speech gestures when audition lacks, particularly in hearing impairment; it improves speech intelligibility in noisy audio condition or with foreign languages; it intervenes in gesture recovery even if the visual input is conflicting with the audio one, as in the famous McGurk effect (McGurk & MacDonald, 1976); and the visual input is implied in the development of speech control, and in the acquisition of phonology in conjunction with cued speech for hearing-impaired people (see Schwartz et al., 2002, for a review of audio-visual fusion in the context of a theory of perception for action control). The visual sensor should be able to capture what can be seen on the speaker's face, that is lip geometry, jaw position, and probably some parts of the tongue. At present, the visual inputs of the robot are the basic lips characteristics: height, width, area and protrusion.

2.2.3. *Tactile model*

The orosensorial channel, that contains the tactile sensation on the lips, the tongue and the palate, is most often absent from the modeling of the perceptual representation of speech gestures. However, Human possesses a highly developed representation of the oral space. This is illustrated by data on oral stereognosis in which subjects are able to integrate tactile and motor information to identify three-dimensional objects placed in their mouths (Bosma, 1967). The tip of the tongue and the lips belong to the most sensitive parts of the body surface, as displayed by two-point discrimination data. The neurophysiology of the tactile orosensory system has been described in a number of reviews (see e.g. Hardcastle, 1976; Landgren and Olsson, 1982; Kent et al., 1990). Most of the oral mucosa, and particularly the tongue, is supplied with mechanoreceptors of different types, able to provide detailed information about the position of the jaw, lips and tongue, and the velocity and direction of movement. Histological data show that the density of sensory endings decreases progressively from the front to the rear of the mouth: the tip of the tongue seems the best endowed with receptors in the oral system, in agreement with its accurate tactile acuity. Several data show the importance of the tactile sensor for speech control. MacNeilage et al. (1967) cited the case of a patient with a generalised congenital deficit in somesthetic perception: she produced totally unintelligible speech though she had no apparent auditory or motor trouble. Hoole (1987) and Lindblom et al. (1977) showed the influence of oral sensitivity for the production of perturbed speech (bite-block experiments).

The above facts motivated the elaboration a model for the prediction of palatal contacts of the tongue from articulatory commands (Schwartz & Boë, 2000). In this model, patterns of palatal contacts are described by five variables (hereafter L_i , $i=1..5$), defining the number of contacts per line along five lines that go from the periphery to the middle of the palate (Recasens, 1991) (Fig. 3). The L_i values are predicted from the articulatory commands P_i by a linear-with-threshold associator:

$$L_i = f(\sum w_{ij}P_j + w_i0)$$

where w_{ij} and w_{i0} are the weights and the bias to learn, and f is a threshold function limiting L_i to their ranges of variation, that is from 0 (no palatal contact in the corresponding line) to their maximal possible value (respectively 9, 8, 7, 5, 4). The values of w_{ij} and w_{i0} were tuned by minimizing a summed square error between observed and predicted L_i values (Fig. 4a).

To test the behavior of this model, a set of predicted palatal contacts were computed for a great number (about 1,000) of articulatory configurations that lead to formant frequencies in the acoustic regions of the vowels [i], [a], [u]. Though these configurations vary largely in their articulatory parameters, it appeared that the predicted palatal contacts were quite coherent (Fig. 4b), and in line with the variability of contacts observed by Recasens (1991) for vowels embedded in various consonantal contexts. Hence, the model seems able to provide useful predictions, adequately linked with the articulatory and acoustical structure of the gesture.

2.2.4. Simplified perceptual models

In order to focus on learning problems, we chose, in this study, to take into account a restricted and simplified set of sensory variables, easily interpretable in phonetic terms.

The auditory variables were the two first formant frequencies (F1, F2) expressed in Bark, that is, a scale of frequency perception (Schroeder et al., 1979):

$$z(\text{Bark}) = 7 \text{Arg sh} \left(\frac{F(\text{Hz})}{650} \right)$$

The simplified tactile system relied on the vocal tract geometry, which can be described by the following systems (Boë et al., 1995b): (i) the area (A_c) and the distance from the glottis (X_c) of the main constriction along the vocal tract, as well as the inter-lip area (A_l) when produced by robot vocal tract [(Fant, 1960), (ii) the coordinates (X_h , Y_h) of the tongue's highest point in a fixed system of reference (Boë et al., 1992). The visual system estimates A_l when it comes from peer's face. This set of variables is displayed on Fig. 5.

2.3. The model of sensori-motor learning

Learning here involves two steps, which may be synchronized in time but are studied separately in a first stage. Firstly, the robot learns basic relationships between motor commands and sensory inputs, by an endogenous exploration process (only driven by internal motivation). Secondly, the robot attempts to reproduce a given sound presented by speaking partners, given the knowledge acquired by exploration. In the future, this exogenous stage (driven by external stimuli provided by the environment) will also contribute to learning so as to focus the robot inventory of actions and percepts on the patterns of its ambient language.

The Bayesian Robot Programming (BRP) environment developed for general robot programming by Lebeltel *et al.* (2003) is capitalized on to implement the learning and the imitation behaviors. The theoretical foundations of BRP come from the analysis of the central difficulty faced by a robot system, namely, how to use an incomplete model of its environment to perceive, infer, decide and act efficiently? To address this problem BRP proceeds in 2 steps:

- The first step (learning) transforms the irreducible incompleteness into uncertainty. Given some preliminary knowledge (supplied by the designer) and some experimental data (acquired by the robot), learning builds a description of the phenomenon, which takes the mathematical form of a probability distribution. The maximum entropy principle is the theoretical foundation of this first step. Given some preliminary knowledge and some data, the probability distribution that maximizes the entropy is the distribution that best represents this couple. Entropy gives a precise, mathematical and quantifiable meaning to the "quality" of a distribution (for justifications of the maximum entropy principle see, for instance

Bessière *et al.*, 1998a & 1998b). Preliminary knowledge, even imperfect and incomplete, is relevant, and provides interesting hints about the observed phenomenon. The resulting descriptions give no certainties, but they provide a means to take the best decision given the available information.

- The second step (reasoning) is in charge of making inferences with the probability distributions obtained at the first step. The BRP formalism is very general and encompassed for instance the following particular cases: Bayesian net (Pearl, 1988), Hidden Graphical Models (Lauritzen & Spiegelhalter, 1988; Lauritzen, 1996; Jordan, 1998; Frey, 1998), Markov Localization (Thrun, 1998) and Partially Observable Markov Decision Processes (Kaelbling, Littman & Cassandra, 1996).

BRP uses a strict and systematic methodology to model a phenomenon. It always proceeds as follows:

A – Learning:

A1. Specification: define the preliminary knowledge

A1.1 - Choose the variables relevant with the behavior to model

A1.2 - Decompose the joint distribution of the set of relevant variables as a product of simple distributions

A1.3 - Define the parametric forms of the simple distributions

A2. Identification: identify the free parameters of the simple distributions

B – Reasoning: Utilization: ask a question about the joint distribution

During specification (A1), the variables that define the problem to be modeled are chosen. In the case of the present work, these variables were articulatory and perceptual parameters dealt with earlier. In order to constrain the problem, the decomposition of their joint distribution takes into account the relationships between the different variables, given physical and phonetic pieces of knowledge. More precisely, the purpose was to build-in their assumed (in)dependencies to each other, be they conditional or not. Distributions affiliated to each variable (or parametric form) were chosen as being a Gaussian or a uniform law. Within this robotic framework, *exploration* takes place during identification, which consists of providing the robot with experimental data for its simple distributions to be actually implemented. For example, if the decomposition of the joint probability contains a Gaussian law, the associated free parameters, that is, the mean and the variance, are worked out from the set of experimental data, and this simple distribution is therefore considered as learned by the robot. At the end of learning, a description of the sensori-motor system is obtained. The robot can use it to solve problems such as inversion. *Imitation* requires inversion whose associated question is mainly "which articulatory configuration could lead to the target percept"?

Section 4 will describe in more details how this framework was made use of in the case of the baby-robot. However, in order to be as realistic as possible the robot had to be specified in connection with actual data. This is the purpose of the next section.

3. Simulating vocal exploration before and at the beginning of babbling

As infants do not start by exploring all possible speech sounds, we first tried to assess articulatory abilities available both before and at the beginning of canonical babbling, that is, at 4 and 7 months. To obtain this information from the two first formant frequencies of vocalizations produced by real subjects at these developmental stages, three specially designed analysis techniques were developed. They were termed acoustic framing,

articulatory framing and geometric framing. Their description and results will be given after the data they processed are presented.

3.1. Phonetic data

We had two sets of data from studies in developmental phonetics. The first one is composed of vowel-like sounds produced by 4-month old subjects, during early vocal imitation tests from Kuhl and Meltzoff (1996), (see Section 4.1 for further details). Matyear and Davis supplied us with the second set of data, collected in order to study syllable-like productions in Canonical Babbling (Matyear, 1997; Matyear *et al.*, 1998). We selected their 7-month old subjects' vowel-like sounds, at Canonical Babbling onset. These two studies present the interest to have been carefully acquired and carefully labeled and analyzed in a series of paradigms and protocols described in great detail in the original publications. In each case, the two first formant values and a phonetic description were available for analysis.

3.2. Acoustic framing

3.2.1. Method

All the sounds generated by the VLAM belong to the Maximal Vowel Space (MVS) (Boë *et al.*, 1989). MVS corresponds to what an infant at a given age would be able to produce if s/he used the complete set of their articulatory commands, defined as all values between -3 and $+3$ times the standard deviation, that is covering the whole range of possible values for each parameter. So, it stands for all “possible speech sounds” plotted on a multi-formant (Fi) map. The (F1, F2) plane displays the vocalic triangle, attested by phoneticians and whose corners include the [i a u] vowels. The acoustic framing consists of superimposing an age-specified set of actual data on the MVS of the VLAM at the same age. Hence, it tests whether actual vocalizations belong to this MVS and assess the acoustic space region(s) explored by 4- and 7- month old infants.

3.2.2. Results

Each set of actual vocalizations did belong to the corresponding MVS (Fig. 6 - 7). Moreover, the actual data did not entirely cover the space they would have if they had corresponded to mature motor control products. More precisely, the 4-month old vocalizations, displayed as black dots superimposed on the MVS in gray in Fig. 6, were relatively centered and mid-high: the most fronted, backed and open productions did not seem to be exploited. At 7-month old (Fig. 7), the vocalic productions exploited the high-low dimension more than at the earlier stage.

3.3. Articulatory framing

3.3.1. Method

Certain regions of the MVS, generated by the 7 articulatory parameters of the VLAM, were not exploited by the actual data. The articulatory framing allowed to evaluate infants' motor abilities by constraining the motor variables of the VLAM. In other words, this method aims at assessing the minimal set of articulatory degrees of freedom required to reproduce the observed vocalic sounds. We built several articulatory sub-models with different sets of the VLAM motor parameters. A given sub-model was therefore characterized by its number of articulators and their ranges of variation. 255 sub-models were comparatively assessed with respect to the efficiency by which they reproduced each collection of phonetic data, using their probabilities given the actual vocalizations: $P(M_i/f1f2)$, where M_i denotes the i^{th} sub-model, characterized by the set of acoustic outputs it generates, while $f1f2$ stands for the

actual data formant values. The winner is the sub-model that fitted the best a given set of actual data: it maximized the conditional probability criterion.

3.3.2. Results

The results for the 4-month old data indicate that exploration at four months is rather reduced around the neutral configuration. It involves at least three articulatory parameters including at least one for the tongue, and the jaw seems to play a minor role in this exploration. The winner sub-model (Fig. 8) exploited the lower lip height (LH), tongue body (TB) and dorsum (TD) degrees of freedom. At seven months, exploration is much larger, and jaw now plays a dominant role leading to a large exploration of the open-close contrast and its associated F1 dimension in the formant space (Fig. 9). This result agrees with babblers' mandible use.

3.4. Geometric framing

3.4.1. Method

Articulatory framing enabled to infer the tongue configurations that could have yielded the acoustic data recorded at 4 and 7 months. The geometric framing is a method of exhaustive inversion: each vocalization corresponds to a set of tract shapes (geometry), produced by the winner and corresponding to acoustically plausible products. Two systems were capitalized on to describe the vocal tract geometry (see Section 2.2.4): { Xc, Ac, Al } and { Xh, Yh }. Thus, a given vocalic sound could be associated with the mean and variance of these geometric variables in the group of corresponding tract shapes. As compensation leads to rather high variances, especially in central vocalizations, for clarity's sake, we only displayed the dispersion ellipses of 4 "prototypes" added to each set of real data: [i a u] had been chosen at a roughly equivalent position to the adult's in the MVS, whereas [ə] was produced by all commands set to 0. So, [i a u ə] served as landmarks.

3.4.2. Results

At 4 months (Fig. 10), the average tract shapes (plotted by gray stars on the figure) had tongue's highest points rather centered and gathered (around [ə]). The constrictions were slightly fronted and fairly wide. At 7 months (Fig. 11), the tongue positions showed a larger exploration of the high-low and front-back dimensions than at the earlier stage. Moreover, we found that, before canonical babbling (4 months), all the articulatory configurations leading to first two formant frequencies falling within the [u] region had palatal constrictions. This result is of interest with regard to how the developmental path followed by articulatory exploration may shape adult speech. Indeed, although three types of tract constriction (palatal, velo-pharyngeal and pharyngeal) should be able to produce the vowel [u] with identical first three formants (Boë et al. 2000), the only to be found in the native (adult) speakers of all the languages tested is palatal (Wood, 1979). The velo-pharyngeal [u] is seldom observed in perturbation experiments (lip-tube, Savariaux et al., 1995) while the pharyngeal one has never been recorded. According to Abry and colleagues (1996), the palatal [u] would be the first [u] production strategy picked during speech development: its dominant position in adulthood would stem from its early sensori-motor mapping. This hypothesis is in agreement with the palatal nature of the productions in the acoustic region around [u] in the simulations at four months.

3.5. Conclusion

The results of the simulation of vocal exploration in infants point out that speech development does not begin with exhaustive exploration of the tract potential. We may suggest that "explore all possible speech sounds, then select what is needed to communicate" would be a

much more time- and energy-consuming strategy than, for instance, “explore, according to currently available abilities, and try to produce what is perceived in the ambient language just to develop the motor skills needed”. The second strategy should provide a higher adaptive value than the first one, as more resources would be left for the development of other biological functions. From an evolutionary point of view, this would account for the first strategy counter-selection.

To sum up, before canonical babbling, infants would use the lower lip height (LH), tongue body (TB) and dorsum (TD) commands, which is coherent with newborn imitation studies. Furthermore, the importance of TD is in agreement with its likely role in suckling. The jaw articulator (J) would play only a minor role at this stage, and become significant in canonical babbling data.

4. Simulating early vocal imitation

In this section, we tried to simulate Kuhl and Meltzoff's experiment on early vocal imitation (Kuhl & Meltzoff, 1996), which takes place, at least, before canonical babbling. The purpose was to gain some insights into the cognitive representations that might be involved in early vocal imitation and to test whether and how the robot is able to reproduce, at least, the actual infants' imitation performance. First, an overview of Kuhl and Meltzoff's experiment as well as a description of how the problem was translated into Bayesian terms will be given. Then, the implementation of imitation and the corresponding results will be presented.

4.1. An overview of Kuhl and Meltzoff's experiment on early vocal imitation

72 subjects aged from 12 to 20 weeks old were exposed to audiovisual adult face-voice stimuli corresponding to the vowels [i], [a] and [u]. Only 45 of them happened to produce vowel-like utterances during the experiment. Their subsequent vowel-like productions were phonetically and acoustically described. The system of transcription was that of the set of English vowels but the transcribed items were merged into three main classes: the /a/-like, including [a æ ʌ], the /i/-like, with [i ɪ ε], and the /u/-like for [u u]. Table 1 provides the resulting confusion matrix, that is the number of “i-like”, “u-like” and “a-like” vocalizations (according to the criterion presented here above) for each of the three possible adult targets [i a u]. In sum, the pre-babblers produced vocalic sounds significantly more often categorized as being like the "target" after they had been exposed to this stimulus than otherwise. Globally, the subjects performed around 59 % of responses that are congruent (hereafter %CR) with an imitative behavior. Further, about 16.5%, 47% and 36.5% of their vocalizations sounded /i/-, /a/- and /u/-like, respectively.

4.2 Specifying the model

In the Bayesian robotics framework, the robot learns a sensori-motor map of its vocal tract behavior corresponding to a probabilistic description of the observable links between its perceptual and its articulatory variables. Then, imitation corresponds to inversion, that is, the conversion of a sensory state into a motor counterpart.

The motor parameters chosen were selected according to the results of articulatory framing at 4 months (Section 3.3), i.e. the lower lip height (LH), the tongue body (TB) and dorsum (TD) commands while the auditory variables (Section 2.2.1) were the first two formant frequencies (F1, F2) expressed in Bark. The formants of a vocalic sound are function of the tract shape whose mid-sagittal section can be described by three variables: the inter-lip

area (Al) and the coordinates (Xh, Yh) of the tongue highest point in a fixed system of reference. As mentioned in section 2.2.3, Xh and Yh are potential outputs of the somesthetic system and Al can be either a somatosensory or a visual variable (depending on whether this piece of information comes from self or the other). All the model variables were supposed to be discrete, varying in a set of mutually exclusive values.

The core of a Baseyan robot is the set of statistical relationships that define the links between variables. {Xh, Yh, Al} were used as pivots of the joint probability decomposition. Indeed, since they provide an intermediate space between the auditory space and the articulatory space, they help reduce the impact of the many-to-one problem on inversion (Boë et al., 1992) when they function as independent variables in the joint probability decomposition. Then, further assumptions lead to the following probabilistic structure:

$$\begin{aligned}
 &P(LH \otimes TB \otimes TD \otimes Xh \otimes Yh \otimes Al \otimes F1 \otimes F2) && (1) \\
 &= P(Xh) * P(Yh) * P(Al) \\
 &* P(LH/Al) * P(TB/Xh \otimes Yh) * P(TD/Xh \otimes Yh \otimes TB) \\
 &* P(F1/Xh \otimes Yh \otimes Al) * P(F2/Xh \otimes Yh \otimes Al)
 \end{aligned}$$

This equation specifies the decomposition of the global probability distribution linking all articulatory (LH, TB, TD), intermediate (Xh, Yh, Al) and auditory (F1, F2) variables (first line of Eq. 1). The first three factors (second line) indicate that (Xh, Yh, Al) are considered as the primary variables, supposed to be independent. The next three factors (third line) indicate the minimum set of links between intermediate and articulatory variables: Al specifies the lips (LH), while (Xh, Yh) specify the tongue (TB, TD). The two last factors (last line) express the links between intermediate and auditory variables, supposed to be independent one of the other. In this equation, the independent variables Xh, Yh, Al were associated with uniform distributions, while all other factors were conditional probabilities supposed to obey Gaussian laws, the mean and variance of which had to be tuned in the learning phase.

4.3. Learning the model

To become an actual (and useful) description of the robot's sensori-motor behavior, the distributions composing this probabilistic structure need to be learnt from a set of experimental data that corresponds, here, to a random exploration of the articulatori-geometrico-acoustic skills of the 4-mth robot specified in Section 3.3.2 (R4m in the following). The robot's "proficiency" in inversion, that is, in exploiting its map via Bayesian inference to draw motor values likely to make it reach a given target-state of its perceptual variables, will mainly depend on the learning database size (DBS) and the degree of discretization of the geometric parameters (GDD). Indeed, as Xh, Yh and Al are the pivot of the description, the GDD partly determines the accuracy of the distributions the robot learns: it sets the minimal gap required to distinguish two items in the geometric domain and the size of the learning space, that is, the number of articulatory and auditory distributions that have to be learned for the description to represent the whole range of the R4m abilities. However, there is a trade-off between the GDD and the DBS because a given geometric box must include enough configurations for the matched motor and auditory distributions to be learned.

In order to evaluate which description could best account for the performance reported in Kuhl and Meltzoff (1996), 4 GDD x 15 DBSs were tested. The DBS ranged from 1 to 60,000 items. The GDD were {16, 16, 8}, {8, 8, 4}, {4, 4, 2} and {2, 2, 1} for the number of {Xh, Yh, Al} classes, which yielded 2048, 256, 32 and 4 "boxes" in the geometric space, respectively. In a first step, the GDD/DBS trade-off was studied through the ability of the model to perform inversion of vocalizations in its exploration domain. Figure 12 illustrates the results for the auditory inversion of 1000 items randomly chosen out of the R4m abilities.

At maximal DBS that is for the largest amount of learning, the error decreases, as the GDD increases, and reaches values lower than 0.5 Bk (roughly, formant *jnd*) for the highest two GDD values. Moreover, for a given GDD, the error tends to decrease, along with the DBS rise, until a limit that is the lowest this GDD can make the robot reach. However, all the GDDs, but the roughest, provide unstable scores as long as the DBS is below a certain value. This is due to the fact that not all geometric boxes are actually learned (*under-learning phase*). Indeed, the smallest DBS that is required to have an error at most 10% from the GDD-matched lowest error was found to be three times the size of the geometric space (in boxes). In other words, the more boxes in the geometric space (the larger the GDD is), *the more precise* its variables are, but the *larger the DBS* must be for the robot's map to be representative of its sensori-motor skills (at least three times larger than GDD).

4.4. Implementing auditory and audio-visual imitation

Once a model, defined by a given GDD, has been learned on a given DBS, it can be submitted to imitation tests. Since the experimental data were obtained in an audio-visual configuration, we submitted the robot to two imitation tasks, that is audio-only and audio-visual imitation, to assess the role of multimodality in this framework. In auditory (hereafter *A*) imitation, the perceptual target was the (F1, F2) pair of a vowel, while in the audiovisual (*AV*) one it was its (F1, F2, A1) values. Two target sets were the focus of imitation experiments, that is, “external” and “internal” [i a u] items. The former corresponded to those of the 4 months old VLAM, the latter were their closest simulations within the R4m capacity. This means that both target sets were adapted to the 4-mths articulatory-acoustic space (“normalized” targets), but the first one consisted in [i], [a] or [u] targets outside the true vocalization space at 4 mths, while the second one consisted in the three corners of this space. For each target, 300 motor configurations were drawn from the $P(LH \otimes TB \otimes TD / PerceptualTarget)$ distribution. The formants produced by each articulatory pattern were computed and the sound was categorized as [i], [a] or [u] according to its nearest target in the (F1, F2) plan, in terms of Euclidean distance. This allowed to compute congruent imitation scores %CR for *A* and *AV* imitation, for both external and internal targets, and for various values of GDD and DBS.

4.5. *A* and *AV* imitation results

The congruent response scores %CR as functions of the GDD and the DBS in the *AV* inversion of the internal and external [i a u] targets are displayed in Figures 13 et 14, respectively. *A* inversion scores, not displayed here, are systematically slightly lower. Furthermore, the following trends appear.

GDD/DBS Trade-off and under-learning

Of course, the same GDD/DBS trade-off as in Fig. 12 is found in all cases. Under-learning happens when the imitation scores are lower than their asymptote for a given GDD (DBS not large enough for this GDD), and results in a rather erratic behavior of %CR scores. Globally, under-learning is greater for external than for internal targets, and ends more quickly for *AV* than for *A* imitation.

External vs. internal targets: the risk of over-learning

The scores for external targets are lower than for internal ones, which is quite understandable, considering that the former are outside the R4m vocalization space, while the latter are not. More surprisingly, in the *A* case the imitation score never reaches 100% with external targets even with the highest GDD and DBS configurations, that is 2048 geometrical boxes and 60,000 items in the learning set. This is ascribable to the *over-learning problem*. Indeed, when the description is completely representative of the robot sensori-motor abilities (e.g. with a maximal DBS), all the distributions of the motor variables have small variances, that is, are

very accurate. However, none of them matches the target the robot tends to imitate. Hence, the system draws articulatory configurations regardless of their irrelevance given the sound. In other words, the GDD *has to* contain a *small* number of *large* boxes for the robot to be able to imitate vocalic sounds that are out of its sensori-motor abilities. The problem is overcome if the visual information is also provided: since the VLAM [i a u] inter-lip areas belong to the R4m ones, the robot is enabled to select configurations that produce the nearest sounds to the target.

Early vocal imitation does not need much learning

Altogether, it is striking to notice that the robot needs *neither a high GDD nor a large DBS*, in order to perform as good as, and even better than, the actual infants. For example, in the case of external targets which are out of its motor abilities (which corresponds more closely with the experimental conditions of the imitation data in the Kuhl & Meltzoff study), it gets 60% CR (as infants did) or more with DBSs of 50 and 25 data and GDD of 32 boxes, in *A* and *AV* inversions, respectively.

4.6. Conclusion

The major lesson in this second study is that a very small number of vocalizations (less than a hundred) are necessary for a robotic learning process to provide imitation scores at least as high as those of 20-weeks infants. This is due to the fact that the imitation task studied by Kuhl & Meltzoff is basically a three-categories problem, which can be described rather simply and roughly in articulatory-acoustic terms, hence the success of the present robotic experiment. This shows that actually, more than learning, the problem is of course *control*, that is achieving to produce a desired articulatory configuration ... which the infant is not able to do easily at four months.

The *A* and *AV* imitation experiments displayed a trade-off between the somesthetic acuity of the tract shape representation (GDD) and the amount of information (DBS) to learn in order to build a sensori-motor map that is representative enough of the robot skills. Further, our results show that the GDD *has to* be rough for the robot to be able to imitate vocalic sounds that are out of its articulatori-acoustic abilities. This is interesting since, in fact, the infants must acquire, by imitation, the speech sounds of their ambient languages although they are not endowed from birth with the matched motor capacity. Moreover, this investigation supports the view that the formation of the cognitive representation likely to underlie early vocal imitation would require less learning with audiovisual speech perception than without vision. This gives some evidence that the latter can facilitate phonetic development and is congruent with the slight differences in speech development between seeing and visually impaired children (Mills, 1987).

Allover, this preliminary work confirms that infants complement the very early visuo-facial imitation abilities by using auditory-to-articulatory relationships, and shows that a very small amount of data is enough to produce realistic imitation scores, if the discretization is rough enough.

5. Perspectives in the study of ontogeny and phylogeny

The experiments described here anchor both the production and the perceptual representations of the baby robot in actual infants' perceptuo-motor ground. The continuation of this work will consist in allowing the robot to grow up, mimicking as much as possible the developmental process at work in human speech acquisition. This involves the various steps described in Section 1.3, and particularly the acquisition of frame and content control in the

production of syllables (Davis & MacNeilage, 1995; MacNeilage, 1998). All over this process, an important output of the work will consist in information about the perceptual and the motor representations acquired by the system at the various developmental stages. In a way, it should provide a window on the representations of speech in the baby' and child's brain, which cannot be directly observed by simple means.

Another challenge will be to study how speech as a linguistic system may be patterned by both perceptual and motor constraints. This route towards a "substance-based" approach of phonology, that simulates speech phylogeny, is not new. One of its precursor is found within the Adaptive Variability Theory by Lindblom and colleagues, with a number of important results on the prediction of vowel systems (see e.g. Liljencrants & Lindblom, 1972; Lindblom, 1986, 1990; and the extension we proposed through the "Dispersion-Focalization Theory", Schwartz *et al.*, 1997) and of consonant systems (e.g. Lindblom, 1997; Boë *et al.*, 2000; Abry, 2003). More recently, Steels and others introduced the concept of speech games in societies of talking agents (e.g. Steels, 1998; Berrah *et al.*, 1996; de Boer, 2000). The definition of more realistic agents, able to act, perceive and learn in a biologically, developmentally and cognitively plausible way, is crucial there.

Integrating perception and action within a coherent computational framework is a natural way to better understand how speech representations are acquired, how perception controls action and how action constrains perception. This provides also a natural framework to integrate various sources of knowledge about the speech process, including behavioral and developmental data, neurophysiological and neuropsychological facts about the neural circuits of perception, action and language, and linguistic knowledge about phonology or syntax, and to attempt to draw some links between these complex ingredients in order to begin to write the story of the emergence of human language. We believe that modeling speech communication in a robotic framework should contribute to a computational approach, which is relevant for future progress in the study of speech and language ontogeny and phylogeny.

Acknowledgements

This work was prepared with support from the European ESF Eurocores program OMLL, and from the French funding programs CNRS STIC Robea and CNRS SHS OHLL, and MESR ACI Neurosciences Fonctionnelles. It benefited from discussions with and suggestions from Louis-Jean Boë, Barbara Davis, Chris Matyear, Emmanuel Mazer and Christian Abry.

References

- Abry, C. (2003) [b]-[d]-[g] as a universal triangle as acoustically optimal as [i]-[a]-[u]. *15th Int. Congr. Phonetics ICPHS*, 727-730.
- Abry, C., Badin, P. et al. (1996). Speech Mapping as a framework for an integrated approach to the sensori-motor foundations of language. *4th Speech Production Seminar, 1st ESCA Tutorial and Research Workshop on Speech Production Modelling: from control strategies to acoustics*, 175-184, May 21-24, 1996, Autrans, France.
- Abry, C., Benoît, C., Boë, L.J., & Sock, R. (1985). Un choix d'événements pour l'organisation temporelle du signal de parole. *14èmes Journées d'Etudes sur la Parole, Société Française d'Acoustique*, 133-137.
- Abry, C. & Boë, L.-J. (1986). Laws for Lips. *Speech Communication*, 5, 97-104.
- Abry, C., Cathiard, M.A., Vilain, A., Laboissière, R., Loevenbruck, H., Savariaux, C., & Schwartz, J.L. (2004). Some insights in bimodal perception given for free by the natural time course of speech production. In (E. Vatikiotis-Bateson, G. Bailly, P. Perrier, eds.), *Audiovisual Speech Processing*. MIT Press (to appear).
- Abry, C., Orliaguet, J.P., & Sock, R. (1990). Patterns of speech phasing. Their robustness in the production of a timed linguistic task: single vs. double (abutted) consonants in French. *European Bull. of Cogn. Psych.*, 10, 269-288.
- Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, 22, 251-268.
- Berrah, A.R., Glotin, H., Laboissière, R., Bessière, P., & Boë, L.J. (1996). From form to formation of phonetic structures: an evolutionary computing perspective. In T. Fogarty & G. Venturini (eds.) *ICML '96 Workshop on Evolutionary Computing and Machine Learning* (pp. 23-29). Bari: Italy.
- Bessière, P. (2000). *Vers une théorie probabiliste des systèmes sensori-moteurs*. HDR, Université Joseph Fourier, Grenoble, France.
- Bessière, P., Dedieu, E., Lebeltel, O., Mazer, E. & Mekhnacha, K. (1998a). Interprétation ou description (I) : proposition pour une théorie probabiliste des systèmes cognitifs sensori-moteurs. *Intellectica*, 26-27, 257-311.
- Bessière, P., Dedieu, E., Lebeltel, O., Mazer, E. & Mekhnacha, K. (1998b). Interprétation ou Description (II) : Fondements mathématiques de l'approche F+D. *Intellectica*, 26-27, 313-336.
- Bladon, A. (1982). Arguments against formants in the auditory representation of speech. In *The Representation of Speech in the Peripheral Auditory System* (R. Carlson & B. Granström, eds.), pp. 95-102. Amsterdam: Elsevier Biomedical.
- Boë, L.-J. (1999). Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults. *Proc. XIVth International Congress of Phonetic Sciences*, San Francisco, USA, 2501-2504
- Boë, L.J., Abry, C., Beautemps, D., Schwartz, J.L., & Laboissière, R. (2000). Les sosies vocaliques – Inversion et focalisation. *XXIIIèmes Journées d'Étude sur la Parole*, Aussois, 257-260.
- Boë, L.J., Gabioud, B., & Perrier, P. (1995a). Speech Maps Interactive Plant « SMIP ». *Proc. XIIIth International Congress of Phonetic Sciences*, vol. 2, 426-429, Stockholm, Sweden.
- Boë, L.-J, Gabioud, B., Perrier, P., Schwartz, J.-L., & Vallée, N. (1995b). Vers une unification des espaces vocaliques. In C. Sorin et al. (eds.) *Levels in Speech Communication: Relations and Interactions* (pp. 63-71). Elsevier Science B.V.
- Boë, L.-J, & Maeda, S. (1998). Modélisation de la croissance du conduit vocal. Journées d'Études Linguistiques “*La Voyelle dans tous ses états*”. Nantes, 98-105.

- Boë, L.J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20, 27-38.
- Boë, L.J., Perrier, P., Guérin, B., & Schwartz, J.L. (1989). Maximal vowel space. *Proc. of Eurospeech 89*, 281-284.
- Boë, L.-J., Vallée, N., Badin, P., Schwartz, J.-L. & Abry, C. (2000). Tendances in Phonological Structures : The Influence of Substance on Form. Current Trends in Phonology and Phonetics II : Relationship between phonetics and phonology. *Les Cahiers de l'ICP, Bulletin de la Communication Parlée*, 5, 35-55.
- Bosma, J.F. (ed.) (1967). *Symposium on oral sensation and perception*. Springfield, Ill.: Charles C. Thomas.
- Bothorel, A., Simon, P., Wioland, F. & Zerling, J.-P. (1986). *Cinéradiographie des voyelles et des consonnes du français. Recueil de documents synchronisés pour quatre sujets: vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques*. Institut de Phonétique, Strasbourg, France.
- Brooks, R.A., Breazeal, C., Marjanovic, M., Scassellati, B. & Williamson M. (1999). The Cog Project: Building a Humanoid Robot. In C. Nehaniv (ed) *Computation for Metaphors, Analogy, and Agents. Lecture Notes in Artificial Intelligence 1562* (pp. 52–87). New York: Springer.
- Campbell, R., Dodd, B., & Burnham D. (eds.) (1998). *Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing*. Hove (UK): Psychology Press.
- Chistovich, L.A. (1976). *Physiology of Speech : Human Speech Perception*. Leningrad : Nauka (in Russian).
- Chistovich, L.A. (1980). Auditory processing of speech. *Language and Speech*, 23, 67-72.
- Davis, B., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Am. SLH Ass.* 38, 1199-1211.
- de Boer, B.G. (2000). Self-organisation in vowel systems. *Journal of Phonetics*, 441-465.
- Delgutte, B. (1984). Speech coding in the auditory nerve II: Processing schemes for vowel-like sounds. *J. Acoust. Soc. Am.*, 75, 879-886.
- Dodd B. , & Campbell, R. (eds.) (1987). *Hearing by eye : the psychology of lipreading*. Lawrence Erlbaum Associates, London.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton: The Hague.
- Gabioud, B. (1994). Articulatory Models in Speech Synthesis. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Recognition. Basic Concepts, State-of-the-Art and Future Challenges* (pp. 215-230). Chichester : John Willey.
- Goldstein, U.G. (1980). *An articulatory model for the vocal tract of the growing children*. Thesis of Doctor of Science, MIT, Cambridge, Massachusetts, USA.
- Guenther, F.H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102, 594-621.
- Guiard-Marigny, T. (1992) *Modélisation des lèvres*. DEA Signal Image Parole, INP, Grenoble, France.
- Hardcastle, W.J. (1976). *Physiology of speech production*. London: Academic Press.
- Jakobson,R.(1968). *Child language aphasia, and phonological universals*. The Hague: Mouton.
- Jordan, M. (1998). *Learning in Graphical Models*. MIT Press.

- Kaelbling, L.P., Littman, M.L. & Cassandra, A.R. (1996). Partially observable Markov decision processes or artificial intelligence; Reasoning with Uncertainty in Robotics. *Proc. International Workshop RUR'95* (pp.146-62). Springer-Verlag
- Kent, R.D., Martin, R.E., & Sufit, R.L. (1990). Oral sensation: a review and clinical prospective. In H. Winitz (ed.), *Human Communication and its Disorders* (pp. 135-191). Norwood, NJ: Ablex Publishing.
- Kent, R.D., & Miolo, G. (1995). Phonetic Abilities in the First Year of Life. In Fletcher, P. & MacWinney (Eds.) *The Handbook of Child Language*, Blackwell Publishers.
- Koopmans-Van Beinum, F., & Van Der Stelt, J. (1986). Early stages in the development of speech movements. In B. Lindblom, & Zetterstrom, R. (eds.) *Precursors of Early Speech* (pp.37-49). New York: Stockton Press.
- Kuhl, P., & Meltzoff, A.N. (1992). The bimodal perception of speech in infancy. *Science*, 218, 1138-1141.
- Kuhl, P., & Meltzoff, A.N. (1996). Infant vocalizations in response to speech : vocal imitation and developmental changes. *J. Acoust. Soc. Am.*, 100, 2425-2438.
- Laboissière, R. (1992). Préliminaires pour une robotique de la communication parlée : inversion et contrôle d'un modèle articulatoire du conduit vocal. Thèse de Docteur de l'INPG, Signal-Image-Parole, Grenoble, France.
- Landgren, S., & Olsson, K.A. (1982). Oral mechanoreceptors. In S. Grillner (ed.) *Speech Motor Control*. Oxford: Pergamon.
- Hoole, P. (1987). Bite-block speech in the absence of oral sensibility. *Proc. ICPhS, Tallinn*, 4, 16-19.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.
- Lauritzen, S., & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B*, 50, 157-224.
- Lebeltel, O., Bessière, P., Diard, J., & Mazer, E. (2003). Bayesian robot programming. *Autonomous Robot* (in press), 2003.
- Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839-862.
- Lindblom, B. (1986). Phonetic universals in vowel systems. In *Experimental Phonology* (J.J. Ohala and J.J. Jaeger, eds.), pp. 13-44. New-York: Academic Press.
- Lindblom, B. (1990). On the notion of possible speech sound. *Journal of Phonetics*, 18, 135-152.
- Lindblom, B. (1997). Systemic constraints and adaptive change in the formation of sound structure. In J. Hurford (ed.) *Evolution of Human Language*. Edinburgh: Edinburgh Univ. Press.
- Lindblom, B., Lubker, J., & McAllister, R. (1977). Compensatory articulation and the modeling of normal speech production behavior. In R. Carré et al. (eds.) *Articulatory modeling and phonetics* (pp. 147-161). GALF.
- Mackenzie Beck, J. (1997). Organic variation of the vocal apparatus. In W.J. Hardcastle & J. Laver (eds.) *The Handbook of Phonetic Sciences* (pp. 256-297). London: Blackwell Publishers.
- MacNeilage, P. F. (1998). The Frame/Content Theory of Evolution of Speech Production. *BBS* 21 (4), 499-511.
- MacNeilage, P.F., & Davis, B. (1990). Acquisition of Speech Production, Frames then Content. In M. Jeannerod (ed), *Attention and Performance, XIII: Motor Representation and Control* (pp.453-476).
- MacNeilage, P.F., Rootes, T.P., & Chase, R.A. (1967). Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. *Journal of Speech and Hearing Research*, 10, 449-467.

- Maeda, S. (1989) Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model. In W.J. Hardcastle & A. Marchal (eds.) *Speech Production and Modelling* (pp. 131-149). Kluwer : Academic Publishers.
- Matyear, C. L. (1997). *An acoustical study of vowels in babbling*. Doct. diss. University of Texas. Austin (unpublished).
- Matyear, C. L., MacNeilage, P. F., & Davis, B. L. (1998). Nasalization of vowels in nasal environments in babbling: evidence for frame dominance. *Phonetica*, 55, 1-17.
- McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
- Meltzoff, A. N. (2000). Newborn imitation. In Min, D. & Blater, A. al. (eds) *Infant development, the essential readings* (pp 165-181). Blackwell.
- Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., & Vallée, N. (2002). Auditory normalization of french vowels synthesized by an articulatory model simulating growth from birth to adulthood. *Journal of the Acoustical Society of America*, 111, 4, 1892-1905.
- Ménard, L., Schwartz, J.L., & Boë, L.J. (2004). The role of vocal tract morphology in speech development: Perceptual targets and sensori-motor maps for French synthesized vowels from birth to adulthood. *Journal of Language, Speech and Hearing Research*, 47, 1059-1080.
- Mills, A.E. (1987). The development of phonology in the blind child. In B. Dodd and R. Campbell (Eds.), *Hearing by eye: the psychology of lipreading* (pp. 145-161). London: Lawrence Erlbaum Associates.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, California, USA.
- Piquemal, M., Schwartz, J.L., Berthommier, F., Lallouache, T., & Escudier, P. (1996). Détection et localisation auditive d'explosions consonantiques dans des séquences VCV bruitées. *Actes des XXIemes Journées d'études sur la parole, SFA*, 143-146.
- Pols, L.C.W. (1975). Analysis and synthesis of speech using a broad-band spectral representation. In *Auditory Analysis and Perception of Speech* (G. Fant & M.A.A. Tatham, eds.), pp. 23-36. London: Academic.
- Recasens, D. (1991). An electropalatographic and acoustic study of consonant-to-vowel coarticulation. *Journal of Phonetics*, 19, 177-192.
- Savariaux, C., Perrier, P., & Orliaguet, J.P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production. *J. Acoust. Soc. Am.*, 98, 2428-2442.
- Schroeder, M.R., Atal, B.S., and Hall, J.L. (1979). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In B. Lindblom and S. Ohman (eds.) *Frontiers of Speech Communication Research* (pp. 217-229). London: Academic Press.
- 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.) *Phonetics, Phonology and Cognition* (pp. 255-280). Oxford: Oxford University Press.
- Schwartz, J.L., Arrouas, Y., Beautemps, D., & Escudier, P. (1992). Auditory analysis of speech gestures. In M.E.H. Schouten (ed.) *The Auditory Processing of Speech – From Sounds to Words* (pp. 239-252). Speech Research, 10, Berlin : Mouton de Gruyter.
- Schwartz, J.L., & Boë, L.J. (2000). Predicting palatal contacts from jaw and tongue commands: a new sensory model and its potential use in speech control. 5th Seminar on speech production : Models and data.
- Schwartz, J.L., Boë, L.J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25, 255-286.

- 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 (UK) : Psychology Press.
- Serkhane, J.E., & Schwartz, J.L. (2003). Simulating vocal imitation in infants, using a growth articulatory model and speech robotics. *Proc. ICPHS*, Barcelona, 2241-2245.
- Serkhane, J., Schwartz, J.L., Boë, L.J., Davis, B., Matyear, C. (2002). Motor specifications of a baby robot via the analysis of infants' vocalizations. *ICSLP'2002*, Denver, Colorado, 45-48.
- Steels, L. (1998). Synthesising the origins of language and meaning using co-evolution, self organisation and level formation. In J.R. Hurford, M. Studdert-Kennedy & C. Knight (eds.) *Approaches to the evolution of language* (pp. 384-404). Cambridge: Cambridge University Press.
- Thrun, S. (1998). Bayesian landmark learning for mobile robot localization. *Machine Learning*, 33, 41-76.
- Vilain, A., Abry, C., & Badin, P. (2000). Coproduction strategies in French VCVC: Confronting Ohman's model with adult and developmental articulatory data. *Proc.5th Seminar on Speech Production*, Munich, Germany, pp.81-84.
- Wood, S. (1979). A radiographic analysis of constriction locations for vowels. *J. Phon.*, 7, 25-43.
- Wu, Z.L., Schwartz, J.L., & Escudier, P. (1996). Physiologically plausible modules for the detection of articulatory-acoustic events. In B. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing, Vol.3 : Cochlear Nucleus* (pp. 479-495). U.K. : JAI Press.

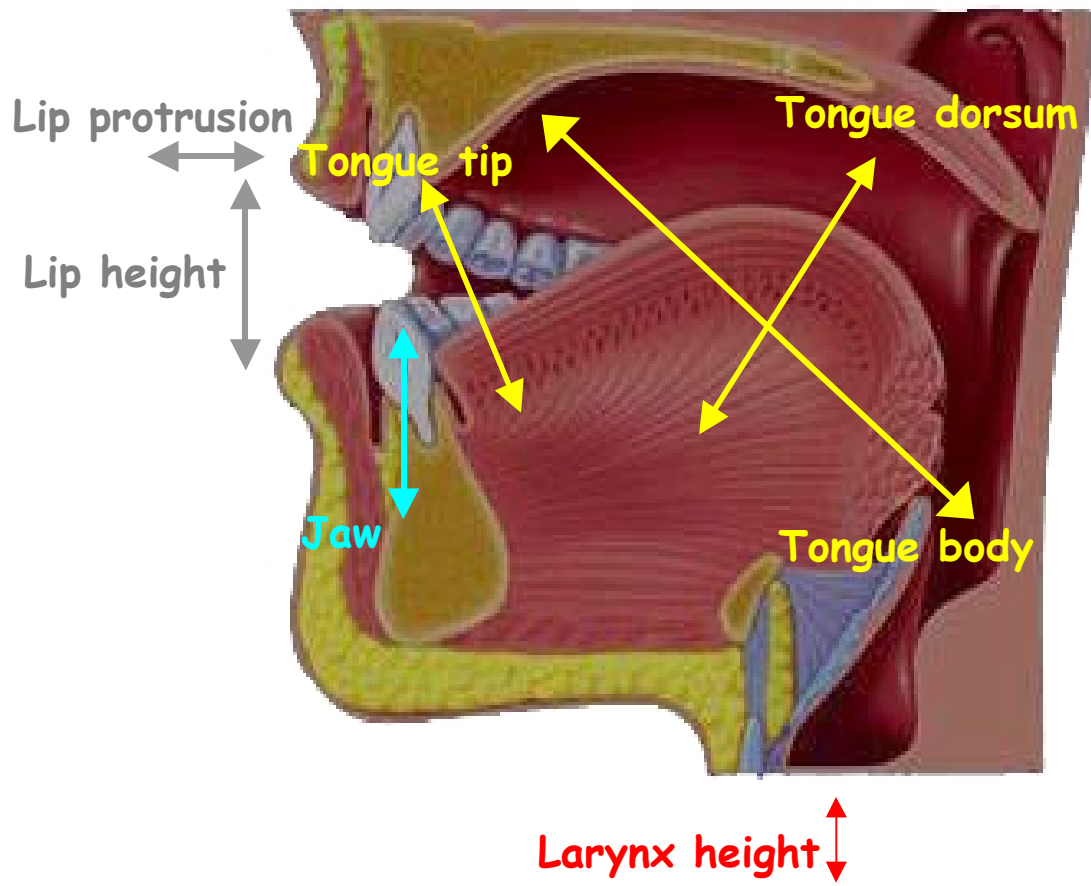


Figure 1: The articulatory model

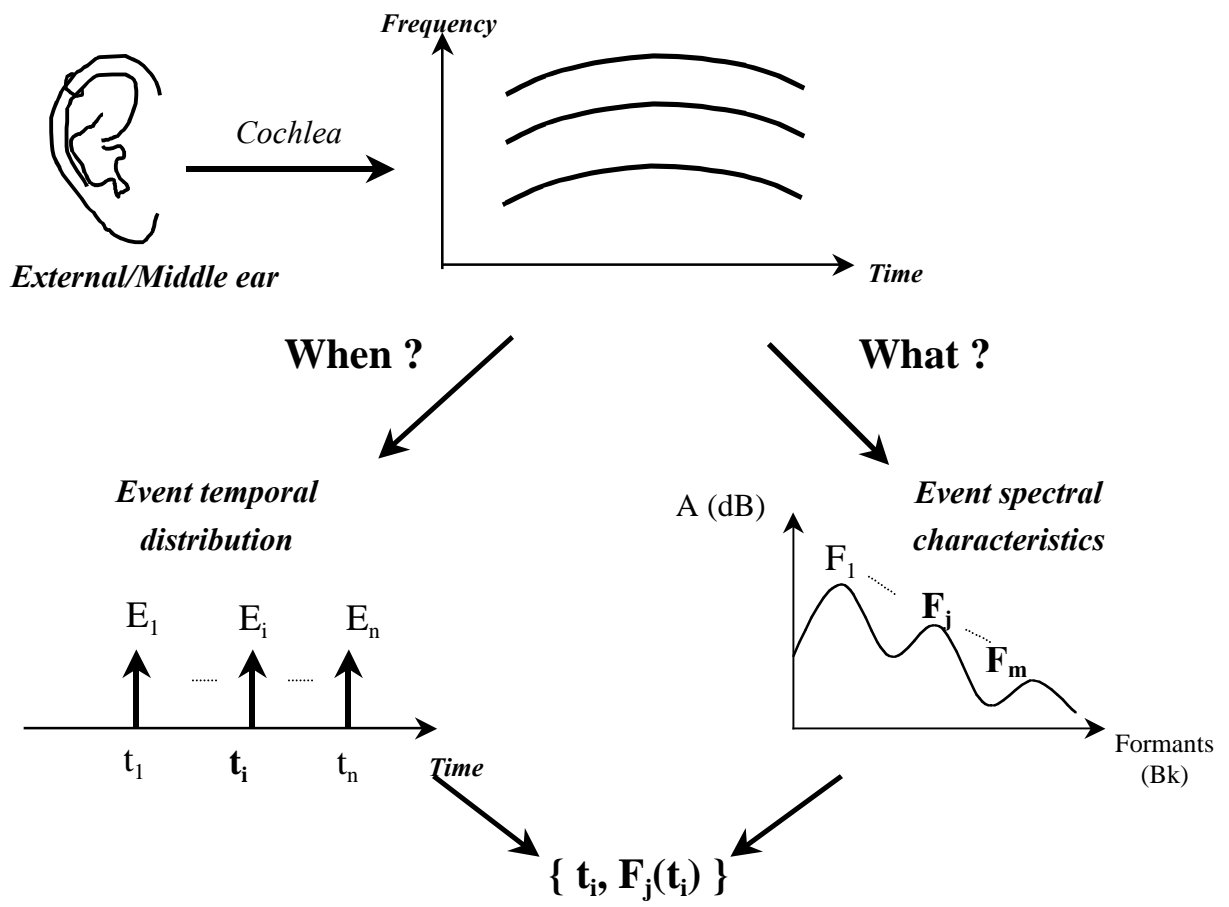


Figure 2: The auditory model.

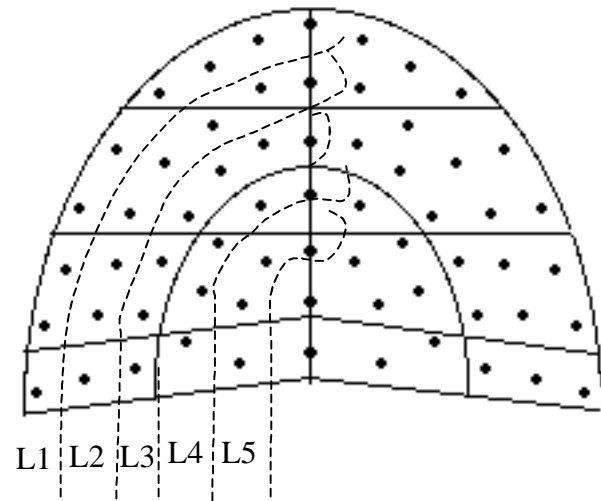


Figure 3: The palatal tactile sensor of the baby robot. See text.

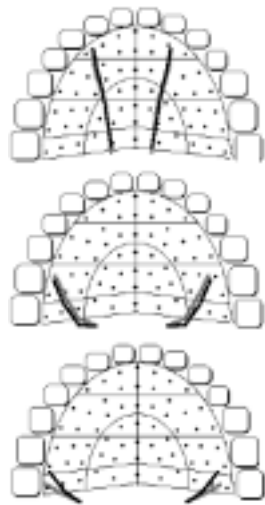


Figure 4a: Predicted (in black) and observed (in gray) palatal configurations for prototypical [i], [a], [o] (from top to bottom)

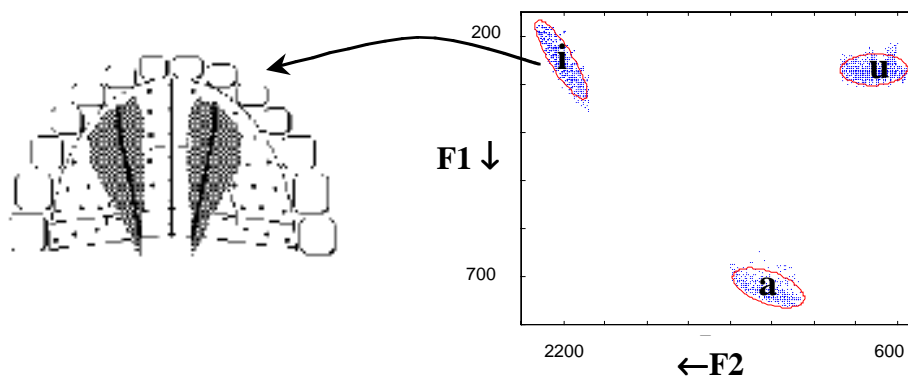


Figure 4b: Predicted palatal configurations (left) for a thousand articulatory configurations around [i] (formants on the right; the same was done for [a] and [u]).

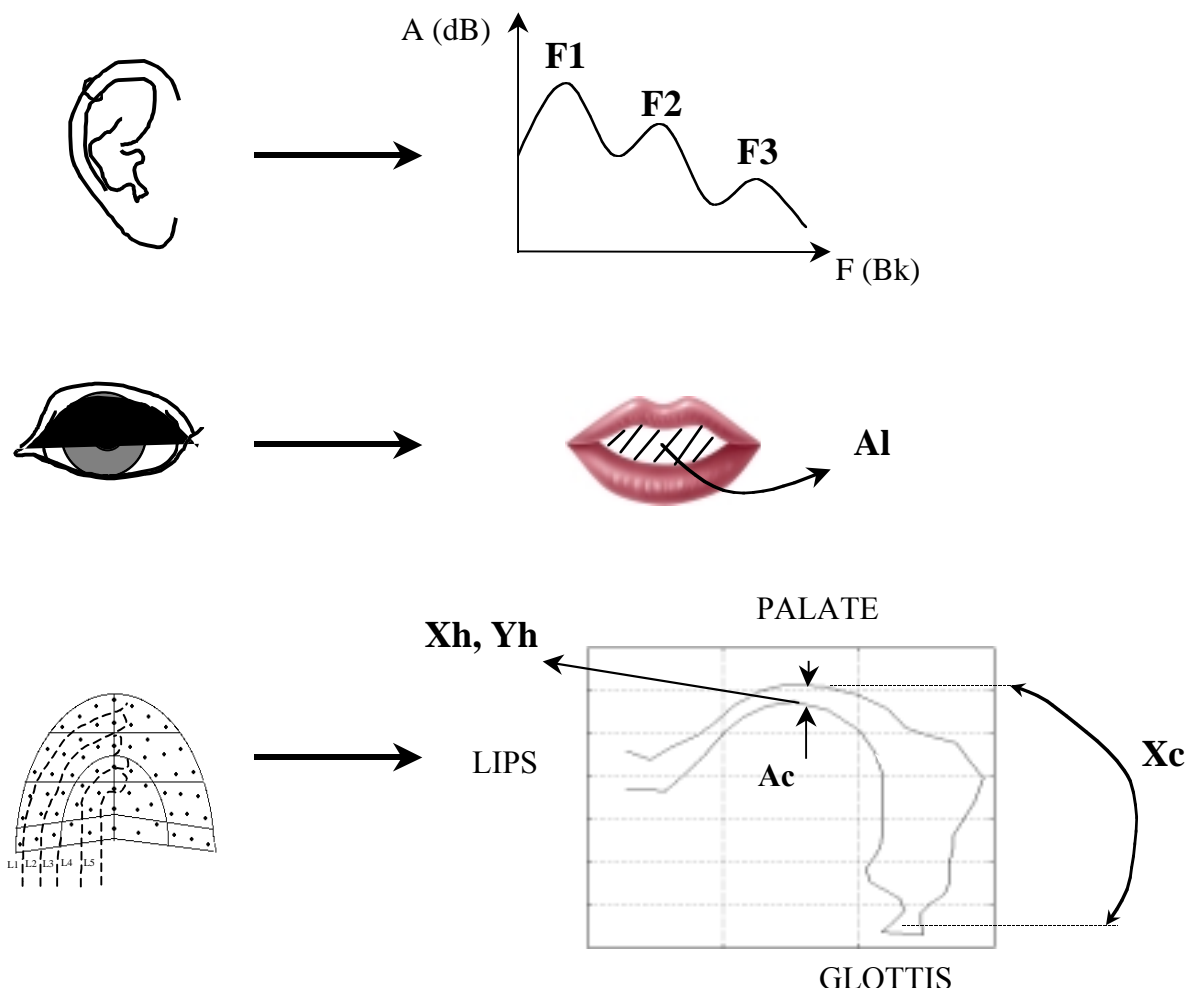


Figure 5: The simplified sensory models.

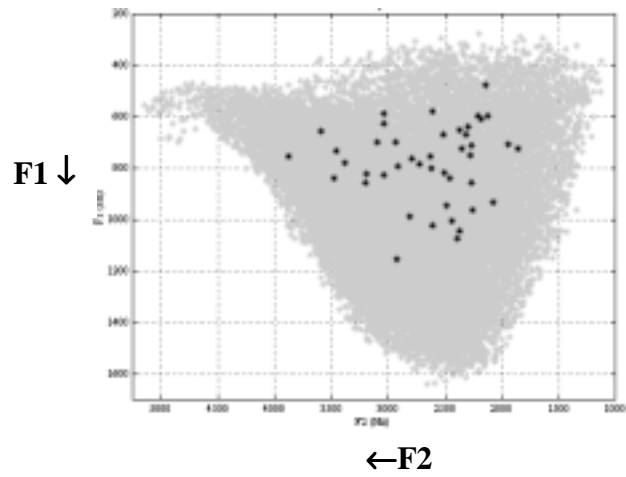


Figure 6: Acoustic framing of 4-month-olds' vocalizations (black dots). Gray dots correspond to the 4-month MVS. The F_i are expressed in Hertz.

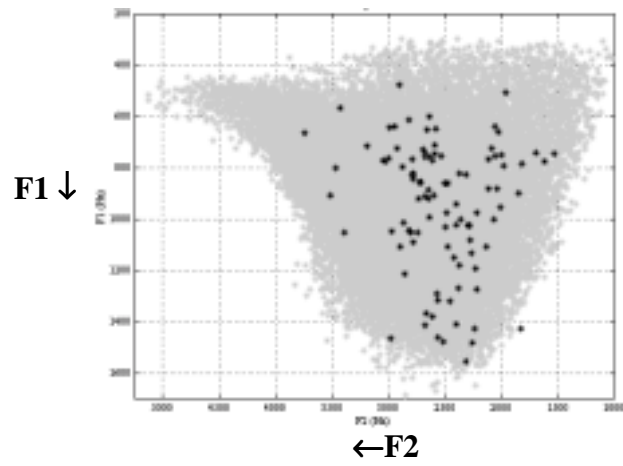


Figure 7: Acoustic framing of 7-month-olds' vocalizations (black dots). Gray dots correspond to the 7-month MVS. The F_i are expressed in Hertz.

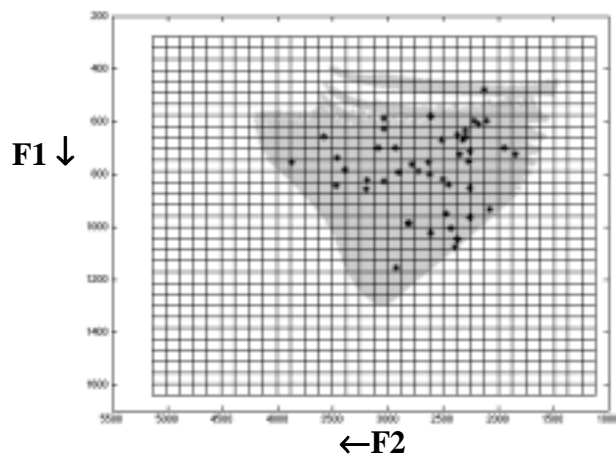


Figure 8: The articulatory framing of the 4-month-olds vocalizations by the selected three-parameters articulatory sub-model. The black dots correspond to the actual data, while the gray ones to the sub-model acoustic outputs. The grid shows the boxes employed to compute the probability criterion. The F_i are expressed in Hertz.

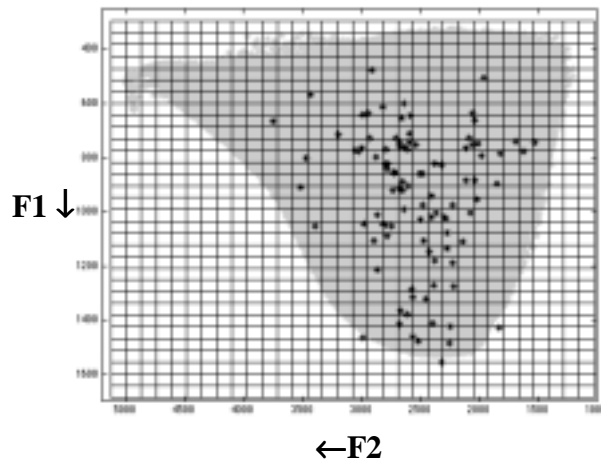


Figure 9: The articulatory framing of the 7-month-olds vocalizations by the selected four-parameters articulatory sub-model. Same caption as in figure 8.

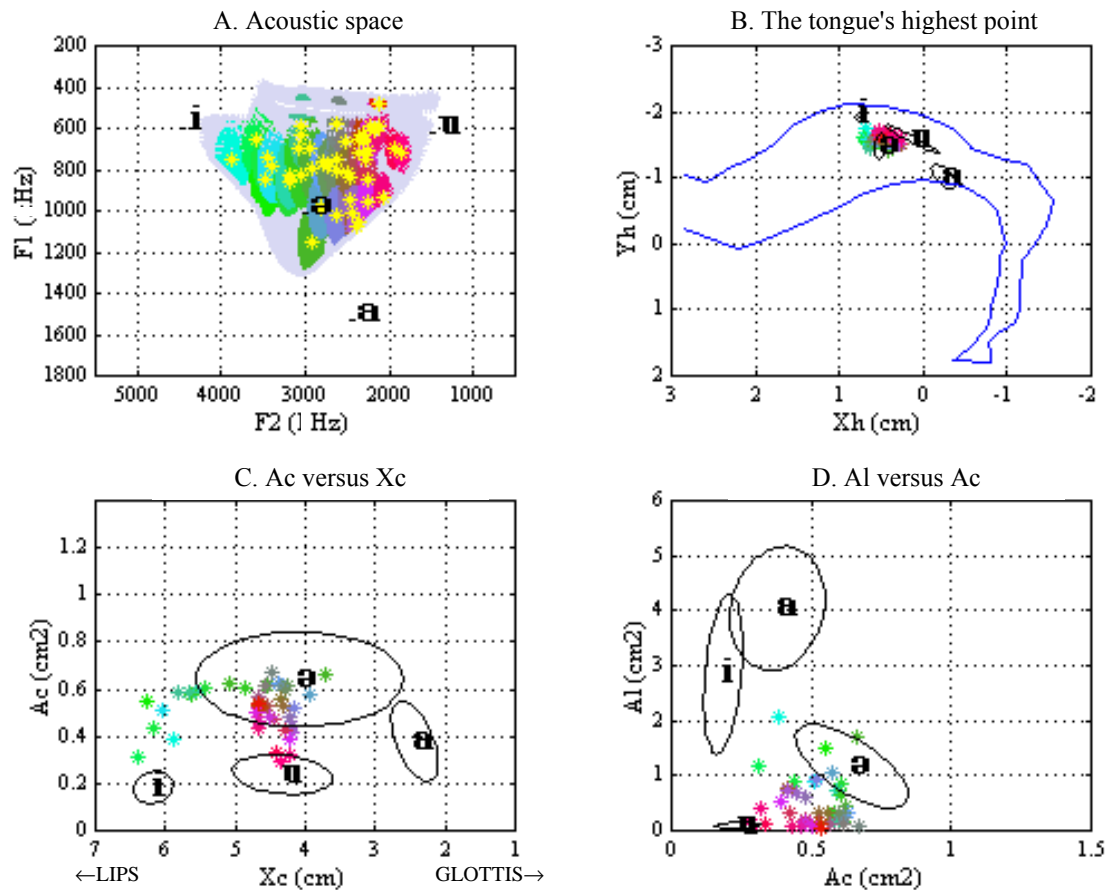


Figure 10: The geometric framing of the pre-babblers' vocalizations with the aged-matched winner (4 months old). In the acoustic domain (Panel A), the yellow stars correspond to the actual acoustic data, the mauve ones stand for the sub-model acoustic simulations from which, around each actual vocalization, a group of sounds was selected to perform the exhaustive inversion. Each group is color coded along the F2 axis (from cold to warm colors) so as to be able to track the means of the geometric characteristics of the resulting shape in the $\{X_h, Y_h\}$ space (Panel B) and the $\{X_c, A_c, A_1\}$ space (Panels C and D).

The points represented by the characters "i a u ə" correspond to "prototypic" formant values of the adult-like vowels (Panel A) that have been exhaustively inverted using the aged-matched VLAM. The dispersion ellipses of the geometric characteristics of their inferred average shapes are the only to be plotted for clarity's sake.

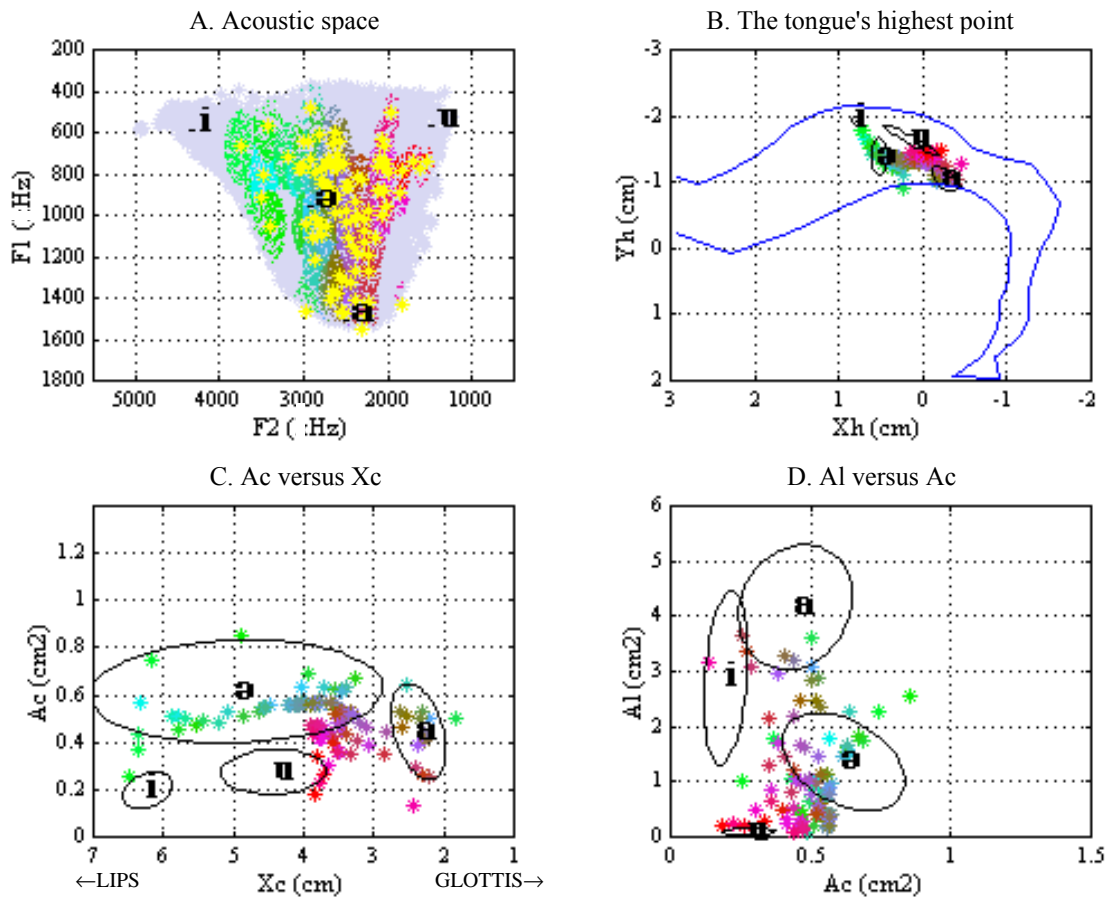


Figure 11: The geometric framing of the (7 months old) babblers' vocalizations with the aged-matched winner. Same caption as in Figure 10.

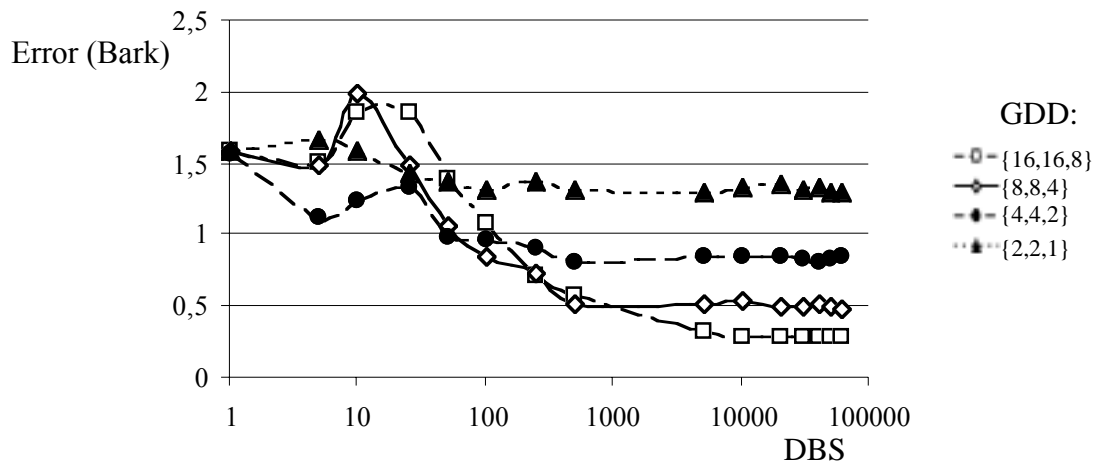


Figure 12: Assessing the GDD/DBS trade-off. Mean formant error at the output of the inversion process (in Bark) as a function of the DBS (GDD as parameter).

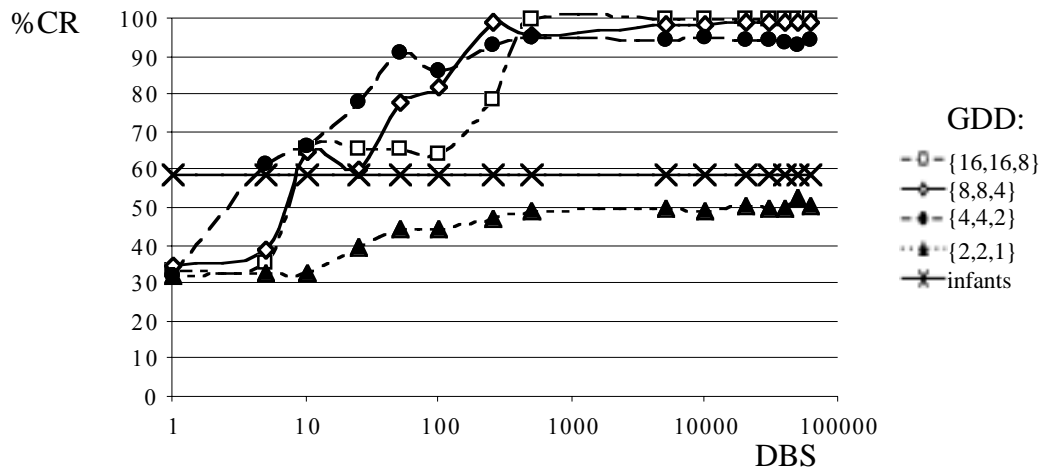


Figure 13: %CR for the *AV* inversion of the “internal” [i a u] vowels, as a function of the DBS (GDD as parameter). “Infants” stands for the score obtained by 12-20 weeks infants in the study by Kuhl & Meltzoff (1996).

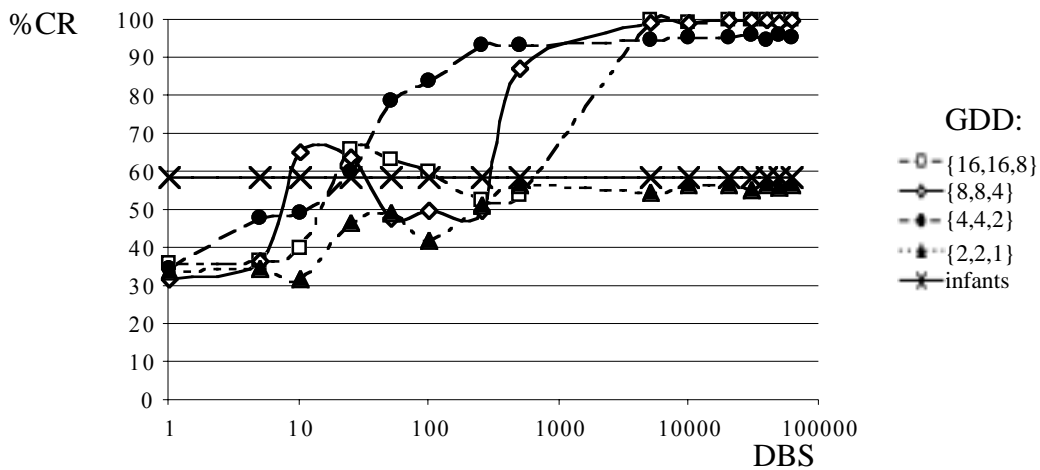


Figure 14: %CR for the *AV* inversion of the “external” [i a u] vowels, as a function of the DBS (GDD as parameter). “Infants” stands for the score obtained by 12-20 weeks infants in the study by Kuhl & Meltzoff (1996).

Table 1: The confusion matrix of early vocal imitation reported in Kuhl and Meltzoff (1996). Each cell provides the number of “i-like”, “u-like” and “a-like” vocalizations (see text) for each of the three possible adult targets [i a u]. Among the 72 infants in the experiment, only 45 produced vowel-like utterances. Altogether the 45 infants uttered 224 vowel-like vocalizations along the experiment.

	i	a	u	Total
i-like	22	11	4	37
a-like	25	66	14	105
u-like	20	18	44	82
Total	67	95	62	224