

Infants' vocalizations analyzed with an articulatory model: A preliminary report

Running title: Infants' vocalizations analyzed with an articulatory model

Key-words: speech acquisition, articulatory exploration, modeling

J. E. Serkhane⁽¹⁾, J.L. Schwartz⁽¹⁾, L.J. Boë⁽¹⁾, B. L. Davis⁽²⁾, C. L. Matyear⁽²⁾

(1) Institut de la Communication Parlée (ICP), CNRS-INPG-Université Stendhal

46, avenue Félix Viallet. 38031 Grenoble Cedex 1, France

jihene, schwartz, boe@icp.inpg.fr

(2) Department of Communication Sciences & Disorders, The University of Texas at

Austin, Austin, TX 78712 USA

babs, matyear@mail.utexas.edu

Abstract

Articulatory exploration enables the infant to discover abilities of the vocal tract and learn relationships between movements and percepts. However, neither direct measurements nor transcription methods have access to tongue configurations in pre-linguistic infant vocalizations. A statistical articulatory-acoustic model integrating the non-linear growth of the human vocal tract was used to describe infant behavior before and at the beginning of canonical babbling. Analyses were developed to assess from a set of (F1, F2) formant frequencies reported at four and seven months in two separate corpora the most likely articulatory degrees of freedom of the model. Results indicate that exploration in the four-month corpus is centered around a neutral configuration. It involves at least three articulatory parameters, including at least one for the tongue. The jaw seems to play a minor role in this exploration. In contrast, in the seven-month corpus, the exploration range increases: in this case the jaw plays a dominant role, leading to a large exploitation of the open-close contrast and associated F1 diversification in formant space. The simulation of co-occurrences between closants and vocants from the seven-month corpus in the framework of the Frame-Content theory provides a portrait largely consistent with previously reported experimental data. Locus scatter-plots were also simulated and compared to available data on development of coarticulation in CV syllables. This kind of analysis could be applied to corpora of infants' vocalizations at various ages to understand the development of speech production in relation to the growth of the human vocal tract.

1. Introduction

The development of speech production skill entails progressive mastery of a complex sensori-motor system. For this aim, two processes seem necessary: (1) an *exploration process* by which the infants should discover the abilities of their vocal tracts, and learn the correspondence between movements and sounds; and (2) a *tuning process* by which they gain control of the articulatory system, in order to produce the movements and sounds of their target language(s). Both processes are precisely defined in computational motor control models (e.g. Jordan, 1990; Jordan & Rumelhart, 1991). The *exploration process* provides sensori-motor inputs enabling the infant to learn a “forward model” that represents the set of associations between actions and percepts, while the *tuning process* enables the infant to learn an “inverse model” which specifies the articulatory commands required to produce a given acoustic output. These mechanisms are generally modeled as *sequential* and *exhaustive*. That is to say, motor control models of speech development generally begin with a first stage of exhaustive exploration of the vocal tract, resulting in a forward model specifying all possible correspondences between movements and sounds (e.g. Laboissière, Schwartz & Bailly, 1991; Markey, 1994; Guenther, 1995; Bailly, 1997). The inverse model has then to solve a complex many-to-one inversion problem, using various heuristics.

In that framework, infants were proposed to start by vocalizing all possible speech sounds in the world's languages (in agreement with Jakobson, 1968, p.21). However, studies by a number of researchers have shown that this is not the case in the first stages of infant vocalization (e.g. Oller, 2000; Stark, 1980). Nor is it the case in canonical babbling, the stage considered to be the first critical step into speech development (e.g.

Davis & MacNeilage, 1995; Kent & Miolo, 1995; Locke, 1983; Oller, 2000). Indeed, whatever their ambient language, babblers produce a certain subset of what can be performed with their vocal tract (e. g. MacNeilage & Davis, 2001). In this context, it is quite important to gain further insights into the vocal tract articulatory dimensions that infants actually exploit, since it should suggest some developmental constraints that may shape speech acquisition. The present paper is focused on the early exploration period extending until the initiation of canonical babbling.

1.1. Basic steps in infants' vocal tract exploration

At birth, infants imitate gestures from vision: tongue and lip protrusion, and mandible depression (Meltzoff, 2000). These movements, which are basic in adult speech, are available from the beginning of life, even though they are clearly not linked with speech production at birth. The imitation behavior itself is part of a general ability to reproduce somebody else's actions in a broad sense (Meltzoff & Moore, 1997). Infants begin to vocalize very shortly after birth as well. The formant range of vocalizations slowly increases across the first year (e.g. Buhr, 1980; Kent & Murray, 1982; Kuhl & Meltzoff, 1996). Moreover, four-month-olds tend to direct their productions towards vowel sounds they hear, illustrating early steps toward vocal imitation (Kuhl & Meltzoff, 1982, 1996).

At about seven months, infants begin canonical babbling (Davis & MacNeilage, 1995; Koopmans-Van Beinum & Van Der Stelt, 1986; Oller, 2000): their mandibles open and close rhythmically, while their vocal folds vibrate. This phenomenon marks the first appearance of vocalizations with speech-like timing. According to the Frame-Content theory (MacNeilage & Davis, 1990; MacNeilage, 1998), rhythmic mandibular cycles serve as a "frame" for the future syllable, by producing an alternation between resonant and non-resonant acoustic

outputs producing the two basic syllabic components: vowel- and consonant-like percepts, referred to as *vocants* and *closants*, respectively (Martin, 1981). In adult speech, the “content” corresponds to controlled segments, generated by movements of the lower jaw, the lips, the tongue and the velum that are independently activated during verbal sequences. In contrast, at the onset of canonical babbling, the only active articulator is the lower jaw since the movements of the articulators it carries, that is, the lower lip and the tongue, as well as the velum and the upper lip, do not seem independent of the rhythmic jaw cycles (see Munhall & Jones, 1998, for the lips; Sussman, Duder, Dalston, & Cacciatore, 1999, for the tongue; and Matyear, MacNeilage & Davis, 1998, for the velum). Thus, in canonical babbling vocalizations, the only significant articulatory difference between contiguous vocalic and consonantal aspects of the babbled sequences is based on the up versus down movement of the lower jaw. A babbling utterance can hence be viewed as a shape (*presetting*) of the vocal tract on which the mandibular oscillation is superimposed. At a phonetic level, this phenomenon would be translated into shared places of articulation between contiguous vocants and closants. The predicted outcome, according to the Frame-Content theory, would be the preponderance of the following *co-occurrence* patterns: front vocalic sounds associated with coronal closants (e.g. /de/), central vocalic sounds with labial closants (e.g. /ba/), and backed vocalic sounds with palato-velar closants (e.g. /gu/). At an articulatory-acoustic level, this rhythmic movement cycle would lead to specific coarticulation patterns, that is predictable relationships between the characteristics of a closant and the following vocant inside a given syllable. This implies, therefore, that babblers' repertoire is *not infinite*, since it would favor certain associations of speech-like sound qualities (MacNeilage & Davis, 2001; see also Davis, MacNeilage & Matyear, 2002, for the babbling-to-speech transition). Later on, development from babbling onset to appearance of first words and until a completely mature control of the vocal tract involves a number of steps that can extend over many years. These

steps include the control of sequences of mandibular oscillation (Green, Moore & Reilly, 2002), of the movements of the articulators carried by this cycle independently one of each other (Munhall & Jones, 1998; Green, Moore, Higashikawa & Steeve, 2000), and of the full shape of the vocal tract (Sussman *et al.*, 1999) to master sounds and sequential patterns of the ambient language (Nittrouer, 1993; Abry, Cathiard, Vilain, Laboissière, & Schwartz, to appear).

1.2. Vocal tract growth

Across the period of progressive development of motor control abilities, major anatomic modifications also occur. Their role is sometimes difficult to disentangle from cognitive factors. More precisely, while the geometry of vocal tract modifications is now reasonably well-known from a number of experimental studies, the acoustic consequences of these modifications have not been systematically considered. Cineradiographic data (Goldstein, 1980) as well as MRI (Fitch & Giedd, 1999; Callan, Kent, Guenther & Vorperian, 2000; Vorperian, 2000; Vorperian *et al.*, 2005) have shown that the adult's vocal apparatus is a complex remodeling of the infant's. At birth, the overall vocal tract length, determined from the larynx to the lips, is about 8 cm, whereas the adult male vocal tract is about 17 cm long (Goldstein, 1980). Furthermore, vocal tract growth is not uniform. Following Goldstein (1980), the ratio of the pharynx length versus the length of the oral cavity varies from 0.5 to 1.1 from birth to adulthood for a male. MRI data (Fitch & Giedd, 1999) confirm this general tendency.

These studies provide a relatively complete description of the anatomical modifications of the vocal tract from birth to adulthood, from which it is possible to estimate the acoustic configurations that can be generated by the vocal instrument all along this growth process. For

example, Ménard, Schwartz & Boë (2004) used an articulatory model of the vocal tract based on such measurements and demonstrated that anatomy does not prevent even the youngest speaker from producing all possible vocalic sounds *if motor control were mature*. Articulatory modeling in this respect may help to disentangle anatomic from cognitive aspects in speech development. The present study used the Ménard et al. (2004) computational model of the vocal tract.

1.3. Articulatory characterization of early vocalizations

Even though many studies have focused on description of infant vocalizations, little is known of actual early articulatory exploration. Indeed, ethical reasons prevent human infants from being exposed to X-ray or any other invasive methods such as non-surface electromyography (EMG) with hooked-wired electrodes or electro-magnetography with coils placed on the tongue. Lingual movements are difficult to record via captor-laden pacifiers (Maeda, personal communication). There is no report on direct measurements of articulatory activities before 2 years of age, except for movements of the lips and/or the lower jaw, using video recording of infrared emitting diodes (Munhall & Jones, 1998) or reflective markers (Green *et al.*, 2000; Green *et al.*, 2002), and for activities of lower jaw muscles, using surface EMG electrodes (Moore & Ruark, 1996). Importantly, current methods allow no direct measurement of the tongue configurations associated with infant vocalizations. Researchers exploring early motor abilities are restricted to indirect investigation into articulatory activity through transcription and acoustic methods (Kent & Miolo, 1995).

Transcription studies employ a grid of symbols, most often the International Phonetic Alphabet (IPA), to code utterances perceived and labeled by adult transcribers. However, the problem with phonetic transcriptions is that "use of consonant and vowel symbols implies

independent *control* [*sic*] of segments," as has been pointed out by Davis and MacNeilage (in press) (see also Kent & Murray, 1982), whereas vocalizations are rather the by-products of an immature motor behavior that has no linguistic interpretation *per se* or even no (self-evident) linguistic intention. Further, perception of phonemic categories represented by the IPA can be biased toward the listener's linguistic background. For instance, native English speakers rarely categorize nasalized vocalic sounds, which are frequent in infant productions (Beddor & Strange, 1982; Matyear, 1997), as there is no vowel contrast based on this characteristic in English. An attempt to free transcription from adult phonemic systems can be found in the sensorimotor classification proposed by Koopman-van Beinum and Van der Stelt (1998) in which vocalizations are sorted according to phonation and articulatory types, with an encoding grid of the vocalic sounds based on broad categories defined along the front-back and the high-low dimensions of the oral cavity. Nevertheless, this system still lacks precision in tracking articulatory activities related to pre-linguistic vocalizations.

It should also be possible (Kent & Murray, 1982) to exploit acoustic analysis in combination with transcription, in order to attempt to estimate articulatory characteristics of vocalizations, capitalizing on the rough articulatory-acoustic relations derived from investigations into adult speech (e.g. Buhr, 1980; Lieberman, 1980; Matyear *et al.*, 1998). However, estimating articulation from sound is a very tough problem, the more so considering that vocal tract growth modifies articulatory-acoustic relationships (Ménard *et al.*, 2004). This raises the problem of normalization, which researchers try to overcome by looking for a way to convert acoustic measurements from the infant vocal tract to the adult's, to make early vocalizations relate validly to adult vowel spaces (see a recent review and set of proposals in Ménard, Schwartz, Boë, Kandel & Vallée, 2002).

1.4. Articulatory modeling as a tool for re-analyzing infants' vocalizations

In this context, articulatory modeling incorporating vocal tract growth could provide a powerful tool for assessing articulatory exploration from acoustic data based on infant vocalizations. This is the core objective of the present paper. It is proposed that it is possible to infer a number of articulatory trends from a set of formant patterns by analyses based on articulatory modeling. For this aim, infant vocalizations were compared to the acoustic and articulatory capacities of an infant vocal tract model.

This work is both preliminary and exploratory, since it involves a new methodology for vocalization analysis, which has the inherent limitations of any modeling approach. The focus is on infant vocalization patterns produced during two developmental periods: before and at the beginning of canonical babbling, corresponding to actual vocalizations produced by 4- and 7-month-olds. Indeed, although most developmental studies deal with canonical babbling productions, a working hypothesis is that infants are likely to gain some motor experience from the pre-babbling period as well. Therefore, the main aim was to assess the extent of vocal exploration in infants before babbling and at babbling onset.

Three basic questions were asked in the present study. First, can an infant vocal tract model account for the range of infant vocalizations reported in the literature? Particularly, does it have geometrical characteristics compatible with the formants of infant utterances? Second, is it possible to quantify the range of articulatory exploration compatible with the displayed range of formant exploration, to estimate the subset of commands that should be exploited in the articulatory model to reproduce as adequately as possible the available corpora of formant data? Finally, are coarticulation patterns (i.e. relationships between articulatory and acoustic characteristics of consecutive consonants and vowels) displayed by the model exploration at seven months, consistent with the Frame-Content theory of speech acquisition?

Section 2 provides all components of the method used in this work, including description of the articulatory model incorporating vocal tract growth, selection of phonetic material gathered in studies published in the literature on speech development, and description of analysis tools aimed at comparing infant vocalizations to the productions of the model. Section 3 provides the results of these comparisons, which are discussed in Section 4, in relation to both the interests and limitations of the proposed method, and the coherence of the inferred results with other approaches and results in the literature.

2. Method

2.1. The Variable Linear Articulatory Model

The *Variable Linear Articulatory Model* (hereafter VLAM) (Boë, 1999) is a version of the *Speech Maps Interactive Plant* model (SMIP) (Boë, Gabioud & Perrier, 1995a) that integrates the non-uniform growth of the vocal tract. The SMIP mainly stemmed from a principal component analysis (PCA) of data describing mid-sagittal cineradiographic sections of a speaking adult's oral tract (Maeda, 1990). This statistical analysis led to 5 relevant factors, which explained 92% of the observed variability in the data. These factors defined the following SMIP commands or “articulatory degrees of freedom”: jaw vertical movement (hereafter Jaw, or J), tongue protrusion-retraction (Tongue Body, or TB), tongue arching-flattening (Tongue Dorsum, or TD), tip of the tongue vertical movement (Tongue Tip, or TT) and larynx height (Larynx, or Lx). These factors may be related to concrete muscular actions (Maeda & Honda, 1994; Payan & Perrier, 1997). A model of lip shape (Abry & Boë, 1986) was adapted to Maeda's model, thereby adding two degrees of freedom, which are the intralabial height (Lip Height, or LH) and the lip protrusion (Lip Protrusion, or LP). The

relative weights of the 7 articulatory variables were normalized through the database that yielded the SMIP: their ranges of variation are expressed in terms of \pm number of standard deviation(s) (*std*) centered on 0. These parameters serve as model inputs to synthesize a two-dimensional mid-sagittal section and the corresponding area function (three-dimensional equivalent), from which it is possible to work out the transfer function, formant frequencies (resonance maxima) and speech signal (Badin & Fant, 1984).

The VLAM is extensively described elsewhere (Boë, 1999; Ménard *et al.*, 2004). In this model, based on the degrees of freedom of the SMIP, the growth process is introduced through two scaling factors that size the length and the width of both the anterior and the pharyngeal parts of the generated adult mid-sagittal section in the SMIP, interpolating the zone in-between. The variations of both scaling factors from birth to 21 years old follow a model derived from the cranio-facial measurements gathered by Goldstein (1980): hence, the scaling factor for the pharyngeal part is much lower in infants than the factor for the anterior part, resulting in a ratio of the pharynx length versus oral cavity length around 0.5, similar to the measurements. Likewise, the value of the fundamental frequency (f_0) varies as a function of the age. It was fitted to Beck's data (Beck, 1996). Thus, the age of the virtual vocal tract sets the sizes of the front and the back cavities as well as the fundamental frequency exciting the vocal resonator. The VLAM has been compared to real data (Ménard *et al.*, 2004), and it generates realistic articulatory and acoustic vowel configurations. Overall vocal tract lengths and cavity lengths are in line with MRI measurements from birth to 6 years of age (Vorperian, 2000), and acoustic values obtained for prototypical vowels are in the range of the mean values \pm 1 standard deviation reported for vowels from 3 years old to adulthood (Lee, Potamianos, & Narayanan, 1999; Hillenbrand, Getty, Clark, & Wheeler, 1995). This model is thus well suited for studying infant vocalizations.

2.2. *Phonetic material*

Two corpora were selected, combining formant values and phonetic transcriptions, and as representative as possible of the 4- and 7-month time points, before and at the onset of canonical babbling. They came from published data, collected for purposes independent of the present study, and hence providing an interesting sample for evaluating the ability of the proposed tools to assess articulatory exploration in the first stages of speech development.

The 4-month-old data is from Kuhl and Meltzoff (1996). The original study dealt with early vocal imitation in 20-week-old infants. The design of this experiment might have encouraged infants to vocalize rather than to produce other comfort sounds, and to explore, as far as they could, their articulatory-acoustic repertoire since the targets corresponded to the extreme articulatory configurations of the vocal tract. This is the primary reason why it was selected.

The second set of data, gathered by Matyear and colleagues (Matyear, 1997; Matyear *et al.*, 1998), aimed at testing for the Frame-Content theory (MacNeilage & Davis, 1990) related to the potential contribution of the soft palate to changes within CVC forms during canonical babbling. This study dealt with spontaneous vocalizations rather than stimulus-driven imitation. All forms analyzed by Matyear and colleagues were rhythmic canonical syllables based on perceptual agreement between two listeners. No pre-canonical vocalizations (see Oller, 2000, for definitions) were included for analysis in the corpus. We shall discuss in Section 4 potential consequences of differences in experimental focus between the two corpora.

2.2.1. *4-month-old corpus*

This corpus is made of the whole set of the vocalizations produced by 24 20-week-olds, born in the Seattle (Washington) area, in the Kuhl & Meltzoff (1996) study. The study, carried out in a laboratory setting, was focused on early vocal imitation of the adult vowels /i/, /a/ and /u/, displayed as audio-visual face-voice stimuli to infants, whose subsequent vowel-like productions were, whenever possible, phonetically and acoustically described. All utterances produced by the infants during the experiment were included, provided that they could be selected as “vowel-like” sounds on the basis of a set of criteria described in the original paper. The formant values (F1, F2) of the vocalic sounds produced by the infants were available (figure 3 in Kuhl & Meltzoff, 1996). There were 45 vocalizations available for analysis¹. The system of transcription employed English vowel symbols but the transcribed items were merged into three categories: the /a/-like, including /a æ ʌ/, the /i/-like, with /i ɪ ε/, and the /u/-like for /u u/. Acoustic signals were digitalized at a rate of 20 kHz, and formant values were assessed through the corroboration of a narrowband spectrogram (114 Hz), a Fast Fourier Transform (256 points) and Linear Predictive Coding (LPC) response (10 ms frame length, filter order 12). The frequencies of the first two formants resulted from the mean of each measurement through five temporal locations across each vocalization (onset, 1/4, 1/2, 3/4, and offset), since no significant differences were found between these successive points.

2.2.2. 7-month-old corpus

The data gathered by Matyear and colleagues consisted of spontaneous vocalizations produced by three infants reared in a monolingual American-English speaking environment. Data were from a larger study of infant vocalizations (Davis & MacNeilage, 1995). Vocalizations were recorded in each infant's home environment from the onset of canonical

babbling to the onset of the single word period. They were transcribed using IPA system. The corpus selected by the authors in the original work included only perceptually rhythmic CVC syllable-like canonical tokens, with reduplicated closants /b d m n g ŋ/ surrounding each of the vocants /i ɪ e ε æ ə ʌ a u ʊ o ə/. Vocalic portions were acoustically analyzed whenever two transcribers agreed on their phonetic transcription. Each token was digitalized at a rate of 16 kHz. Formant frequencies were estimated using a spectrogram (100-Hz bandwidth) permitting selection of a 200 ms portion of the vocant steady state from which a Linear Predictive Coding (LPC) at 50 ms intervals generated a spectrum averaged across the steady-state area.

The corpus analyzed in the present study included all canonical syllable-like vocalizations produced at seven months by the two participants who began babbling at seven months (the third infant began later). There were ninety-eight tokens in the corpus, corresponding to 7 vocant categories: /i ɪ e ε æ ə a o/.

2.3. Acoustic framing

This first analysis tool was developed to normalize infants' acoustic productions while taking into account vocal tract growth. This was accomplished by finding their position in an acoustic system of reference equivalent to what has been referred to as the vowel space in adults. The VLAM was used as a model of oral vocalic production. For a specified age, all the articulatory configurations of the VLAM produced sounds that fell within an acoustic space called the *Maximal Vowel Space* (hereafter MVS) (Boë, Perrier, Guérin & Schwartz, 1989). The MVS corresponds to what an age-matched infant would be able to utter if she used the complete set of the VLAM articulatory commands. MVS therefore stands for all oral vocalic

speech sounds *acoustically* achievable at a given age considering vocal tract shape, plotted on a multi-formant (Fi) map. The (F1, F2) plane depicts the age-matched vocalic triangle that phoneticians use to investigate adult speech: at its corners there are the /i a u/ vowels.

The acoustic framing consisted of superimposing the set of actual vocalizations produced by the 4- and 7-month-olds on the MVS at 4 and 7 months, respectively. To generate both MVS, the input grid of command parameters was randomly drawn from a uniform distribution between -3 and $+3$ *std* for the 7 parameters and applied to the VLAM at the corresponding ages. The minimal intra-oral and inter-lips areas were constrained to 0.1 cm^2 and 0.01 cm^2 respectively (these thresholds were used for the simulation of newborn's vowel production in Ménard *et al.* 2002). The computed formant values were stored, with a total of 30, 000 simulations for each of the two tested ages.

2.4. *Articulatory framing*

The next step consisted of attempting to determine, on the articulatory model, the range of articulatory exploration corresponding to the observed acoustic exploration in each infant corpus. The precise estimation of an articulatory configuration from an acoustic output, referred to as “acoustic-to-articulatory inversion,” is an insolvable problem, because of the many-to-one relation between the articulatory and the acoustic domains in speech production (e.g., Atal, Chang, Mathews & Tukey, 1979; Boë, Perrier & Bailly, 1992). However, it is possible to propose a global articulatory characterization of an acoustic corpus of infant vocalizations within the VLAM. The principle is to take into account the result of acoustic framing, generally showing that the extent of actual vocalizations does not cover the full range of possible sounds for the VLAM at the corresponding age. Then, articulatory framing

consists of finding a reduced articulatory space that produces a reduced acoustic space as close as possible to the actual data space in terms of formant repartition.

From a probabilistic viewpoint, the articulatory framing procedure looked for the likeliest minimal set of articulatory parameters (or "sub-model") on the basis of the distribution of its acoustic realizations (theoretical distribution), given the distribution of the age-matched infants' vocalizations (actual distribution) in the plane of the first two formants. In other words, the purpose was to find the sub-model that maximized the probability $P(M_i/D)$, where M_i denotes the i^{th} sub-model, characterized by the distribution of the acoustic outputs it generated, while D stands for the distribution of the actual data in the formant space. This technique required (1) the definition of the articulatory sub-models competing one with each other, (2) the generation of their acoustic output spaces as well as the computation of the corresponding statistical distribution in the (F1, F2) plane, and (3) the selection of the best sub-model given the statistical distribution of the age-matched infants' vocalizations.

Four articulatory parameters, namely LH, J, TB and TD, were selected to complete the procedure. These parameters are the most relevant with regard to speech acoustics for related reasons. Firstly, J, TB and TD account for most of the observed variance in the tongue profiles of the statistical analysis the VLAM originated from, as they explain 81% of total inertia. Furthermore, they provide the major parameters to specify the position, X_c , and the area, A_c , of the tongue-palate constriction in the vocal tract, whereas J along with LH drive the inter-lip area, A_l , and $\{X_c, A_c, A_l\}$ are the main correlates of the formant frequencies (Fant, 1960; Boë, Gabioud, Perrier, Schwartz & Vallée, 1995b). The three parameters not considered in the articulatory framing stage, LP, TT and Lx, play only a minor role in speech acoustics: they were neglected here to make the analysis tractable.

For each of those four parameters, the range of variation varied from zero to the adult range defined by the interval $[-3, +3]$ *std*. The step size for range variations was chosen as one

unit standard deviation and always included the 0 value considered as the neutral position in the VLAM. Hence ranges were of the form $[-m, +n]$ *std*, m and n being positive or null integers lower than or equal to 3. The values of the remainder, that is, LP, TT and Lx, were set to 0. This provides 65,536 sub-models altogether (that is 16^4 : for each command, 4 values for m and 4 for n , hence 16 tested ranges). Then, for a given corpus of infants' vocalizations providing a set of acoustic data D , the sub-model maximizing the *a posteriori* probability $P(M_i/D)$ over the 65536 sub-models M_i was selected (see Appendix for further elaboration on the computation of $P(M_i/D)$).

2.5. Simulating closant-vocant co-occurrences in babbling in the context of the Frame-Content theory

The output of the articulatory framing stage provides an estimate of available articulatory exploration at seven months for the vocant component of sequences in babbling. In the Frame-Content theory the articulatory configuration of the vocal tract is conceived as a “presetting” on which mandibular cyclicities are superimposed in babbling. A percept of vocants and closants emerges as a result of these cyclicities. The purpose of this last step was to simulate this mechanism, to determine what would be the co-occurrences between closants and vocants in this simulation, and to compare model co-occurrences with actual ones observed during this period (Davis & MacNeilage, 1995).

For each vocalization of the 7-month corpus, the possible articulatory configurations corresponding to the acoustic configuration were searched. Since this is an ill-posed problem with many solutions, an “exhaustive inversion” process was implemented. All articulatory configurations, generated by the sub-model selected by articulatory framing at seven months

that produced an acoustic output close enough to the analyzed acoustic configuration³ were selected.

Then, these inferred configurations had their jaw moved upward till closure, whatever its position (between the lips, or between the tongue and the dorsal wall of the tract) based on a procedure from Vilain, Abry, Brosda & Badin (1999). The corresponding closants were classified as being labial, coronal or palato-velar depending on the position of the closure. If closure happened at the lips (with null inter-lip area A_l), the closant was labeled as labial. If it happened inside the vocal tract (with null constriction area A_c), the closant was labeled as either coronal or palato-velar, depending on the constriction place: a border between both groups was defined 6 cm from the glottis ($X_c = 6$ cm), with coronals more anterior ($X_c \geq 6$ cm) and palato-velars more posterior ($X_c < 6$ cm). The boundary value was adapted from studies on consonant-vowel sequences on the adult articulatory model SMIP (Berrah, 1994). It was chosen so as to separate the coronal and the palatal closures induced from the articulatory configuration of an /i/ prototype in the 7-month VLAM.

The proportion of labials, coronals and velars generated by all articulatory configurations for all vocalizations in a given phonetic category were computed, in order to compare simulated closants for each vocant category. Altogether, this procedure matched each phonetic category of vocant to the set of closant places of articulation it would yield in the sub-model resulting from articulatory framing at seven months, in the framework of the Frame-Content theory.

Finally, for each closant-vocant pair generated by the previous procedure, the formant values were computed by the VLAM. For the closant, the values were computed just before the closure (that is, with a value of A_c or A_l equal to 0.01 cm²). This enabled the display of locus scatter-plots, relating F2 for the closant to F2 for the vocant. These plots are claimed to provide a good representation of vowel-plosive coarticulation (Sussman, Fruchter, Hilbert &

Sirosh, 1998). Sussman *et al.* (1999) studied locus scatter-plots for a female between 7 and 40 months. A total of 7,888 utterances were analyzed to obtain F2 values for closants and vocants for each utterance. The simulated locus scatter-plot was compared with the first scatter-plot provided by Sussman *et al.* (1999) when the child was 10 months old.

3. Results

For each corpus, acoustic framing allowed comparison of actual productions with the acoustic possibilities of the VLAM at the corresponding age. Articulatory framing resulted in reducing the VLAM articulatory exploration and hence its acoustic exploration, in order to match the set of productions at each age as well as possible. The sub-model selected by articulatory framing at seven months was then used to simulate closant-vocant co-occurrences and compare them with observed co-occurrences.

3.1. Acoustic framing

Each set of actual vocalizations belonged to the age-matched MVS⁴ (See Figures 1 & 2). Moreover, the actual data did not cover all the acoustic area they would if infants had used the whole range of articulatory configurations, according to the VLAM. More precisely, for the 4-month-old corpus, as shown in Figure 1, the vocalizations were grouped around the neutral position defined by zero values of all articulators on the VLAM. In phonetic terms, these correspond to central, mid-high to mid-low configurations: the most fronted, backed and open productions were not exploited. For the 7-month-old corpus displayed in Figure 2, the vocalic

productions explored more of the lowest part of the acoustic space: the vertical dimension seemed dominant at this age.

Please insert Figures 1 and 2

3.2. *Articulatory framing*

Let us first consider results for the 4-month-old corpus. In Figure 3, actual vocalizations are plotted on the same diagram as the whole range of vocalizations produced by sub-models with one or two non-zero articulatory dimensions. The one-dimension plots display the articulatory-acoustic sensitivity for each parameter, P_i , around the neutral configuration with P_i varying from -3 to $+3$ *std*. These sensitivities are globally of the same magnitude, though with quite different F1 and F2 extents (see also Ménard *et al.*, 2004). Altogether, the plots in Figure 3 indicate that two dimensions are not enough to reproduce the range of acoustic productions displayed in the 4-month-old corpus.

In Figure 4, the results of the best sub-models (in terms of $P(M_i/D)$ maximization) respectively for 3 and 4 dimensions are displayed showing that three dimensions are the minimum set required for reproducing acoustic exploration in the 4-month-old corpus. Within these three dimensions, at least one tongue parameter (either TB or TD or both) is required to adequately simulate infants' vocalizations. The best sub-model with three dimensions does not incorporate the jaw parameter. Among three-dimension sub-models,

models incorporating the jaw are systematically associated with lower a posteriori probabilities than models without the jaw. Among models with the four articulatory dimensions involved, large variations of range for each dimension are possible because of articulatory compensations.

For each tested sub-model, it is possible to estimate the global range of articulatory exploration by computing the number of articulatory configurations produced by the sub-model, and comparing with the number of articulatory configurations produced by the complete four-dimensional VLAM, that is if each of the four selected articulatory parameters is systematically varied between -3 and $+3$ *std*. It appears that the best sub-models, for three or four articulatory dimensions, typically exploit around 10% of the whole range of possible articulatory variations. This means that the whole volume of articulatory exploration in the 4-month-old corpus represents about 10% of the available 4-dimensional volume for J, LH, TB, TD varying in their whole possible range between -3 and $+3$ *std*.

Please insert Figures 3 and 4

For the 7-month-old corpus of canonical babbled syllables, acoustic exploration is more diverse. It is quite likely that more dimensions and larger articulatory ranges are required. Indeed, simulations indicate that three dimensions are not enough to reproduce the range of acoustic productions displayed in the corpus. On Figure 5, the results of the best sub-model (in terms of $P(\text{Mi}/\text{D})$ maximization) for 4 dimensions are displayed. In this sub-model, the jaw is the only parameter exploiting the whole $[-3, +3]$ *std* range, while the three other

dimensions occupy a reduced range. Altogether, the best sub-models, for four articulatory dimensions, typically exploit around 50% of the whole range of possible articulatory variations.

Please insert Figure 5

3.3. Simulated closant-vocant co-occurrences at seven months

In Figure 6, the percentages of closant-vocant co-occurrences in the simulations are displayed for the seven categories of vocants in the 7-month-old corpus. Front vocalic sounds and the central /a/ were most often associated with coronal closants, and also with a significant amount of palatal closures (around 20% of the cases). The central /ə/ and the back /o/ were most often associated with labial closants and also to a large though slightly lesser extent with palato-velars (around 35% of the cases). Hence, there is a set of co-occurrences, more or less in agreement with predictions provided by the Frame-Content theory. We shall come back on the fit and discrepancies between theory and simulations in the Discussion Section.

Please insert Figure 6

Figure 7 displays the simulation of locus scatter-plots at the onset of canonical babbling (Fig. 7a), compared with data for the female infant studied by Sussman *et al.* (1999), at 10 months (Fig. 7b). The simulation provides values grouped around the diagonal (F2 values for the vocant and the closant close to each other). Configurations for labial closants are rather below the diagonal and correspond to low F2 values for the vocant, while configurations for coronal closants are rather above the diagonal and correspond to higher F2 values for the vocant. Looking more closely at the distribution of vocant F2 values (displayed along the horizontal axis for each plot in Fig. 7), it appears that the range of vocant F2 values respectively for labial and coronal closants is quite similar between simulations and actual data, and it corresponds to the co-occurrences displayed in Fig. 6: back and central vocants with low F2 for labial closants, central and front vocants with higher F2 for coronal closants. Values for palato-velar closants are more widely distributed in the simulations (Fig. 7a, with both back and front vocants, as in Fig. 6) than in real data (Fig. 7b, where they mainly correspond to front vocants). The repartition of F2 values for the closants is displayed along the vertical axis for each plot in Fig. 7. It appears that there is much less clustering within and separation between labials and coronals in simulations than in real data. We shall discuss in Section 4 the possible reasons for the various discordances between model and data.

Please insert Figure 7

4. Discussion

Altogether, these analyses based on the matching of acoustic corpora of infants' vocalizations, and the productions of an articulatory model of the growing vocal tract, VLAM, suggest a progressive articulatory exploration from four to seven months, with a strong jaw involvement at the onset of babbling at 7 months. The results should be considered as preliminary, considering both the difficulty of matching actual infant productions with a model, and the large inter-individual variability in the various available corpora of infant vocalizations.

4.1. A coherent portrait of progressive articulatory exploration

To estimate their reliability, these results should be related to available knowledge about this developmental period. The results of acoustic and articulatory framing indicate that exploration at four months is reduced around a neutral vowel configuration generally considered as a rest position. Exploration involves at least three articulatory parameters, including at least one for the tongue. Moreover, the jaw seems to play a minor role at this age: it does not seem to account for a large part of the acoustic variance. Altogether, the whole volume of articulatory exploration is around 10% of the available range in the four-dimensional articulatory space provided by J, TB, TD, LH. At seven months, exploration is much larger, around 50% of the available range in the same space, and the jaw plays a dominant role leading to a large exploration of the open-close contrast and its associated F1 dimension in the formant space.

The two corpora analyzed were gathered in quite different experimental conditions, vocal imitation for the 4-month data and spontaneous canonical babbling vocalizations for the 7-month data. Vocal imitation should have lead infants to explore, as far as they could, their

articulatory-acoustic repertoire since the targets corresponded to the extreme articulatory configurations of the vocal tract. Hence, exploration in the first corpus should probably be considered as a rather maximal – and perhaps exaggerated – picture of what a 4-month-old infant might utter. In this sense, the increase of articulatory range from the 4- to the 7-month-old corpora is likely to have been *underestimated* in the present study, with regard to "spontaneous" production. The 7-month-old corpus includes only canonical babbling: this probably explains the importance of the jaw parameter in articulatory framing as this is a fundamental characteristic of this stage in speech development. Furthermore, the present results are consistent with phonetic data showing that the “vertical” dimension explains a larger part of the variance than the “horizontal” dimension in infants’ vocalizations at this stage (Davis & MacNeilage, 1995).

4.2. Presetting and co-occurrences in relation to the Frame-Content theory

The two stages of articulatory exploration analyzed in this study may be discussed in the framework of the Frame-Content (F-C) theory. First, it might appear puzzling and even counter-intuitive that the jaw seems not much involved at four months. At least, it seems to be contradictory with the jaw primacy in the first stages of speech development according to the theory. The F-C theory takes as its starting point the onset of *speech-like movements*, that is rhythmically *coordinated actions involving the vocal source and the vocal tract*. Perceptually apparent speech-like movements begin at the onset of canonical babbling around seven months. The jaw is seen as the core component of vocal tract close-open alternations. However, the F-C theory also considers that these “frames” produced by the jaw are superimposed on *presetting* of the vocal tract, that is a vocal tract shape, which involves a range of possibilities for the placing of the tongue and the lips. The presetting is stable across

jaw cycles (frames) in canonical babbling vocalizations, resulting in the production of labial-central, coronal-front or velar-back vowel depending on the presetting of the tongue. Across the period of acquisition, control of independent segmental or content elements is mastered as the child learns to control articulators independently of the jaw in vocal sequences.

In this respect, it is not contradictory with the F-C theory that presetting might occur before onset of a babbling episode, and various data on early vocal imitation suggest that presetting could even be to a certain extent *controlled* by the infant before seven months. Moreover, there is no prediction in the theory about the possible role of the jaw in early presetting. The goal of the present simulations was precisely to attempt to better analyze the presetting range. The simulations of exploration at four months indicate that the tongue is involved in presetting, and that the jaw does not seem to play a strong role. In some sense, there could be a natural unfolding, from the 4-month-old stage centered on tongue-lips presetting, towards the 7-month-old canonical babbling stage basically driven by jaw cycles. This hypothesis should be considered in further studies incorporating other experimental material.

The co-occurrence simulations in Section 2.5 provide a portrait largely compatible with F-C predictions. There is a global co-occurrence gradient in the simulations from front to back, with coronals associated with front vocants and labials and palato-velars more with central or back vocants. This patterning fits with F-C predictions.

There are however two discrepancies relative to F-C predictions. Firstly, /a/, predicted to be a central vocant predominantly associated with a labial closant, appears to be mostly associated with coronals in our simulations. Model morphology is crucial to understand the difference. Vilain *et al.* (1999) showed that various articulatory models, slightly differing in the morphology of the palate, could produce /da/ as well as /ba/ or /bda/ co-occurrences, depending on the fact that, when the jaw is closed from a neutral position, contact happens first between the lips or first between tongue tip and anterior palate. Furthermore, Vilain and

colleagues showed that the particular morphology of the VLAM, leads to coronal frames with /a/. The second discrepancy concerns the back vocant /o/, predominantly associated with labial closants and only to a lesser extent with palato-velars, which are predicted as the favored co-occurrence. Simulations on VLAM show that when the closant is labial, there is also an almost complete closure inside the vocal tract, at a position compatible with a palato-velar closure. Hence, back vocants are characterized by a kind of joint labial-velar closure. The discrepancy with predictions is hence not so large, and it could also depend on individual morphology of the palate.

The last simulation concerned locus scatter-plots. Three major inferences emerged from the pattern of results in Fig. 7. Firstly, the front-back co-occurrence gradient appears once again, with labial closants associated with back or central low-F2 vocants, and coronal closants associated with front high-F2 vocants (see displays along the horizontal axis in Fig. 7). This is also coherent with the acoustic data obtained by Sussman et al. (1999). Secondly, palato-velar closants are more widely dispersed, with both front and back vocants. This is actually both inconsistent with the data of Sussman et al. (1999) – with palato-velars mostly associated with front vocants – and with F-C predictions – with palato-velars mostly associated with back vocants. In fact, our simulations reveal that there is a significant amount of back closure for all vocants (Fig. 6), though never in a majority. As mentioned previously, individual morphology might tune this pattern differently from one child to the other.

Thirdly, there was a clear discrepancy between data and simulations concerning the distribution of F2 values for labial and coronal closants, the values being well separated in data, and not in simulations (see displays along the vertical axis in Fig. 7). This result suggests that actual data could have been produced by more complex movements than just a jaw upward trajectory. Indeed, Sussman et al. (1999) indicate that there is already a significant modification of locus equations from 7 to 12 months, and this change could induce the onset

of a control of the labial and coronal articulators involved in mature speech. Adding such specific articulators for each plosive category (i. e., LH for labial closure in labials and TT for tongue tip elevation in coronals) leads to a higher separation of closant F2 values between labial and coronal closant configurations in VLAM simulations (Serkhane, 2005). Another possible cause of the discrepancies between data and model could be due to the data themselves. There is probably a range of closant F2 values providing for a given vocant F2 a rather ambiguous stimulus, and it should be rather hard to categorize such ambiguous stimuli as either labials or coronals. In this ambiguous region, it is not impossible that the phonetician's ear naturally separates the patterns into two different classes in a perceptually consistent way, though inconsistently in articulatory terms. This could produce an *artificial increase* of the acoustic separation between articulatory clusters.

4.3. Interest and limitations of the modeling approach

Considering the difficulty of acquiring articulatory data for infants, and estimating robust acoustic speech parameters (Van der Stelt, Wempe & Pols, 2003), it is helpful to match actual data to an articulatory-acoustic model of the vocal tract, particularly considering the importance of growth mechanisms and the difficulty of integrating them into the phonetic analysis (see Ménard *et al.*, 2004).

However, it is also necessary to consider that the VLAM has the classical limitations of speech production models. It disregards inter-individual variability, which could be quite important in the simulation of closant-vocant co-occurrences (Vilain *et al.*, 1999), as we noted in the previous section. VLAM blurs morphological details and linearizes or simplifies the three-dimensional changes of the jaw, the tongue, the lips and the larynx that are due to the growth process, though staying within a reasonable approximation (Ménard *et al.*, 2004).

Most importantly, it makes use of degrees of freedom from *adults* to study *infant* vocalizations. Of course, this is arguable: in fact, nobody knows exactly what the degrees of freedom of the vocal tract at birth are, and how they evolve with age. The goal of the present analysis is to make no assumption about these unknown data, and to consider acoustic data in terms of progressive exploration of a growing tract. Considering that the remodeling of the vocal tract between 4 and 7-months is minor and cannot explain satisfactorily the vowel space expansion, the acoustic and articulatory framing results show that changes in pre-linguistic vocalization inventories over time cannot be explained by the growth of the vocal tract only, but that they also express changes in the *articulatory exploitation* of the vocal apparatus which may point to other aspects of the developmental process.

Conclusions of the present study are tentative, and constrained by the limitations of the data themselves. This analysis matrix provides preliminary confirmation of the realism of this model, considering that acoustic framing displays coherence between the VLAM productions and actual vocalizations, and that articulatory framing and clesant-vocant simulations are consistent with previous knowledge and published acquisition data. Further analyses of other experimental corpora could support the findings from this analysis in a more general way.

4.4. A preliminary step in a computational program of speech development

The present study is consistent with attempts to model speech development through the construction of a virtual robot endowed with a growing vocal tract, basic systems of perception and a learning mechanism in order to simulate the way infants progress from speech-like vocalizations to mastery of their ambient language (Serkhane, Schwartz, Boë & Bessière, 2005). These results provide the first articulatory specifications of the virtual robot.

The major result of this analysis suggests that exploration should be conceived as a progressive process. A possible developmental schedule, compatible with the results of articulatory framing, could first involve three articulatory parameters (one for the lips, two for the tongue) best able to reproduce 4-month-olds' vocalizations. From seven months, rhythmic syllables in the *canonical babbling* period would chiefly be characterized by the superimposition of rhythmic jaw cyclicity on the tract pre-settings produced by the three previous parameters. This gradual exploration would enable learning a relatively accurate sensori-motor representation of robot skills, consisting in the correspondence between the articulatory configurations of the robot vocal tract and the perceptual consequences they yield. This "map" would then adapt to the changes the robot undergoes and enable it to imitate the speech sounds perceived in its environment according to its *current* perceptuo-motor skills (Serkhane, Schwartz & Bessière, 2003).

In such a "speech robotics" project (Abry & Badin, 1996), the most efficient way to program a robot is to follow the time course of speech production development. Indeed, *developmental plausibility* is one of the basic principles of "cognitive robotics". As Brooks has pointed out, "[b]uilding robots developmentally facilitates learning both by providing a structured decomposition of skills and by gradually increasing the complexity of the task to match the competency of the [final] system" (Brooks, Breazeal, Marjanovic, Scassellati & Williamson, 1999; see also Scassellati, 1998).

5. Conclusion

This paper presents a preliminary approach in which infants' vocalizations were matched with an articulatory-acoustic model based on statistically available speech data that integrated

the non-linear growth of the vocal tract. The goal of this analysis procedure was to better characterize infants' articulatory skills. An important aspect of this modeling strategy is that it helps to disentangle morphology from control; that is, to separate acoustic variations due to the growth process from those due to changes in the way articulatory degrees of freedom of the vocal tract are exploited.

Results suggest that articulatory exploration tends to increase from four to seven months. The jaw plays a minor role before babbling, but a major role at onset of rhythmic syllable-like output in canonical babbling. The Frame-Content theory was tested in the framework of the VLAM, exploiting the tongue configurations inferred at seven months as providing possible pre-settings, predicted by the theory, for the proposed "frame" based on rhythmic jaw cyclicity. The articulatory and acoustic simulations were largely compatible with actual data, though exact replication is impossible, partly because of the role of individual variations in vocal tract morphology.

The same kind of analyses should be undertaken on a variety of other available corpora of infant vocalizations to test the generality of these findings. Further testing the validity of the VLAM for the speech acquisition period could provide an extended framework for the continuation of the developmental speech robotic program initiated with the present work.

Acknowledgements – This program was prepared with support from the European ESF Eurocores OMLL, and from the French funding programs CNRS STIC Robea and CNRS SHS OHLL, and MESR ACI Neurosciences Fonctionnelles.

References

- Abry, C., & Boë, L.-J. (1986). Laws for lips. *Speech Communication* 5, 97-104.
- Abry, C., & Badin, P. (1996). Speech Mapping as a framework for an integrated approach to the sensori-motor foundations of language. *Proc. 4th Speech Prod. Sem., Autrans*, 175-184.
- Abry, C., Cathiard, M.-A., Vilain, A., Laboissière, R., & Schwartz, J.-L. (to appear). Some insights in bimodal perception given for free by the natural time course of speech production. In G. Bailly, P. Perrier, & E. Vatikiotis-Bateson (Eds.), *Festschrift Christian Benoît*. New-York: MIT Press.
- Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *Journal of the Acoustical Society of America*, 63, 1535-1555.
- Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, 22, 251-267.
- Badin, P., & Fant, G. (1984). Notes on vocal tract computations. *STL QPSR*, 2-3, 53-108.
- Beddor, P. S., & Strange, W. (1982). Cross-language study of perception of the oral-nasal distinction. *Journal of the Acoustical Society of America*, 71, 1551-1561.
- Beck, J. M. (1996). Organic variation of the vocal tract apparatus. In W. J. Hardcastle, & J. Laver (Eds.), *Handbook of Phonetic Sciences* (pp. 256-297). London: Blackwell.

Berrah, A. R. (1994). *L'émergence des structures sonores : les syllabes consonnes-voyelles*.
Diplôme d'Etudes Approfondies, Institut National Polytechnique de Grenoble, France.

Boë, L.-J., Perrier, P., Guérin, B., & Schwartz, J.-L. (1989). Maximal Vowel Space. *Proc. Eurospeech89*, 281-284.

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., Gabioud, B., & Perrier, P. (1995a). Speech Maps Interactive Plant « SMIP ». *Proc. XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, 426-429.

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. (1999). Modeling 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

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.

Buhr, R. D. (1980). The emergence of vowels in an infant. *Journal of Speech, Language, and Hearing Research*, 23, 73-94.

Callan, D. E., Kent, R. D., Guenther, F. H., & Vorperian, H. K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research*, 43, 721-736

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research*, 38, 1199-1211.

Davis, B. L., MacNeilage, P. F., & Matyear, C. (2002). Acquisition of serial complexity in speech production: A Comparison of Phonetic and Phonological Approaches to First Word Production. *Phonetica*, 59, 75-107.

Davis, B. L., & MacNeilage, P. F. (2004). The internal structure of the syllable. An ontogenetic perspective on origins. In B. Malle (ed.) *The rise of language* (pp. 133-152) Amsterdam: Benjamin.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 106, 1511-1522.

Goldstein, U. G. (1980). *An articulatory model for the vocal tract of the growing children*.
Doct. diss: MIT, Boston (unpublished).

Green, J. R., Moore, C. A., Higashikawa, M., & Steeve, R. W. (2000). The physiologic development of motor control: lip and jaw coordination. *Journal of Speech, Language, and Hearing Research*, 43, 239-255.

Green, J. R., Moore, C. A., & Reilly, K. J. (2002). The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research*, 45, 66-79.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102, 594-621.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099-3111.

Jakobson, R. (1968). *Child language, aphasia, and phonological universals*. The Hague: Mouton.

Jordan, M. I. (1990). Motor learning and the degrees of freedom problem. In M. Jeannerod (Ed.), *Attention and performance* (Vol. 13). Hillsdale, NJ: Lawrence Erlbaum.

Jordan, M. I., & Rumelhart, D. E. (1991). *Forward models: supervised learning with a distal teacher*. Occasional paper 40. Cambridge, MA: MIT, Center for Cognitive Sciences.

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6 and 9 months. *Journal of the Acoustical Society of America*, 72, 353-365.

Kent, R. D., & Miolo, G. (1995). Phonetic abilities in the first year of life. In P. Fletcher, & MacWinney (Eds.), *The Handbook of Child Language* (pp. 303-334). Blackwell Publishers.

Koopmans-Van Beinum, F. J., & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom, & R. Zetterstrom (Eds.), *Precursors of Early Speech* (pp. 37-49). New York: Stockton Press.

Koopmans-van Beinum, F. J., & van der Stelt, J. M. (1998). Early speech development in child's acquiring Dutch mastering general basic elements. In S. Gillis, & A. de Hower (Eds.), *The Acquisition of Dutch* (pp.101-425). Amsterdam/Philadelphia: Benjamins.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, 218, 1138-1141.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. *Journal of the Acoustical Society of America*, 100, 2425-2438.

Laboissière, R., Schwartz, J.L., & Bailly, G. (1991). Motor control for speech skills: a connectionist approach. In D.S. Touretzky, J.L. Elman, T.J. Sejnowski, & G.E. Hinton (Eds.),

Connectionist Models, Proc. 1990 Summer School (pp. 319-327). San Mateo CA: Morgan Kaufmann Publishers.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, 105, 1455-1468.

Lieberman, P. (1980). On the development of vowel productions in young infants. In G. Yeni-Komshian, J. Kanagh, & C. Fergusson (Eds.), *Infant Phonology, Vol. 1: Production* (pp. 23-42). New York: Academic Press.

Locke, J., L. (1993). The beginnings of phonology in the child. In *Phonological acquisition and change* (pp. 1-49). New York: Academic Press.

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). Hillsdale, NJ: Lawrence Erlbaum.

MacNeilage, P. F. (1998). The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-511.

MacNeilage P. F., & Davis, B.L. (2001). Motor mechanisms in speech ontogeny: phylogenetic, biological and linguistic implications. *Current Opinion in Neurobiology*, 11, 567-569.

Maeda, S. (1990). 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.

Maeda, S., & Honda, K. (1994). From EMG to formant patterns of vowels: the implication of vowel systems and spaces. *Phonetica*, 51, 17-29.

Markey, K. L. (1994). *The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development*. Doct. diss. University of Colorado. Chicago.

Martin, J. A. M. (1981). Voice, speech and language in the child; Development and disorder. In G. E. Arnold, F. Winckel, & B. D. Wyke (Eds.), *Disorders of Human Communication*, vol. 4. New-York: Springer-Verlag.

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.

Meltzoff, A. N. (2000). Newborn imitation. In D. Min, & A. Blater (Eds.), *Infant development the essential readings* (pp 165-181). Blackwell.

Meltzoff, A.N., & Moore, M.K. (1997). Explaining facial imitation: a theoretical model. *Early Development and Parenting*, 6, 179-192.

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 Speech, Language, and Hearing Research*, 47, 1059-1080.

Moore, C. A., & Ruark, J. L. (1996). Does speech emerge from earlier appearing oral motor behaviors? *Journal of Speech, Language, and Hearing Research*, 39, 1034-1047.

Munhall, K. G., & Jones, J. A. (1998). Articulatory evidence for syllabic structure. *Behavioral and Brain Sciences*, 21, 524-525

Nittrouer, S. (1993). The emergence of mature gestural patterns is not uniform: evidence from an acoustic study. *Journal of Speech, Language, and Hearing Research*, 36, 959-972.

Payan, Y., & Perrier, P. (1997). Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. *Speech Communication*, 22, 185-205.

Scassellati, B. (1998). Building behaviors developmentally: A new formalism. In *Integrating Robotics Research: Papers from the 1998 AAAI Spring Symposium*. AAAI Press

Schroeder, M. R., Atal, B. S., & Hall, J. L. (1979). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In B. Lindblom, & S. Ohman (Eds.), *Frontiers of Speech Communication Research* (pp. 217-229). London: Academic Press.

Serkhane, J. E., Schwartz, J.-L., & Bessière, P. (2003). Simulating vocal imitation in infants, using a growth articulatory model and speech robotics. *Proc. XVth International Congress of Phonetic Sciences*. Barcelona, Spain.

Serkhane, J. E. (2005). *Un bébé androïde vocalisant : Etude et modélisation des mécanismes d'exploration vocale et d'imitation orofaciale dans le développement de la parole*. PhD dissertation in Cognitive Sciences, National Polytechnic Institute (INPG), Grenoble France.

Serkhane, J. E., Schwartz, J.-L., Boë, L.-J., & Bessière, P. (2005). Building a talking baby robot: A contribution to the study of speech acquisition and evolution. *Interaction Studies*, 6, 253-286.

Stark, R.E. (1980). Stages of speech development in the first year of life. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds). *Child phonology*, vol. 1: Production (pp. 73-90). New York: Academic Press,

Sussman, H. M., Duder, C., Dalston, E., & Cacciatore, A. (1999). An acoustic analysis of the development of CV coarticulation: a case study. *Journal of Speech, Language, and Hearing Research*, 42, 1080-1096.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: the orderly output constraint. *Behavioral and Brain Sciences*, 21, 241–299

Van der Stelt, J. M., Wempe, A. G. & Pols, L. C. W. (2003). Progression in vowel production: comparing deaf and hearing children. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 25, pp. 197-206.

Vilain, A., Abry, C., Brosda, S., & Badin, P. (1999). From idiosyncratic pure frames to variegated babbling: Evidence from articulatory modeling. *Proc. International Conference of Phonetic Sciences*, 2497-2500. San Francisco, USA.

Vorperian, H. K. (2000). *Anatomic development of the vocal tract structures as visualized by MRI*. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America*, 117, 1, 338-350.

Appendix

$P(Mi/D)$ is decomposed as follows:

$$P(Mi/D) = P(D/Mi) P(Mi)/P(D).$$

The sub-models have a priori the same probability to occur: $P(Mi)$ is uniform. Moreover, the distribution of the actual data $P(D)$ is the same whatever the sub-model, and can thus be discarded. The focus is therefore on $P(D/Mi)$.

$$P(D/Mi) = \prod_j P(D_j / Mi) \quad j \in \{1, \dots, N\}$$

where D_j is the j^{th} vocalization ($F1, F2$) and N is the total number of vocalizations in a given set of actual data, D . At four months, $N = 45$, and at seven months $N = 98$.

The computation of the above probability requires a two-dimensional discretization of the ($F1, F2$) plane, which was shaped into a grid of 32 boxes by 32 of constant sides along each formant axis – expressed in Hertz – and bounded by the acoustic limits of the age-matched MVS. In the MVS at four months, $F1$ varied in the range [276 ; 1640] Hz and $F2$ in [1121 ; 5141] Hz. In the MVS at seven months, $F1$ varied in the range [297; 1686] Hz and $F2$ in [1022; 5128] Hz.

Then, in a given sub-model Mi , the probability for each box to occur, $P(\text{box}_k / Mi)$, is computed:

$$P(\text{box}_k / Mi) = \text{Nb_sim_box}_k / \text{Nb_sim_tot_Mi} \quad k \in \{1, \dots, 32*32\}$$

wherein Nb_sim_box_k denotes the number of simulations falling within the k^{th} box and Nb_sim_tot_Mi the total number of simulations generated by the sub-model Mi .

$P(D_j/Mi)$ is provided by the frequency of the box where D_j is located, box_{k_j} :

$$P(D_j / Mi) = P(\text{box}_{k_j} / Mi).$$

It can be shown that $\log P(D/M_i)$ corresponds to a Kullback-Leibler distance between both distributions of D and M_i among the 32×32 boxes.

The only frequencies to be taken into account in this calculation are those of the boxes where there is at least one vocalization. However, as the vocalizations falling outside the acoustic space of a given sub-model are unlikely to be produced by this sub-model, their $P(D_j/M_i)$ were set to 10^{-200} . Hence, the more a sub-model fails to include vocalizations in its space of realization, the more its score is penalized. Conversely, since the articulatory parameters are uniformly distributed, increasing their ranges of variation, other things being equal, implies a rise in the total number of acoustic realizations, and thereby a reduction of the $P(\text{box}_{kj}/M_i)$ values. Thus, the scores of the sub-models decrease as their acoustic spaces tend to go over the edge of the acoustic regions where the actual vocalizations are. Altogether, the procedure looks therefore for models best fitting the acoustic distribution of the actual vocalizations, D .

FOOTNOTES

Note 1

Acoustic data were gathered from the published figure in the original paper.

Note 2

The corpus of formant data, together with the corresponding phonetic transcriptions, were directly provided by Chris Matyear, an author in the present paper.

Note 3

In the present investigation, the neighborhood was defined by a circle centered on the vocalization, with a radius of 0.4 Bark. The Bark scale is a perceptually motivated semi-logarithmic frequency scale defined in this work by the formula proposed by Schroeder, Atal & Hall (1979).

Note 4

The reader can note that one vocalization was out of the 7-month MVS. This is ascribable to the formant measurements and/or to the modeling choices, as this vocalization could have been included in the MVS if the ranges of variation of the command parameters had been slightly widened around the $[-3, +3]$ *std* interval.

CAPTIONS

Figure 1: Acoustic framing of 4-month-olds' vocalizations (black dots). Gray dots correspond to the 4-month MVS. Formants are expressed in Hertz.

Figure 2: Acoustic framing of 7-month-olds' vocalizations (black dots). Gray dots correspond to the 7-month MVS. Formants are expressed in Hertz.

Figure 3: Articulatory framing of 4-month-olds' vocalizations by (a) one- and (b) two-dimensional models (one or two parameters vary between -3 and 3 *std*, the other parameters are set to zero). For each plot, gray dots correspond to acoustic stimuli generated by the model and black dots correspond to actual vocalizations.

Figure 4: Articulatory framing of 4-month-olds' vocalizations by the best three-dimensional (a) and four-dimensional (b) models. For each plot, gray dots correspond to acoustic stimuli generated by the model and black dots correspond to actual vocalizations. The best three-dimensional model does not involve the J parameter. Both models exploit about 10% of the available 4-dimensional articulatory volume.

Figure 5: Articulatory framing of 7-month-olds' vocalizations by the best four-dimensional model. For each plot, gray dots correspond to acoustic stimuli generated by the model and black dots correspond to actual vocalizations. The best four-dimensional model involves the whole range of the J parameter. It exploits about 50% of the available 4-dimensional articulatory volume.

Figure 6: Percentage of closant-vocant co-occurrences generated by jaw (J) upward movements in the sub-model resulting from articulatory framing at seven months.

Figure 7: Locus scatter-plots produced by jaw (J) upward movements in the sub-model resulting from articulatory framing at seven months (a) compared with those of actual vocalizations in the Sussman *et al.* (1999) study at 10 months (b). In each sub-plot and for each of the three closant groups, the repartition of F2 values for vocants is plotted along the horizontal axis and the repartition of F2 values for closants is plotted along the vertical axis. Repartitions are plotted as Gaussian distributions, the mean and variance of which are estimated from the data distributions.

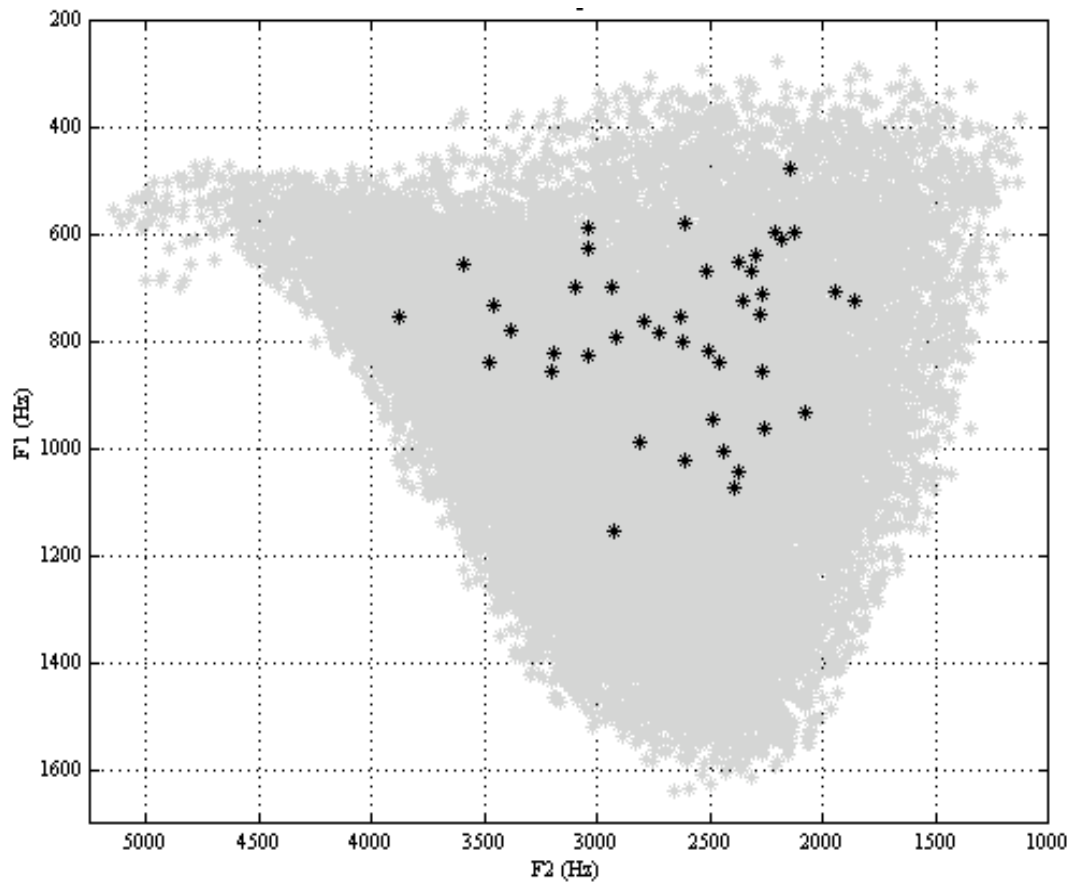


Figure 1

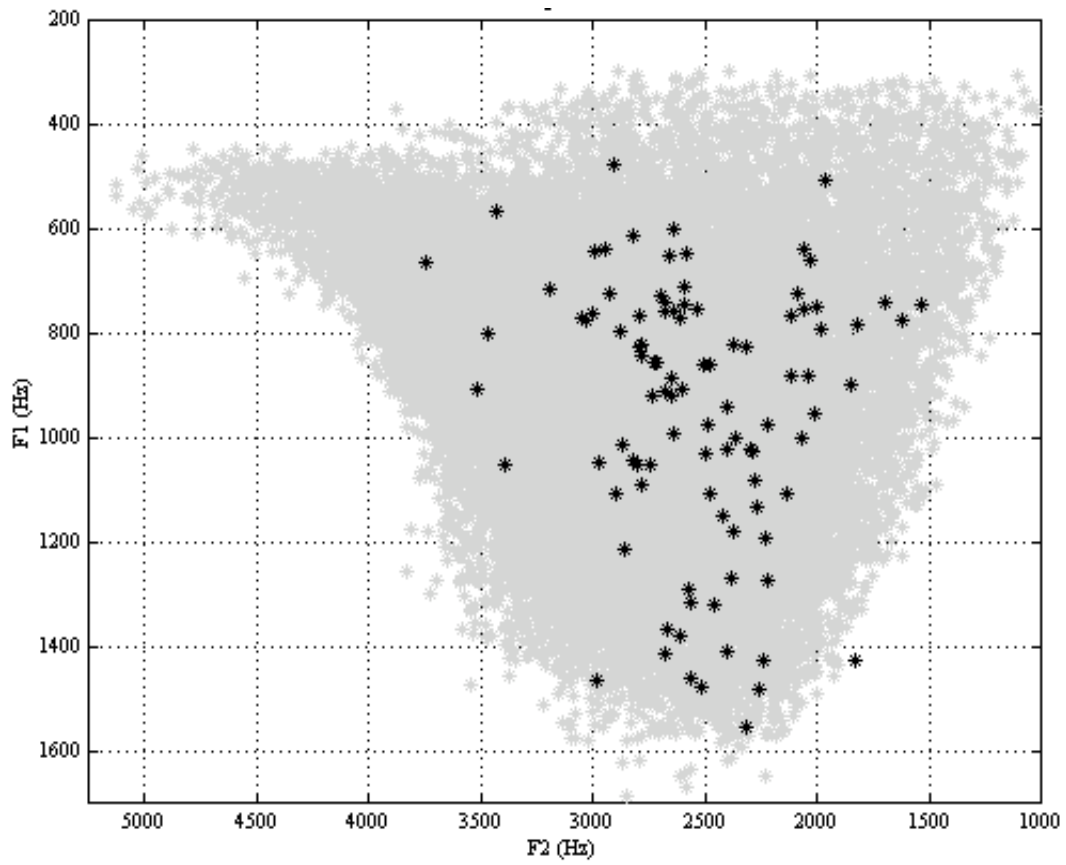


Figure 2

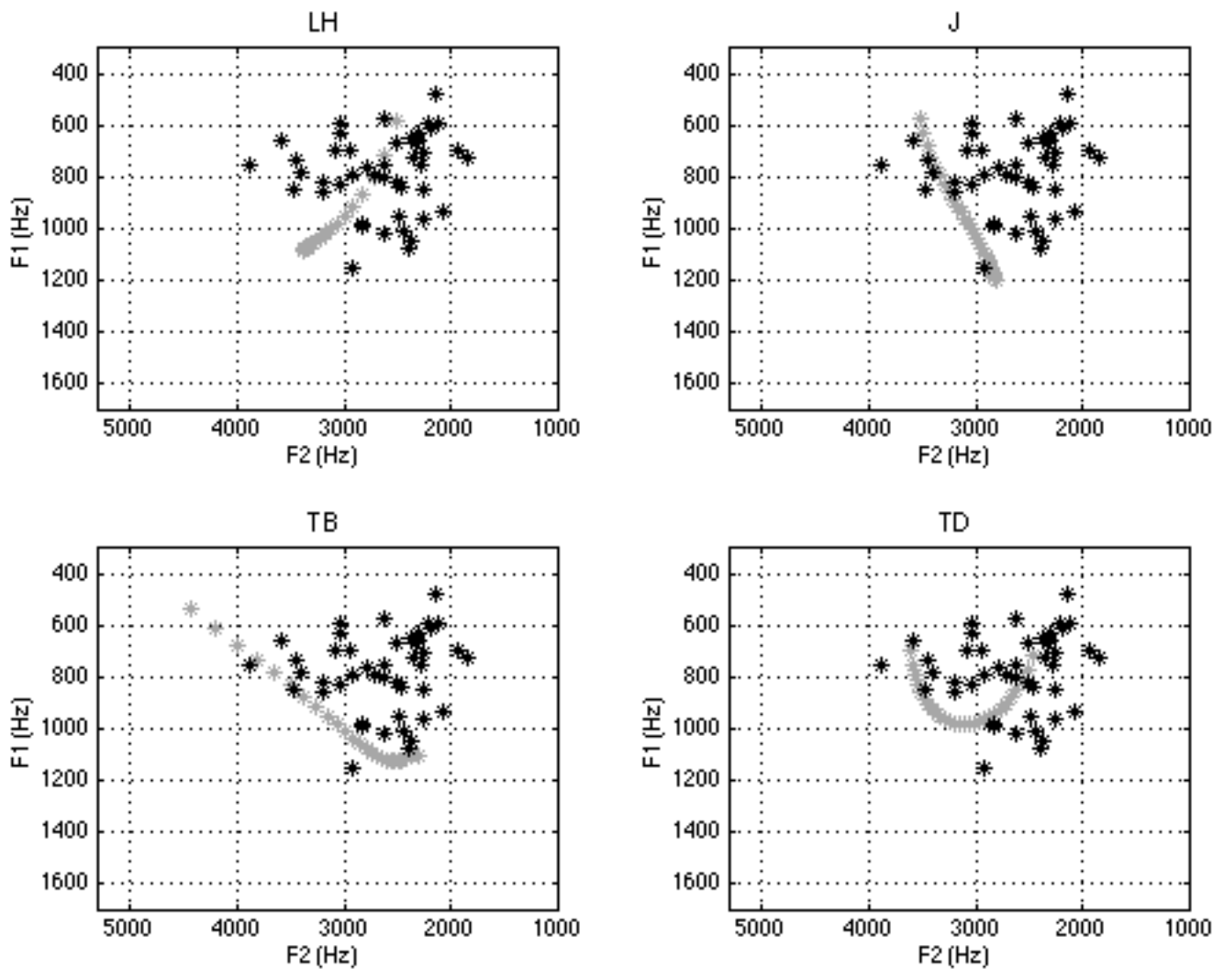


Figure 3a

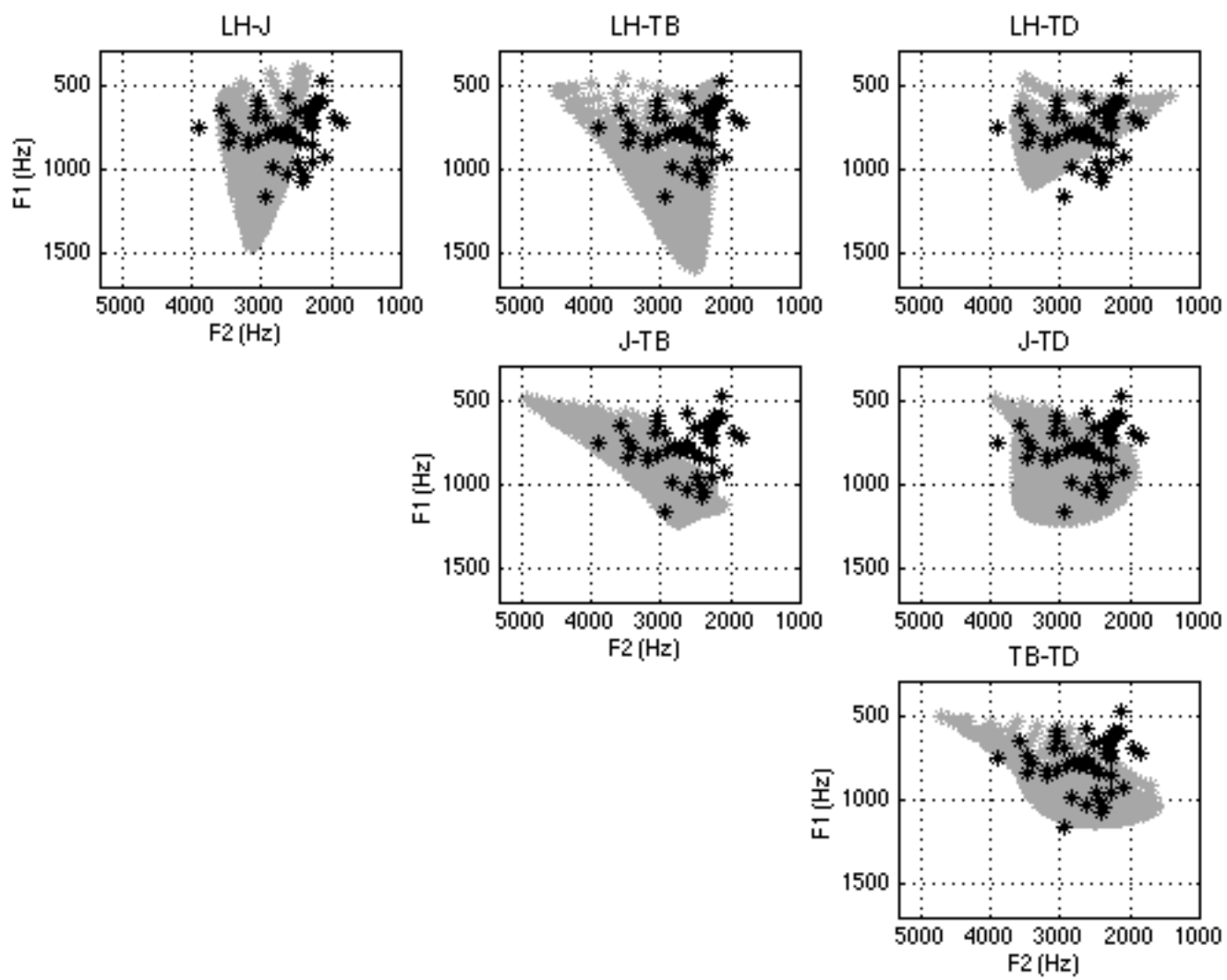


Figure 3b

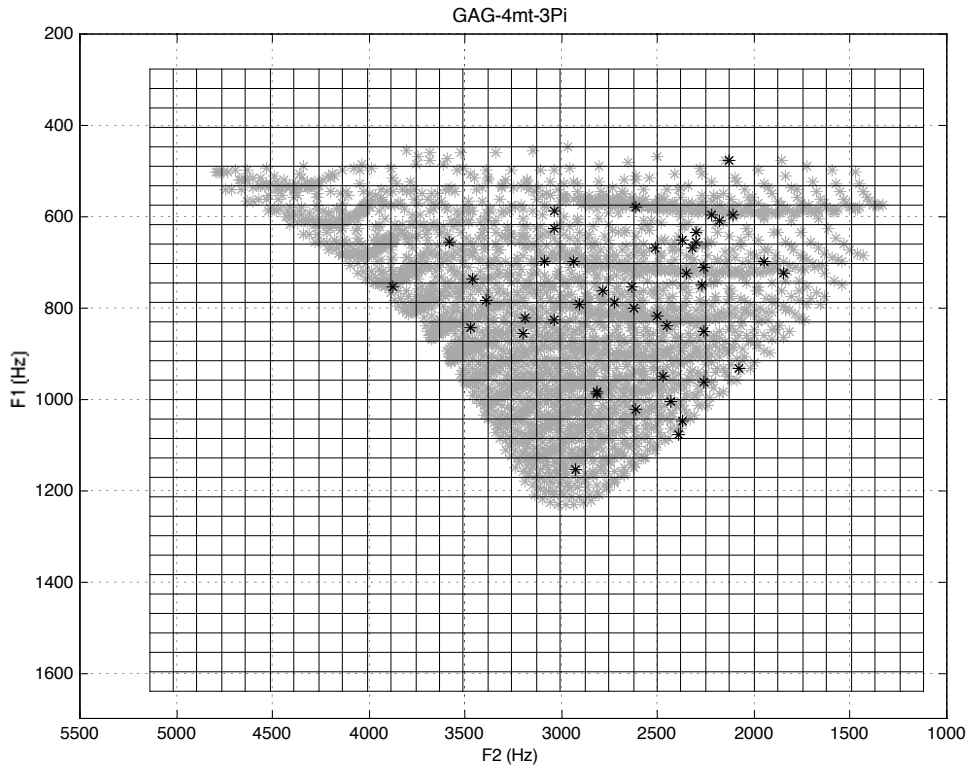


Figure 4a

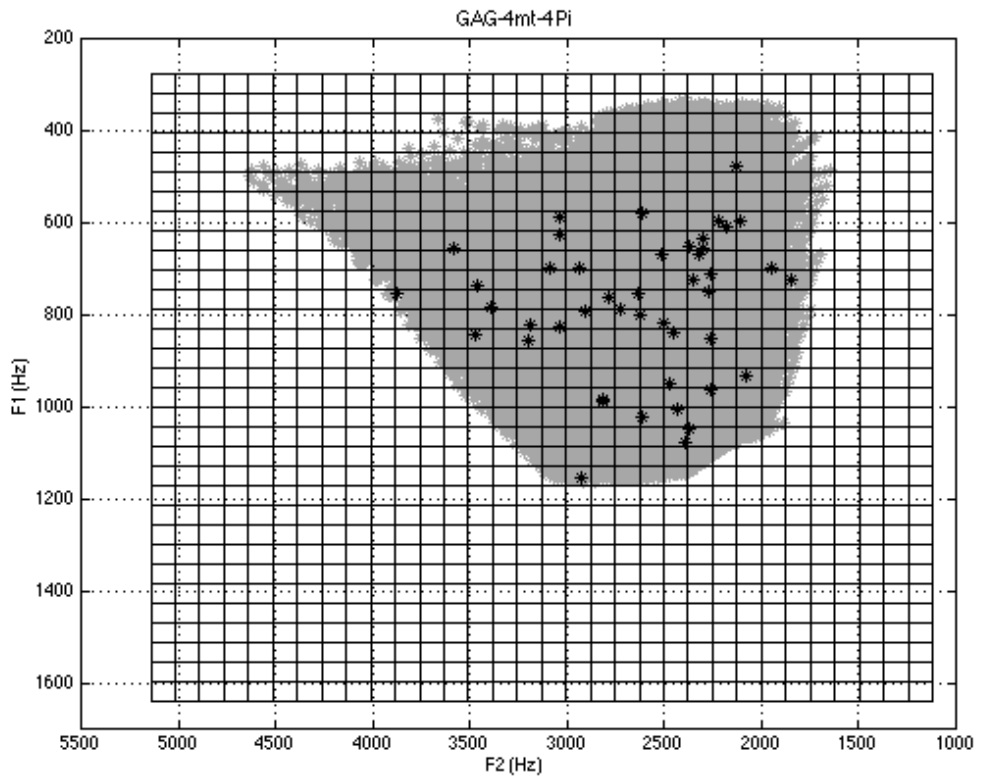


Figure 4b

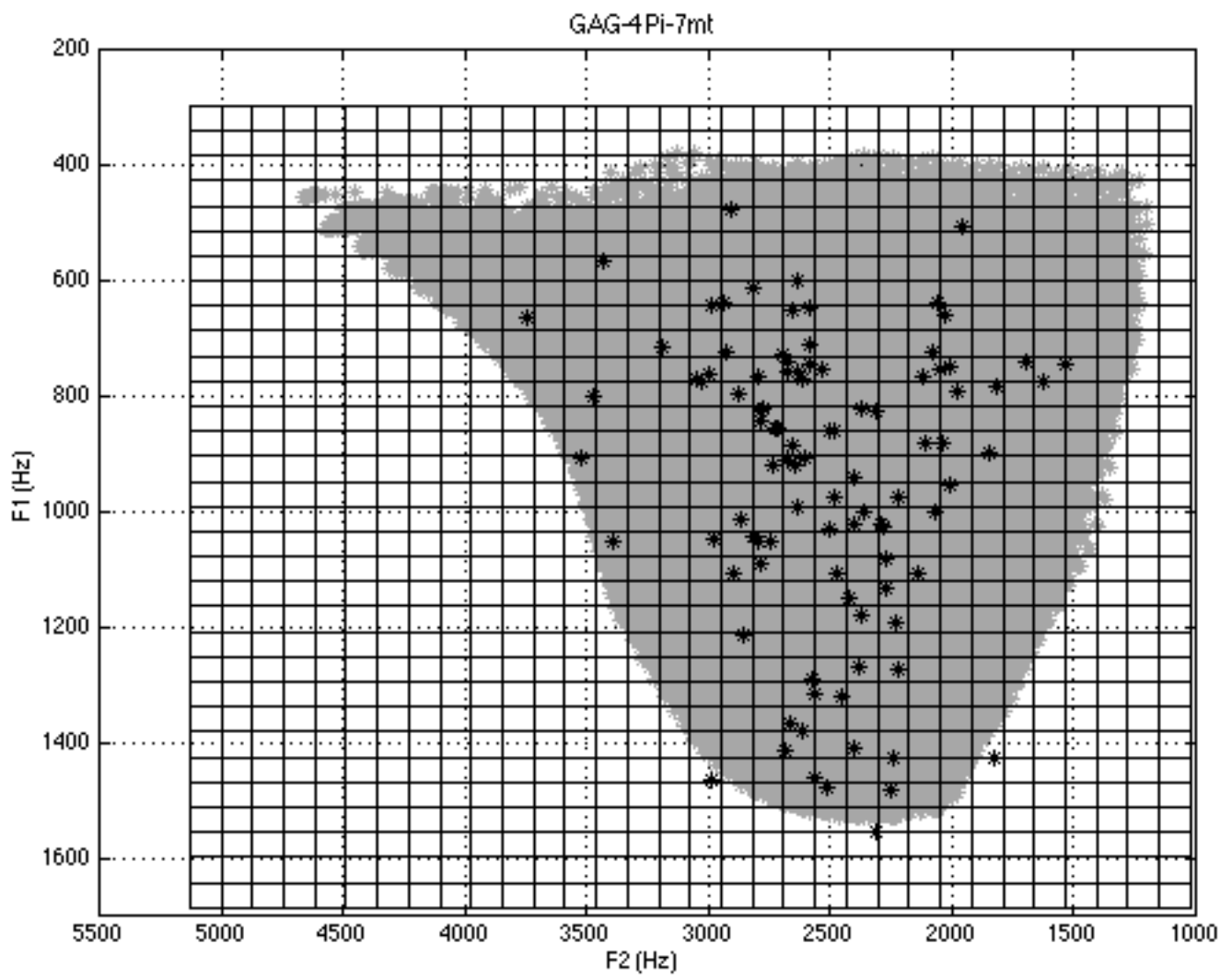


Figure 5

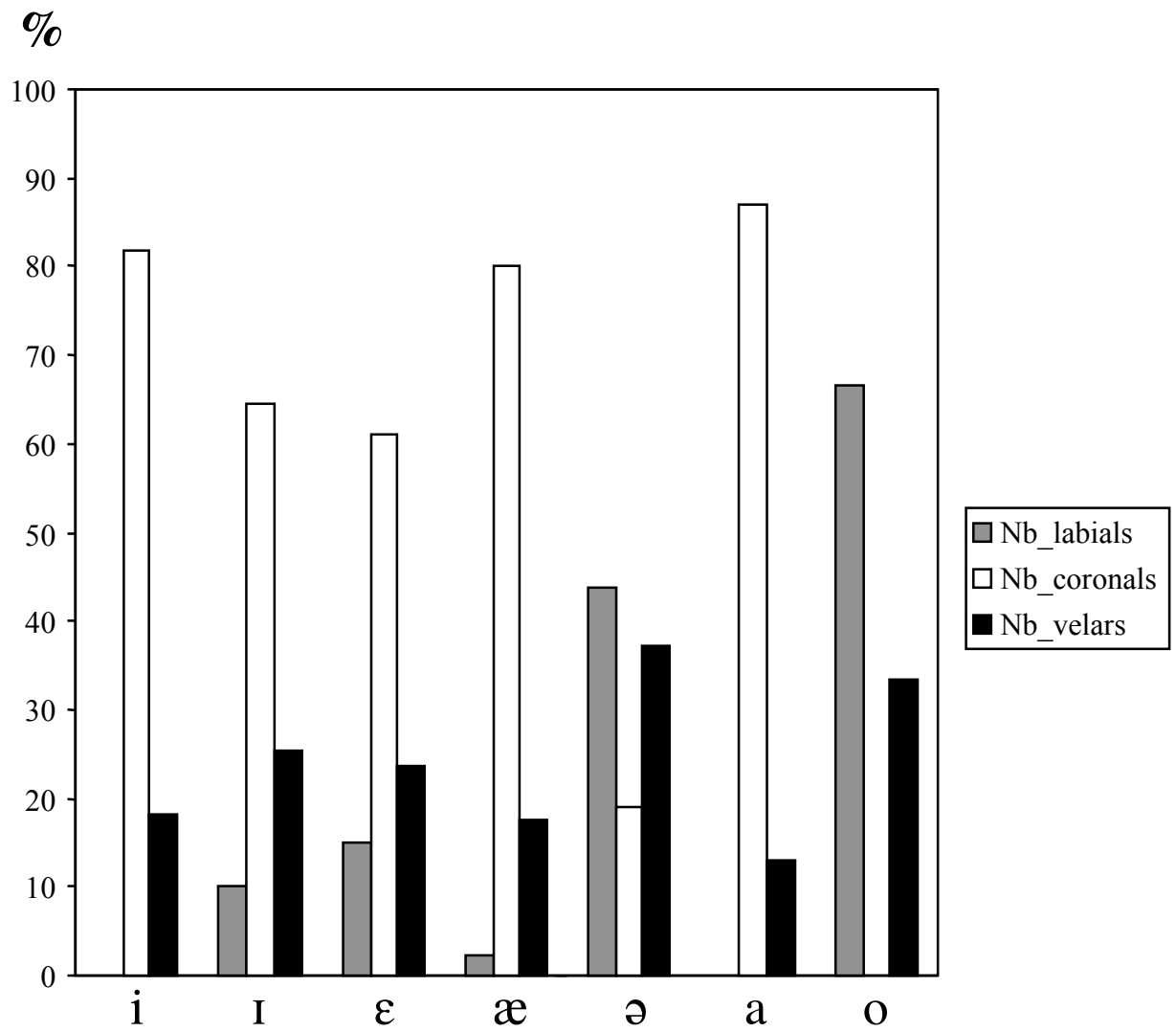
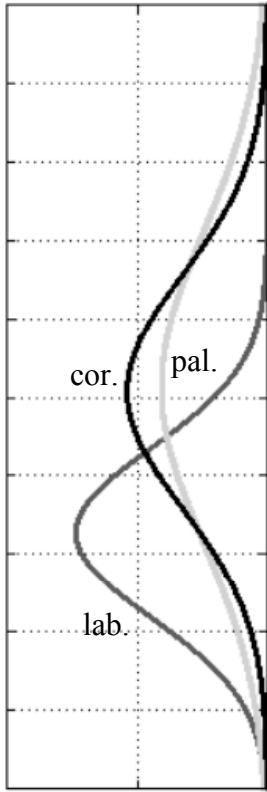


Figure 6

- : labial closure
- + : coronal closure
- : palatal/velar closure



Prob. occurrence

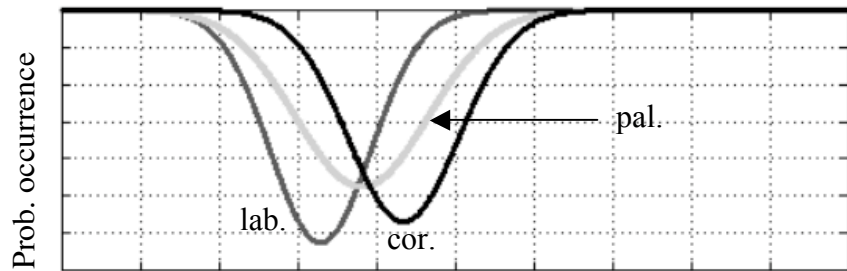
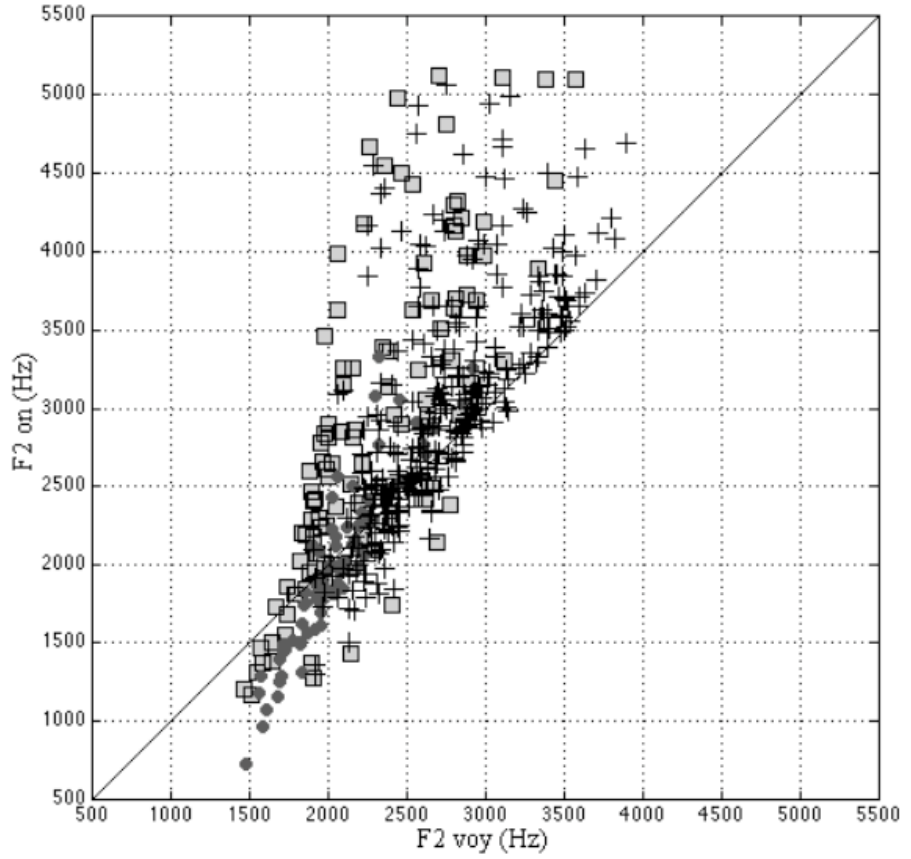
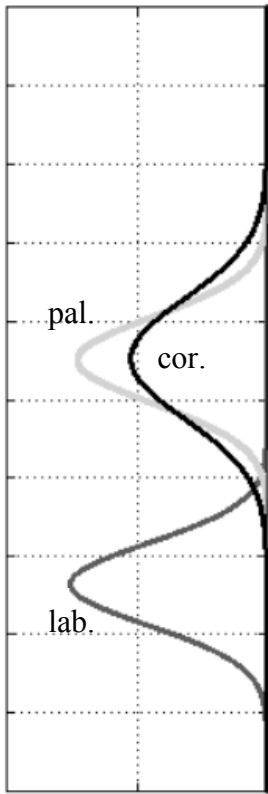


Figure 7a

- : labial closure
- + : coronal closure
- : palatal/velar closure



Prob. occurrence

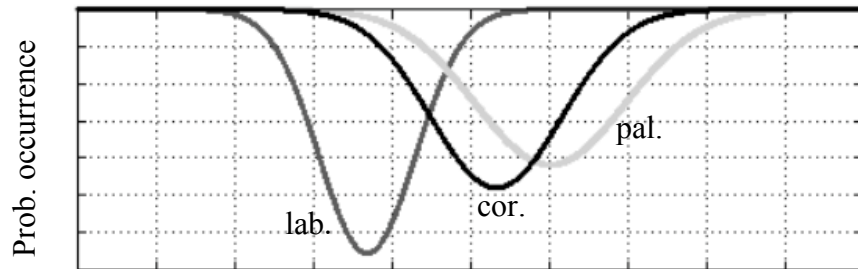
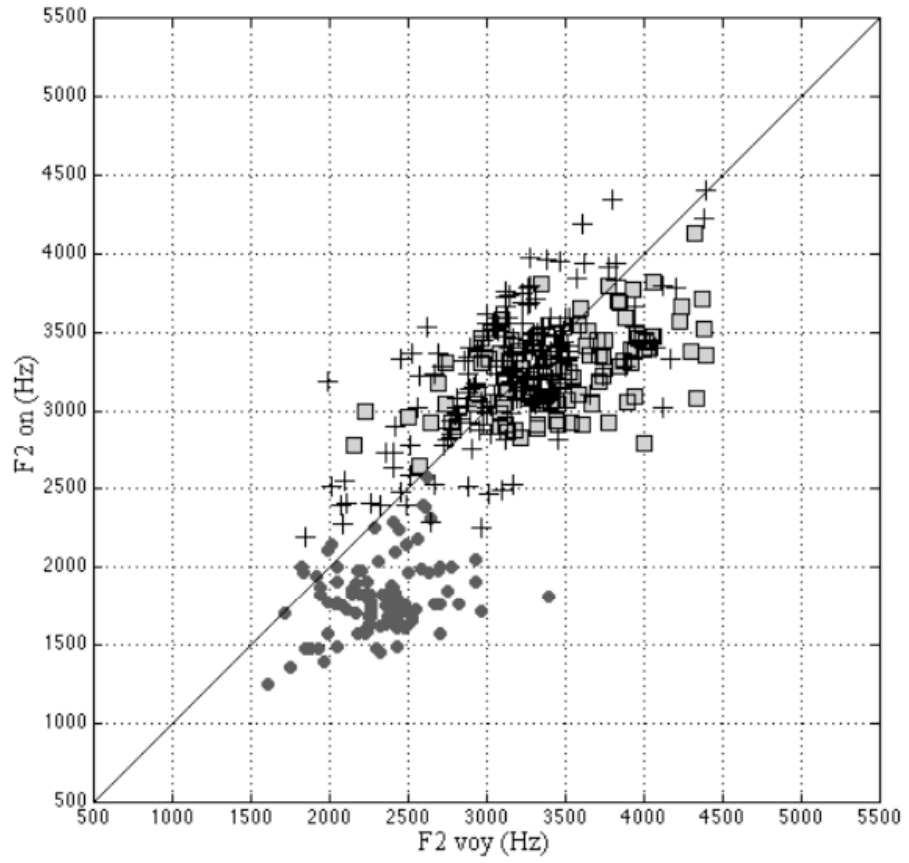


Figure 7b