Journal of Phonetics 41 (2013) 379–392

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Letter to the Editor

Anatomy and control of the developing human vocal tract: A response to Lieberman

A B S T R A C T

Since Lieberman and Crelin (1971), the question of vocal tract abilities and the link between anatomy and control has been the object of a number of conflicting papers. Part of the debate concerns the acoustic possibilities of the Variable Linear Articulatory Model (VLAM), an articulatory model that has provided the foundation of our own work for many years. VLAM is considered by Lieberman and some others as misleading because of its supposed overestimation of phonetic capabilities of human newborns. In this paper, we compare the VLAM simulations between 0 and 5 years with acoustic data on infant and child vocalizations from a number of studies in the literature. We show that the agreement is globally quite good, with no hint of overestimation above the age of 6 months for first formant and 15 months for second formant, while on the contrary simulations assessing the hypothetical role of proportions in an angled vocal tract with another model clearly diverge from ground truth child data. We conclude that limitations in infancy are a matter of control rather than anatomy. Then we lay a framework to situate "efficient acoustic modulation" within speech communication in general. We propose that the Frame-Content (FC) Theory by MacNeilage and Davis (2000) provides the basis of a *vertical* first component of a "principle of efficient modulation," giving birth to *manner of articulation*. We further propose that constriction control is the basis of the *horizontal* second component of efficient modulation, giving birth to *place of articulation*. These linked components provide a valid foundation for exploring the development of human vocal tract anatomy and control, now in two dimensions. We close by summarizing our own perspective on the possible role of swallowing in the evolution of this control, as a possible extension of the role of mastication in FC.

1. Introduction

In a recent paper in this journal, Lieberman (2012) returns to a recurrent theme in his work, the hypothesis that evolution of the acoustic potential of the human vocal tract toward a robust speech communication system must have involved modifications of vocal tract anatomy, and in particular changes to the proportions of the angled vocal tract including the descent of the larynx. He focuses a significant part of his paper on our work with the VLAM articulatory model and argues that it overestimates phonetic capabilities of newborn humans. This he considers a crucial point, since VLAM produces the same vocalic distinctions regardless of larynx height and without his assumptions of vocal tract angle or proportions. Lieberman also capitalizes on a paper by de Boer and Fitch (2010) claiming that VLAM "does not incorporate a level of physiological detail adequate to know (or even surmise) the tongue deformations available to a nonhuman or a very young infant."

As our response to Lieberman (2012) (unless noted, further Lieberman references are to this article), rather than contributing further to the polemics – which we find sometimes quite harsh in his paper – we prefer to attempt to advance the scientific debate along three major lines of investigation.

First, in Section 2, we compare acoustic data on infant and child vocalizations from the literature with age-appropriate VLAM simulations. We show that the agreement is globally quite good, with no overestimation of the vowel space beyond the 1-year-old simulations. By contrast, de Boer's simulations (2010) to assess the effect of larynx height on formant ranges, which provide the underlying computational argument for the papers by Lieberman and de Boer and Fitch (2010), clearly diverge from the ground truth of actual child data.

Then, in Section 3, we situate Lieberman's questions about the link between anatomy and control within a global framework for communication based on a "principle of efficient modulation". MacNeilage and Davis' Frame-Content (FC) Theory (2000; MacNeilage, 1998) offers the initial pillar supporting that principle. We demonstrate that the range of formant variations in the vowel space is a direct consequence of constriction control, which thereby constitutes the second pillar of efficient modulation.

Finally, in Section 4, we return to the question of anatomical constraints on constriction control, and we propose a mechanism whereby accurate control of place of constriction could have evolved during the emergence of speech communication. This in our view provides a possible extension to the FC Theory.

2. On the acoustic capacities of infant vocal tracts: back to ground truth

To estimate the acoustic capacities of the vocal tracts of infants and children, we would require certain kinds of information. We first need a description of vocal tract size and morphology from birth to adulthood, and this is now available thanks to a number of morphological studies (e.g.

DOI of original article: http://dx.doi.org/10.1016/j.wocn.2012.04.001

0095-4470/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2013.04.001

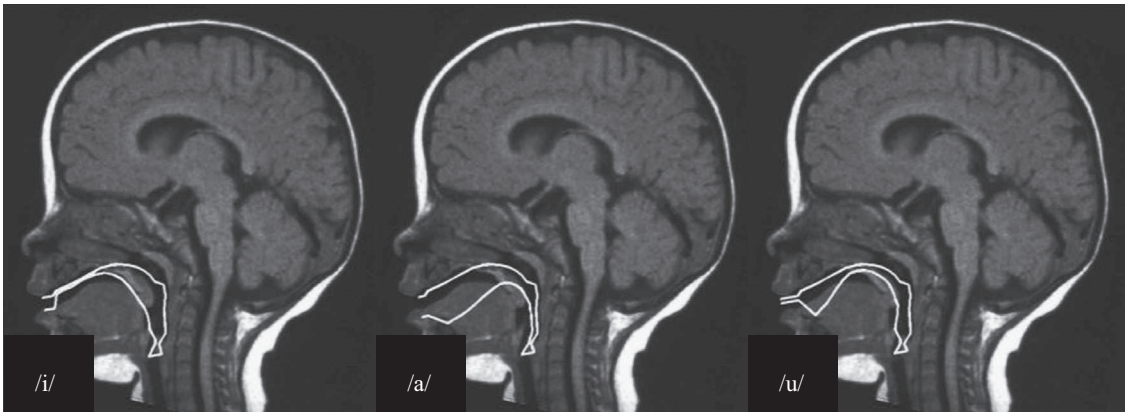

Fig. 1. VLAM-simulated vocal tract shapes for a 7-month-old child uttering [i], [a] and [u], superimposed successively on an MRI view of a 7-month-old child (Vorperian, personal communication—note that the velum is low and the nasal passage open for respiration in the MRI view, which explains the visual discrepancy between VLAM and MRI in that region).

Captier, Boë, & Barbier, 2010; Fitch and Giedd, 1999; Goldstein, 1980; Vorperian et al., 2005). Second, we need data on the possible tongue configurations in the vocal tracts of infants and young children, and/or descriptions of the underlying musculature and its potential effects on the orofacial system at the same ages. Such data are lacking, and this gap is where articulatory models are of invaluable help, since they provide predictions relating vocal tract deformations and acoustic outputs. However, they are also speculative, since they are based on assumptions about vocal tract shapes that cannot be verified against real data on possible tongue shapes.

We do have at our disposal, though, numerous studies providing data about infant vocalizations at various ages. In the following, we compare the data from these studies with the predictions of two articulatory models: VLAM, and the model introduced by de Boer (2010). De Boer's model provides the basis for his claim "that there is an optimal larynx height at which the largest range of signals can be produced and that at this height, the vertical and horizontal parts are approximately equally long" (de Boer, 2010, p. 351), a claim which is then extensively exploited by de Boer and Fitch (2010) and by Lieberman to argue that an adult-human-like vocal tract configuration is required to produce point vowels [i], [a], [u]. In this section, we attempt to compare simulations with both models to data from the literature.

2.1. Articulatory models

2.1.1. The variable linear articulatory model

The Variable Linear Articulatory Model (VLAM) (Boë, 1999; Boë, Heim, & Abry, 2002a, 2002b) is based on the model developed by Maeda (1990) from the statistical linear analysis of vocal tract shapes extracted from cineradiographic data on one adult female French speaker. Maeda's model is controlled by 7 articulatory parameters (mandible height, lip height and width, larynx height, and tongue body, dorsum, and tip), which enable efficient reconstruction of the original tongue shapes in the cineradiography corpus. VLAM assumes that the scaling down from an adult to a child of any age just implies a linear stretching of the front and back parts of the vocal tract according to morphological data on maturing vocal tracts (from Goldstein, 1980). VLAM's initial version was developed by Maeda (Boë & Maeda, 1997), and it has subsequently been systematically tested and improved (Boë, 1999; Boë, Heim, Honda, & Maeda, 2002b; Ménard, 2002; Ménard, Schwartz, Boë, & Aubin, 2007; Serkhane, Schwartz, & Bessière, 2003, 2007).

The misrepresentation of VLAM shapes as presented by Lieberman in Fig. 2 of his paper must be addressed. In his figure, Lieberman displays what he claims to be the palatal region generated by VLAM, and notes that it seems to extend through the nasal cavity of an alleged newborn.[1] Actually, the VLAM shape Lieberman chooses to present was never published, and comes from a preliminary figure that was only presented orally in a conference without published proceedings (Boë et al., 2002a). Soon thereafter, we realized that the palate vault was too high for newborns and infants, and we corrected the model accordingly (Boë et al., 2007a, 2007b). Specifically, the palate was slightly flattened to simulate ages under 2 years in order to fit the posterior nasal spine to average data from Fenart (2003). Note in passing that the precise shape of the palate, though important for realism of the articulatory model, actually plays almost no role in determining the acoustic output of the simulations (Fant, 1960; Lieberman & Blumstein, 1988). Regardless, corrected shapes have been presented in various publications, beginning in 2007 (e.g., Boë et al., 2007a, 2007b, Fig. 9, Boë et al., 2008, Fig. 7). To again illustrate the articulatory acceptability of the corrected shapes, Fig. 1 presents three VLAM simulations of a 7-month-old's vocal tract, each superimposed on an identical mid-sagittal MRI image of a 7-month old female. While no model will track real images perfectly, due to individual variations in size or soft tissue for instance, these VLAM simulations are clearly compatible with this infant's fixed contours at the hard palate and the posterior pharyngeal wall. The simulations are articulations of the point vowels [i, a, u] that a 7-month-old could produce – according to the model – if it had the appropriate vocal tract control.

2.1.2. De Boer's implementation of Mermelstein's model

The model used by de Boer (2010) is a highly simplified adaptation of Mermelstein's (1973) articulatory model. He retains only four parameters to generate midsagittal vocal tract contours. "Jaw angle" allows positioning the height of the lower incisors edge. The tongue is represented by a circle with a constant radius of 2 cm. The location of its center is determined by a "tongue angle" relative to the "jaw angle" and a "tongue displacement" parameter that is the distance of the tongue center from the jaw hinge. The hyoid bone's horizontal coordinate constitutes one more control parameter. The hyoid's vertical position is adjustable, but is not a control parameter: it depends on a "Larynx Depth" parameter considered by de Boer as an anatomical parameter fixed once for each instance of the model. The posterior-superior (palatopharyngeal) wall is modeled by 3 fixed elements,

[1] Actually, the supposed newborn in Fig. 2 of Lieberman's paper cannot in fact be a newborn, since it already has canine dentition. Notice further that the infant pictured also has the larynx positioned much too low for a newborn. We measured the Larynx Height Index (LHI, cf. Boë et al., 2002b) on Lieberman's figure, and found a value of 0.6, which corresponds to a larynx height plausible not for a newborn, but for a child between 1 and 2 years old (Boë et al., 2007).

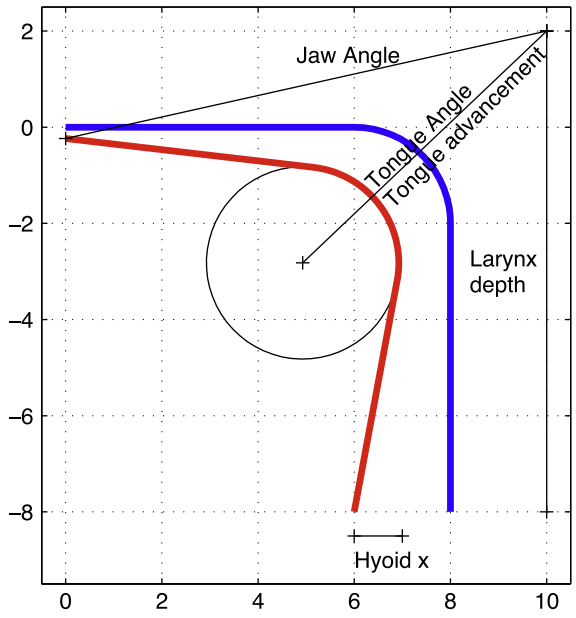

Fig. 2. De Boer's adaptation of Mermelstein's model.
(From de Boer, 2010.)

horizontal and vertical lines in the superior and posterior positions respectively, joined by a quarter circle of the same 2 cm radius as the tongue. The anterior-inferior wall (tongue surface) is modeled by three similar elements: the tongue circle; a tangent off the tongue circle to the edge of the lower incisors, representing the tongue blade; and another tangent from the tongue circle to the hyoid bone, representing the tongue back and root. In his study, de Boer (2010) explores the influence on the F1/F2 formant space of the "Larynx Depth" parameter that sets the larynx vertical position and thereby controls the length of the vertical tube. He maintains the horizontal tube at a constant length.

2.1.3. *Model validation*

It is fair to say that none of these models – not VLAM, Mermelstein's, nor de Boer's, nor in fact any other – is based on firm articulatory evidence, considering the lack of articulatory and/or musculature data for infants and children. Regarding VLAM, the same degrees of freedom – and the same range of possible configurations of the vocal tract — are presumed to hold regardless of age, the only difference being in scaling. This is however merely an assumption which remains to be experimentally validated. As for de Boer's adaptation of Mermelstein's model, de Boer claims that: "A geometrical articulatory model, such as Mermelstein's model and the simplified model studied here implements a number of important constraints that are universal to all mammalian vocal tracts" (de Boer, 2010, p. 356). However, neither argument nor evidence from data is provided to support this claim. VLAM produces simulations pointing to the conclusion that vocal tract shape is not relevant, and that the acoustical potential for contrasting vocalic contrasts is already available soon after birth – if control were available. De Boer on the other hand argues from his simulations that "there is indeed a larynx depth that results in the largest possible acoustic area covered [, which] corresponds to an almost equal length of the horizontal and vertical parts of the vocal tract. … For smaller larynx depths, the decrease in area is caused by a dramatic reduction in the extent of the second resonance, while the extent of the first resonance stays more or less the same" (de Boer, 2010, p. 361). Can acoustic data produced by infants and young children shed some light on this controversy?

2.2. *Acoustic data*

We have already published (Ménard, Schwartz, & Boë, 2004) a comparison between VLAM simulations and acoustic data at various ages, focusing on minimum and maximum F1 and F2 values as modeled by VLAM versus those measured in 5- to 19-year-old American English speakers (previously reported in Lee, Potamianos, and Narayanan (1999)). Below, we compare data and models for even younger talkers, below 5 years of age.

To that end, we have analyzed data available in a number of descriptive studies of vocalizations in infants and young children.[2] In the present study, we have compiled formant values reported in these studies (see Table 1).

Utterances were spontaneous for most children. Exceptions were the three groups of 4-year-olds, who did either repetitions (Barbier, Perrier, Ménard, & Boë, 2012; Ménard et al., 2007) or imitations (Kent & Forner, 1979). Note that we have also included data on two children 1–3 years of age published by Lieberman (1984): 223–224 himself. Also, data from the studies by Sussman, Minifie, Buder, Stoel-Gammon, and Smith (1996, 1999) are focused on F2 and hence provide no F1 data (they appear in a different color in the figures on the PDF version).

Two parameters affected our analysis of the studies in Table 1, with an impact on how we extracted the formant values we use in Fig. 2. One factor was method of reporting, and the other was data reported. Regarding reporting method, only Lee et al. (1999), Ménard et al. (2007) and Barbier et al. (2012) reported numeric formant values, while the others reported their data in graphic form. For those reporting graphically, we scanned the figures to accurately determine numeric values, and converted the values to Hertz where necessary.

As to data reported, some studies documented either formant values for a limited number of sample vocalizations, or one set of formant values for each pertinent category, taking the mean over multiple tokens. For these studies, we used the reported values directly and took their minimum and maximum values for F1 and F2. Other studies reported formant values for a large number of tokens. For these studies, to derive a single value for

[2] Some of these studies were analyzed by Vorperian and Kent (2007), but their focus was on vocal tract size and normalization rather than on formant extrema and ranges.

Table 1
Summary of descriptive studies of vocalizations in infants and young children.

	Reference	*N*	Ages	Data reported	Value used
×	Kuhl and Meltzoff (1996)	3×24	12, 16, 20 weeks	37–105 samples per age	Extrema
○	Buhr (1980)	1	16, 24, 41, 62 weeks	Representative samples	Extrema
+	Matyear et al. (1998)	3	7 months	>100 samples	Means of next 3
□	Kent and Murray (1982)	7	9 months	>50 samples	Extrema
✱	Boysson-Bardies de et al. (1989)	20	10 months	>100 samples	Means of next 3
✡	Sussman et al. (1996)	1	12, 21 months	>100 samples, F2 only	Means of next 3
✸	Sussman et al. (1999)	1	10, 17, 37 months	>100 samples, F2 only	Means of next 3
★	Lieberman (1984)	1	66, 147 weeks	40–50 samples per age	Extrema
▲	Kent and Forner (1979)	9	4 years	Mean values	Mean values
▼	Barbier et al. (2012)	4	4 years	30–60 samples per child	Means of next 3
■	Ménard et al. (2007)	5	4 years	>100 samples for each child	Means of next 3
●	Lee et al. (1999)	10–25	5–19 years	Means, each vowel at each age	Extrema of means±2SD

Note. Each reference is preceded by a symbol linking the study to its graphic representation in Figs. 3 and 4.
"N" is the number of subjects in the study. Kuhl and Meltzoff (1996) had 3 age groups of 24 each; Buhr (1980) had a single subject studied longitudinally; and Lee et al. (1999) had between 10 and 25 subjects per age group.
"Ages" is the age value provided by the author of the study in the corresponding publication.
"Data reported" describes the formant values published in the listed studies.
"Value used" describes the method used to derive minimum and maximum formant values from the data reported. See text for discussion.

the formant extrema, we pooled across the multiple talkers and multiple vowels reported, discarded the most extreme value, and took the mean of the next three values. For Boysson-Bardies de et al. (1989), the "next three" values for 4 languages were pooled and averaged. Lee et al. (1999) reported means and standard deviations by vowel for each age group, so we calculated the range of mean±2SD and scanned across vowels for minima and maxima by age. The final result was that we were able to ascertain formant extremes from each of Table 1 studies to use in Fig. 3.

Of course, data in all these studies have the possible limitations classic to all studies concerning formant measurements at such young ages, and in particular (i) difficulties in formant measurements, (ii) small number of participants, and (iii) inter-individual variability, which may be quite large among children at these ages. Nonetheless, they still provide invaluable ground truth data for comparison, and we shall see that they reveal a clear and coherent pattern.

2.3. Comparing simulations with actual data from infants and children

2.3.1. Comparing children's measured formant minima and maxima with VLAM simulations

A systematic exploratory survey through the entire range of all seven VLAM degrees of freedom was done for VLAM versions varying in age from "birth" to adulthood. For each VLAM configuration, the model generated a two-dimensional mid-sagittal section, as well as the corresponding area function (three-dimensional equivalent). The acoustic transfer function was generated by means of a model with standard losses (Badin & Fant, 1984) and the formants were extracted from this transfer function. Maximum and minimum values of F1 and F2 were then extracted from the formant spaces obtained from the VLAM configurations at each tested age, and used for comparison with ground truth data, the child formants measured in the studies of Table 1. Unfortunately, de Boer (2010) does not provide analogous maximum and minimum formant values for his simulation, but only their ranges (i.e., the difference between maxima and minima for either F1 or F2), which we discuss in the following section.

In Fig. 3, we compare VLAM predictions (solid lines) with the data outlined in the previous section (symbol key provided in Table 1). The top panels provide an overview from birth to adulthood, and suggest that the global agreement between data and VLAM is quite good. The bottom panels zoom in on the lower ages, from 0 to 5 years of age, which are our focus in this study. Let us analyze the data in the bottom panels more in detail.

First, consider the data for F1 on the left. For F1 minimum values, data and simulations are quite close. Maximum values are more dispersed, but it appears that experimental data are close to or above VLAM simulations in a number of studies below 1 year of age, in particular for four studies below 10 months. A polynomial fit of the experimental data, providing a kind of mean trend line, confirms that at one year of age, surveyed F1 reaches maximum values conforming closely to VLAM predictions.

The situation is a bit different for F2, shown on the right. While minimum values are quite close for experimental data and VLAM simulations at all ages, maximum values for experimental data are below VLAM values for most studies under 1 year of age, and apart from two studies (including one by Lieberman himself) they only approach or exceed VLAM values around 15–18 months. This is confirmed by the mean trend provided by the polynomial fit.

2.3.2. Comparing surveyed formant ranges with both de Boer and VLAM simulations

From the data discussed above and in Fig. 3, we can compute the *ranges* of formant exploration attested in children from birth to 5 years, and compare them with predictions both from VLAM and from de Boer's paper, which provides precisely this information.

In Fig. 4 we plot F1 and F2 ranges in Bark[3] for actual child data and for VLAM predictions. We again analyze F1 and F2 separately.

For F1 (Fig. 4, left), the range is more or less stable around 6 Bark, and the ranges for data and model are very similar, beginning with the very first studies before 6 months and throughout the period from 0 to 5 years.

For F2 (Fig. 4, right), the range for VLAM is stable around 8 Bark (and remains stable at ages above 5 as well). In experimental data, the F2 range is above 6 Bark for the majority of studies in the first year of life, but it does not attain the 8-Bark level until 15 months (specifically, at 66 weeks in Lieberman (1984)) and it stabilizes around the VLAM range somewhere between 1.5 year and 3 years.

[3] The Bark formula used in this paper, as in our previous papers and in de Boer (2010), is from Schroder, Atal, and Hall (1979): Bark = 7 ArgSh (Hz/650).

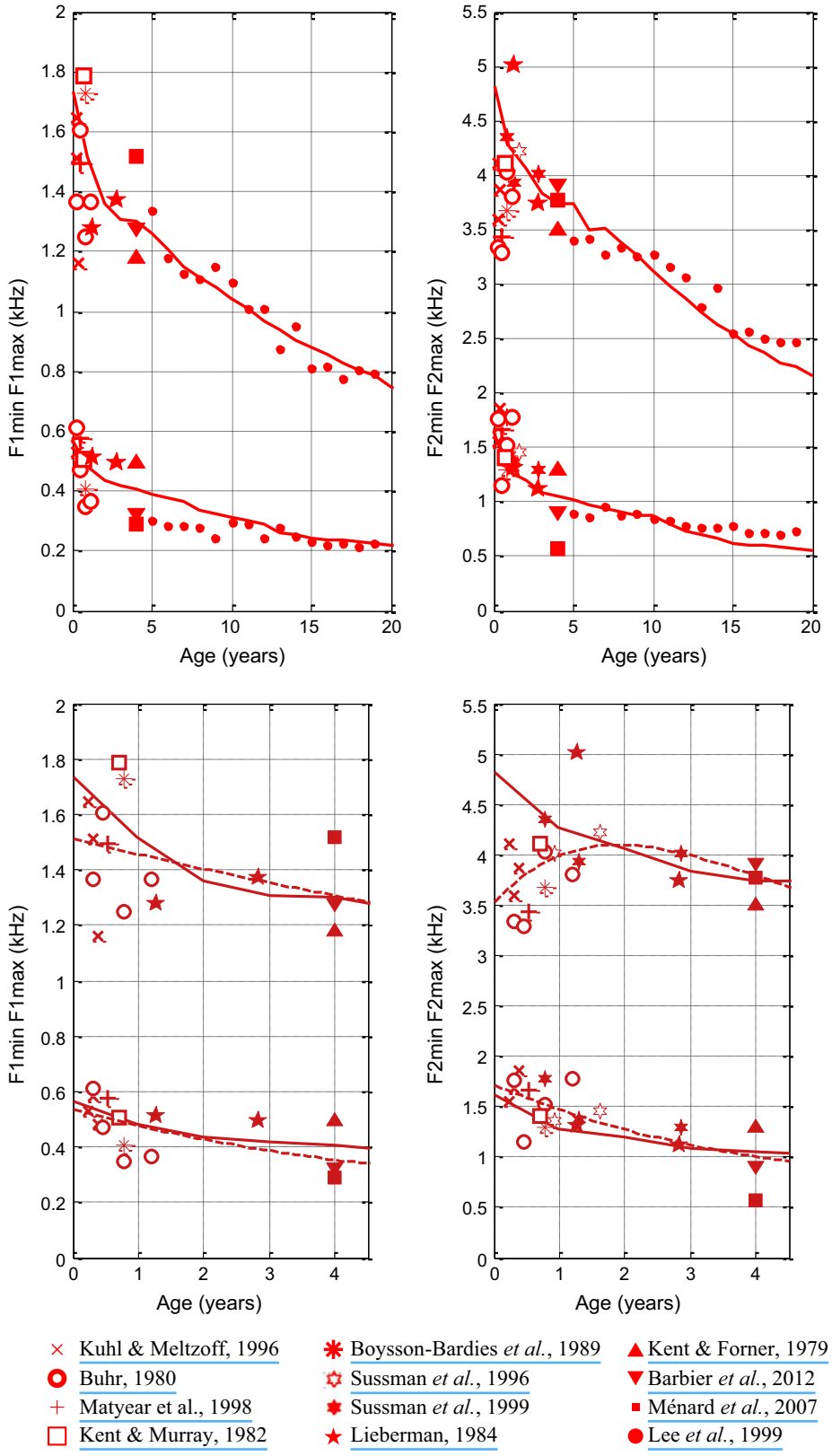

Fig. 3. Minimum and maximum values of F1 (left) and F2 (right) for vowel spaces from VLAM simulations across different ages and from data from in vivo studies. The top panels show the data from birth to 20 years, and the bottom panels show a close-up of the data from birth to 5 years. The VLAM values are drawn in solid lines. In the bottom panels, a polynomial fit to study data is added in dashed line. The values from studies are represented by symbols keyed to Table 1 and repeated here for convenience.

As we discuss in the next section, this outcome is perfectly in line with the Frame-Content Theory (MacNeilage, 1998; MacNeilage & Davis, 2000). According to FC predictions, formant exploration should occur primarily along F1 in the first babbling stages from 7 to 12 months, and then progressively extend along F2 in later stages.

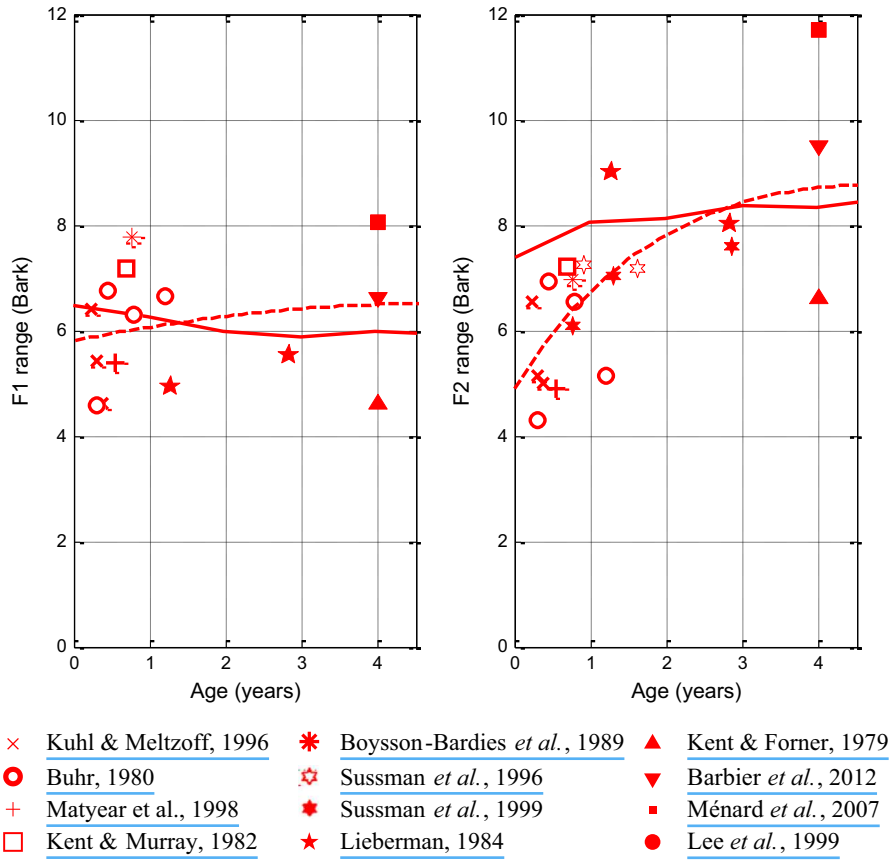

Fig. 4. Formant ranges for F1 (left) and F2 (right) for different ages from 0 to 5 years, from VLAM (solid line), and from the same studies as in Fig. 3. As in Fig. 3, a polynomial fit to study data is added as a dashed line. The values from studies are represented by symbols keyed to Table 1 and repeated here for convenience.

In summary, vowel formant exploration attested in actual child data is compatible with VLAM simulations for F1 from 6 months of age on, while the F2 range predicted by VLAM is probably not reached until 15 months. We consider it quite unlikely that this reflects an anatomical limitation, given the very short duration of this alleged limitation in acoustic potential – much shorter than Lieberman's claim (Section 3.1), that "children younger than age 6 years do not produce the full range of vowels predicted by the VLAM technique". On the other hand, this result is perfectly in line with the Frame-Content Theory, so we feel it is best understood as a lack of acoustic exploration in infants of that age, rather than an anatomical impossibility.

This conclusion is reinforced by the comparison of actual child data with de Boer's simulations. de Boer (2010) explores the F1–F2 acoustic range that can be reached by his simplified version of Mermelstein's model (1973). He sets the model at larynx depth values ranging from 6 to 16 cm, which correspond in his view to lengths of the vertical tube varying from 4 to 14 cm, and he fixes the horizontal tube at 8 cm. At each larynx depth, de Boer computes the possible F1 and F2 values, from which he extracts (but does not report) maximum and minimum values. Ranges (which he does report) are defined as the distance in Bark from the minimum to the maximum for the given formant. It appears that in this model, while the F1 range decreases smoothly as larynx depth increases, the F2 range peaks at a larynx depth which approximates an equal length of the horizontal and vertical parts of the vocal tract, while for smaller larynx depths, there is a dramatic reduction in of F2 range (Fig. 5).

The formant ranges reported by de Boer assume a fixed horizontal vocal tract length of 8 cm, which is typical of an adult-like vocal tract. To compare his range predictions with data from infants or children, a procedure for simulating the shorter vocal tracts typical of infants or children is required. Again, de Boer does not discuss any scaling, nor provide the minimum and maximum formant values from which it could be inferred, but rather only the ranges in Bark. Therefore, in the following evaluation, we assume that the formant range provided by de Boer for F2 is roughly independent of overall vocal tract length. This is reasonable, considering that scaling a vocal tract should proportionally scale its resonances (technically an inverse proportion: smaller resonators give higher resonances), and hence keep the same formant range if frequency is expressed in Bark, provided that the frequencies are above 1 kHz where the Bark scale operates logarithmically, as is typically the case for F2. The same does not hold for F1, in the lower frequencies where the Hz-to-Bark conversion is not logarithmic, so rescaling F1 values is impossible.

In Fig. 5 (adapted from de Boer's Fig. 3), we display F2 extents (i.e., ranges) he predicts for varying depth values. From these depth values, we can estimate the vertical/horizontal ratio LHI ("larynx height index", see Boë et al. (2002b)), computed from de Boer's specification of the horizontal dimension at 8 cm and the vertical dimension as larynx depth minus 2 cm. This allows us to indicate the human configurations corresponding to particular depth values: Boë et al. (2002b) found that typical LHI values reach 0.6 for infants, 0.8 for 4-years-olds, 0.95 for female adults and 1 for male adults, and these values are also compatible with data used by de Boer (2010) and Lieberman.

De Boer predicts 6 Bark as the maximum possible range, and that range is associated in his simulations with near-equal dimensions of the horizontal and vertical regions of the vocal tract (typical for a female adult according to de Boer (2010)). In fact, the F2 ranges predicted by de Boer's model for small vertical dimensions diverge dramatically from the exploration attested in the studies analyzed above of infant and child vocalizations. At these ages, children actually produce much larger F2 ranges than de Boer's model would allow based on his claim of "a dramatic reduction in the extent of the second resonance" at high larynx heights: recall that F2 ranges reach more than 6 Bark in many studies for infants below one year of age, and reach the 8-Bark VLAM value around 1.5 year. Replicating de Boer's simulations – which is not a simple matter considering the limited methodological detail in his paper – suggests a very simple cause to this strong departure from actual speech data: the lack of lips in his model.

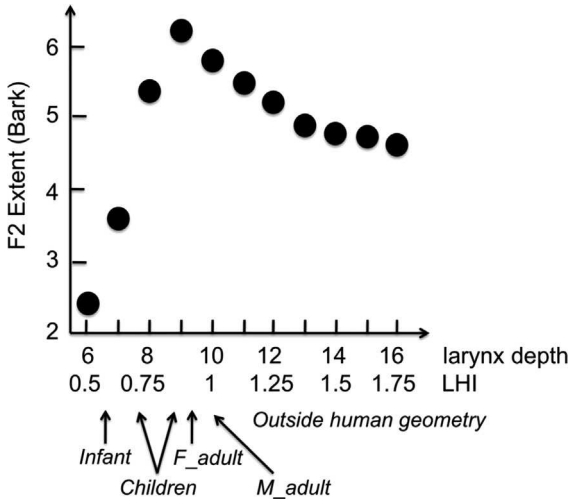

Fig. 5. F2 ranges from de Boer's simulations. Black circles and larynx depth values are directly from de Boer (2010), Fig. 3. We have added the vertical/horizontal ratio LHI computed from de Boer's specification of the horizontal dimension at 8 cm and the vertical dimension as larynx depth minus 2 cm. Typical human configurations associated with each LHI region (infants, children at increasing ages, and adults, noting F for female and M for male) are from Boë et al. (2002b).

Further simulations adding lips (Badin et al., in preparation) actually provide similar F2 ranges from his model and VLAM, regardless of larynx depth, both comparable with child data. Though such detailed simulations are outside the scope of this paper, this provides a likely explanation of the inaccuracy of de Boer's model relative to the present data, and leaves VLAM the better choice for simulating child formant characteristics.

2.4. Child data corroborates only VLAM predictions, not claims based on SVTh/SVTv ratio

The implication of this comparison of articulatory models with ground truth data from studies on infant and child vocalizations seems clear: VLAM, and only VLAM, adequately predicts formant values for human speakers at all ages from 15 months on. Specifically, all children are plainly capable as of 15 months of producing the full range of F1 and F2 values as modeled in VLAM, including point vowels. A full range of F1 values is produced even earlier, so the reduced F2 range before 15 months is quite likely a matter of control rather than anatomy, in agreement with the Frame-Content theory. This is so even though, quoting Lieberman, “The newborn tongue is flat and is positioned almost entirely in the oral cavity. It achieves its posterior rounded contour as it descends down into the pharynx, carrying the larynx down with it” (Section 2). He later continues, “These two anatomical factors – the human tongue's oral and pharyngeal proportions and shape – explain why only adult humans can produce [i], [u] and [a]” (Section 2). His assertion is refuted by observation of child data in vivo. The anatomical facts simply do not produce the acoustic consequences he claims. The “larynx hypothesis” and it is supposed consequences limiting the available acoustic space in infants and children under 6 years old are clearly refuted, hopefully once and for all.

Analysis of ground truth acoustic measurements of infants and children casts serious doubt on Lieberman's reasoning about vocal tract anatomy and tongue shape. As a consequence, we must be wary of his application of similar reasoning elsewhere. For instance, in one of his concluding comments in Section 7 on “Speech anatomy and the evolution of the human brain”, Lieberman considers the ratio in Neanderthals between horizontal and vertical components of the vocal tract, SVTh/SVTv. He states that “a 1.3 SVTh/SVTv ratio [is] in the range of 2-year old children”, a claim (corresponding to an LHI value around 0.76) that is entirely compatible with our own data (Fig. 5, Boë et al., 2002b; Boë et al., 2007a, 2007b). He claims, though, that this is a “problem precluding Neanderthals having a 1:1 SVTh/SVTv ratio necessary to produce quantal vowels”. We have shown here that despite the supposedly prohibitive 1.3 ratio, 2-year old children produce perfectly adequate quantal point vowels, as do even 15-month-olds, presumably with an even higher ratio. Since so much of Lieberman's argumentation about Neanderthal speech is dependent on the SVTh/SVTv ratio, the falsification by child data of his claims about its effect must be considered as a refutation of all his analogous deductions about its effect in Neanderthals, including those in his Section 7.

3. Constriction control as one component of a general principle of efficient modulation for speech communication

A strong limitation of Lieberman's contribution in our view is the rather restrictive way he addresses the problem of acoustic distinctiveness. Communication involves changing the status of the partner's brain, at distance. For this aim, the communication partners need a robust system of contrastive communication stimuli, which involves selecting a channel and a source and controlling efficient modulations of the source. Speech capitalizes on an acoustic source, which must be efficiently modulated by appropriate orofacial gestures. The continuing effect of efficiency in modulation can be seen in the development and optimization of contrasts between acoustic outputs, providing the basis for phonological distinctions in human languages (see Lindblom, 1984; Schwartz, Boë, Vallée, & Abry, 1997).

However, Lieberman's treatment of the evolution of speech is extraordinarily selective, since of the two main categories of speech sounds, he only mentions vowels, and ignores consonants. This is a bit like presenting a treatise on the diurnal cycle and leaving out either the day or the night. However, consonants and vowels are related such that they tend to alternate, and indeed, both result from the universal trend of the supralaryngeal VT to alternate between closed and open configurations, a superordinate regularity that has led to the concept “syllable”, which is also not mentioned by Lieberman. In the following, we propose to restate Lieberman's assumptions about the control of vocal tract constrictions inside a more general “efficient modulation” framework.

3.1. The frame-content theory: the first pillar of a theory of efficient modulation

It is striking that in his proposals about the evolution of speech motor control Lieberman never cites the Frame Content (FC) theory (MacNeilage, 1998; MacNeilage & Davis, 2000). However, this theory provides what can be considered as the *first pillar* of a principle of efficient modulation. FC theory explains

speech modulation as an exaptation (Gould & Vrba, 1982) from feeding behavior. Specifically, MacNeilage suggests (1998, p. 499) that: "syllabic 'frames' and segmental 'content' elements are separately controlled in the speech production process. The frames may derive from cycles of mandibular oscillation present in humans from babbling onset, which are responsible for the open–close alternation. These communication-related frames perhaps first evolved when the ingestion-related cyclicities of mandibular oscillation (associated with mastication [chewing] sucking and licking) took on communicative significance."

An essential part of FC, in our view, is that single mandible movements are easily and efficiently transformed into contrasting acoustic objects, consonants and vowels, thanks to nonlinear "quantal" articulatory-to-acoustic relationships (Stevens, 1972, 1989). Indeed, as the mandible moves smoothly through opening or closing gestures, the acoustic output switches abruptly rather than smoothly between states of vowel-like harmonic behavior, fricative-like turbulence, and stop-like closure (followed in the opening phase by a burst due to accumulated pressure). This is displayed in Fig. 6, which schematizes a VCV sequence, first in its smooth articulatory gesture, then in its quantal acoustico-auditory result, in order to show the non-linearity of the articulatory-acoustic transform. The figure shows the FC division of the mandible gesture into vowel and consonant poles, and also the division of the consonant pole into stops and fricatives. In essence, *manner of articulation* is derived as a direct outcome of the interaction of the FC and quantal theories. The modulation is efficient because stable acoustic outputs can be achieved through relatively large articulatory ranges of mandible movement, leaving fine adjustments to the more mobile articulators, the tongue and lips. Thus can the complex phenomena of consonants and vowels be controlled and sequenced into syllables, all emerging from simple articulatory gestures.

To summarize, in FC, simple mandibular oscillations are produced by the speaker and transformed by the laws of physics as described in quantal theory into rich and variegated sounds for the listener to provide different kinds of acoustic objects (plosives, fricatives, vowels), and this process constitutes a first dimension of efficient modulation of the vocal source.

Moreover, recent data suggesting that monkey lipsmacking develops like speech provide additional support for the theory, with Morrill, Paukner, Ferrari, and Ghazanfar (2012) concluding that "monkey lipsmacking and human babbling share a homologous developmental mechanism, lending strong empirical support to the idea that human speech evolved from the rhythmic facial expressions of our primate ancestors". Also of importance is the recent identification of endogenous cortical rhythms in humans that Giraud et al. (2007) associate with speech-related functions. Among other things they found a 3–6-Hz power band in the lower part of the motor cortex which, in their opinion, "offers a direct neural underpinning for the Frame-Content theory of speech that assumes that syllables are phylogenetically and ontogenetically determined by natural mandibular cycles occurring at about 4 Hz" (p. 1132). They consider that, overall, their findings "emphasize the role of common cortical oscillatory frequency bands for speech production and perception and thus provide a brain-based account for the phylogenetic emergence and shaping of speech from available neural substrates" (p. 1133). This suggests that the first dimension of modulation, phylogenetically, developmentally, and cortically conceptualized by FC, could also capitalize on specific cortical rhythms preexistent in the primate brain, and optimally exploited for communication through auditory-motor coupling mechanisms (Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012).

3.2. *Constriction control: the second pillar of efficient modulation*

Now that basic object categories (i.e., manners of articulation) are provided by the FC theory, finer distinctions require modulating the objects themselves by modifying their resonance frequencies. This is where Lieberman's question about "point vowels" and orofacial control becomes interesting in our view—though we of course conceive of things a bit differently.

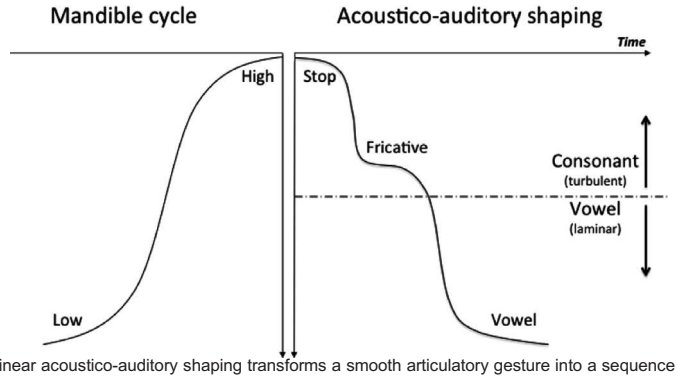

Fig. 6. Two views of a VCV sequence, showing how nonlinear acoustico-auditory shaping transforms a smooth articulatory gesture into a sequence of contrasting auditory units. The left side shows the smoothness in the articulatory point of view, in the closing phase of the mandible cycle, with the jaw moving from low to high. The right side shows the acoustico-auditory point of view, as the opening phase of the mandible cycle is transformed quantally into plateaux corresponding to the three acoustically distinct, auditorily contrasting major classes of speech sounds. The aerodynamics of interrupted or turbulent airflow (either brief or sustained) distinguishes stops and fricatives as consonantal from the laminar flow of vowels.

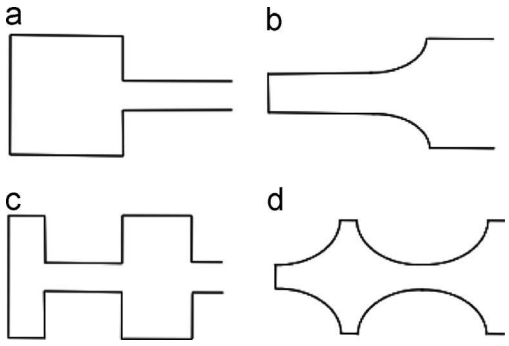

Fig. 7. Tube models with extreme resonances. Top: shapes with minimum (a) and maximum (b) first resonance F1. Bottom: tracts with minimum (c) and maximum (d) second resonance F2. Excitation ("glottis") is on the left and radiating end ("lips") is on the right.
Adapted from de Boer (2008).

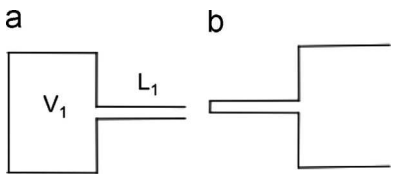

Fig. 8. Two-tube configurations minimizing (a) and maximizing (b) F1 in the Min–Max Area Model.

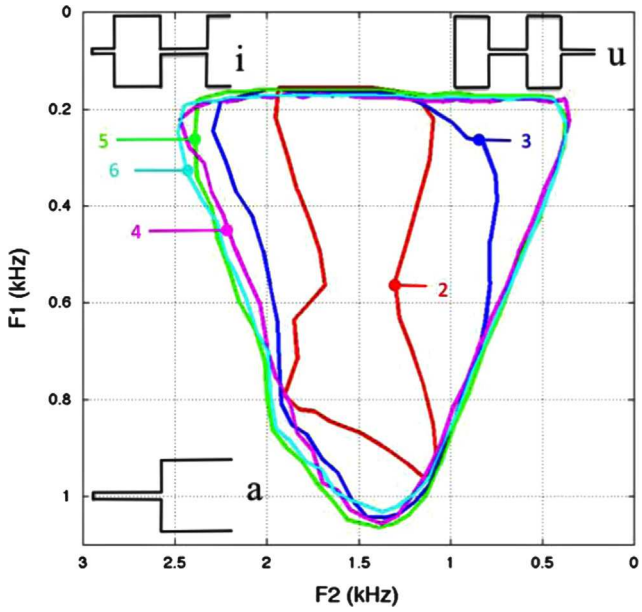

Fig. 9. Limits of the F1-F2 space for Min–Max Area Models configured with the number of tubes varying from 2 to 6. Border contours are labeled according to number of tubes. Configurations typical for [a] (two-tube, maximum F1), [i] (four-tube, minimum F1, maximum F2) and [u] (four-tube, minimum F1, minimum F2) are also displayed. Notice that with 2 or 3 tubes, the full range for F1, but not for F2, is already more or less achieved, while the maximum (F1, F2) space is essentially complete with 4 tubes.

3.2.1. Resonance modulation in the vocal tract: the "Min–Max Area Model"

Resonance frequencies must vary to produce acoustic modulations, and thereby instantiate phonemic contrasts, within what we have called "object categories," i.e., vowels, plosives and fricatives. The problem we are faced with next is determining what modifications of vocal tract shapes produce the greatest modulation of those resonance frequencies. A number of studies have explored and assessed "maximal acoustic spaces" using various kinds of vocal tract models and methodologies (e.g. Boë, Perrier, Guerin, & Schwartz, 1989; Carré, 2004).

In one very enlightening modeling study about the minimum and maximum resonance frequencies of a tube, de Boer (2008) showed that the minimum first resonance is obtained for a tube with a cavity of maximal area (and therefore volume as well) at the ("glottis") end where acoustic excitation is applied and minimal area at the radiating ("lip") end. He also showed that the maximum first resonance is obtained for the mirror configuration of a tube with minimal area at the excited end and maximal area at the radiating end (Fig. 7a, b). Configurations generating minimum vs. maximum resonance frequencies for higher-order formants can be constructed analogously by adding a pair of "min-max" tubes per resonance (e.g., minimum vs. maximum F2 in Fig. 7c, d).

The general principle underlying de Boer's simulation can be illustrated in a simplified model we call the "Min–Max Area Model," which we have developed for estimating the acoustic capacities of a tube of constant length L (i.e., 17.5 cm for a standard male vocal tract). It is based on a simplification of shapes generated by de Boer using sequences of tubes with any length but set to either minimum or maximum cross-sectional areas. The minimum area (A_{min}) is set to the lowest value still free from turbulence, which thus maintains a vocalic configuration with no shift to a fricative. The maximum area (A_{max}) is bounded by realistic vocal tract dimensions. The simplification is designed to enable manipulation of the individual resonances associated with each tube. This is accomplished through maximal decoupling between consecutive tubes, by maintaining a high ratio between maximum and minimum areas (A_{max}/A_{min}). We illustrate the functioning of this model on the configurations proposed by de Boer.

We start by looking at the two configurations of the Min–Max Area Model respectively minimizing and maximizing F1 in de Boer's computations. The former (see Fig. 8a) comprises two tubes, the first (i.e., glottis-side) with a maximum cross-section area and the second (lip-side) with a minimum cross-section area. This produces three basic resonance modes: (1) a "half-wave" resonance for the first tube, which is "closed" at both ends (here and elsewhere, we use "closed" to mean just sufficiently open physically to avoid turbulence); (2) a "half-wave" resonance for the second tube, which is open at both ends; and (3) an additional Helmholtz resonance for the whole configuration, which is characterized by the standard Helmholtz formula $F_H=(c/2\pi)\sqrt{(A_2/(l_2V_1))}$.

In this formula, $V_1=A_1l_1$ represents the volume of the first tube, A_2 is the area of the second tube, l_2 is its length, and c is the speed of sound in air. The Helmholtz resonance frequency F_H can thus be decreased by either increasing the volume V_1 (by increasing A_1 and/or l_1), or by increasing l_2, or by decreasing A_2. Since in the model the areas are fixed and $l_1+l_2=L$ must stay constant, the only manipulation possible is changing the proportions of l_1 and l_2 within L. Therefore, increasing l_2 to lower F_H must entail decreasing l_1, and thereby V_1 as well, which would then raise F_H, and vice versa. The effects on F_H of changing the proportions of l_1 & l_2 are contradictory, so a compromise must be found. It turns out that the optimal l_2 value

maximizing the product l_2V_1 is exactly $l_2=l_1=L/2$. Simply put, F1 is minimized when the vocal tract is divided in two equally long tubes. The resulting Helmholtz resonance approaches 0, the more so when the area value A_2 is set as low as allowed, to A_{min}.

The configuration maximizing F1 can be approximated by a mirror configuration made of two tubes, the first with a minimum cross-section area and the second with a maximum cross-section area (see Fig. 8b). Both tubes can be considered as closed at one end (glottis-side) and open at the other end (lip-side) and are thus characterized by "quarter-wave resonances". There is no enclosed volume, so no Helmholtz resonance. As before, with total length fixed, increasing the length of one tube necessitates decreasing the length of the other. Again the configuration with the highest F1 value is obtained with two tubes exactly the same lengths, equal to half the total length. For example, with a 17.5 cm total length and a speed of sound equal to 350 m/s, this provides quarter-wave fundamental modes at 1 kHz for each tube (assuming no radiation end correction at the lips), which sets the F1 value also to 1 kHz. Setting areas respectively at their minimum and maximum values also minimizes the coupling between the resonances of the two tubes and hence maximizes F1. Indeed, increasing the coupling would increase the separation of the two global resonances below and above 1 kHz, which would thus indirectly lower F1.

Therefore, in two-tube configurations, the F1 value extrema are determined by min–max and max–min configurations. Specifically, area values for each tube must be kept at their minimum or maximum values to lower the Helmholtz resonance frequencies (when present) and to reduce the coupling phenomena that would otherwise moderate the extreme resonance values in the global configurations.

A crucial point is that in the Min–Max Area Model, the maximum F1 value produced in the two-tube configuration in Fig. 8b cannot be increased by any three- or four-tube configuration. Adding more tubes would reduce the length of existing tubes and increase their resonance frequencies, but would also necessarily produce at least one cavity closed at both ends, thus introducing a Helmholtz resonance with a frequency approaching 0 Hz as in Fig. 8a and setting F1 to that value. As a consequence, the F1 range obtained with the two-tube configurations shown is substantially equivalent to that for any larger number of tubes.

For higher formants, the model can be easily generalized to any number of tubes. With four tubes, the extreme (F1, F2) values of the vowel triangle are obtained for three corresponding configurations as follows: (1) Four tubes of equal lengths configured max–min–max–min for two coupled Helmholtz resonances provide F1 and F2 values at their lowest possible values, so this is how the Min–Max Area Model simulates the [u] (see Fig. 9). (2) A min–max–min–max configuration with successive lengths (1/6, 1/3, 1/3, 1/6) provides a single Helmholtz resonance, while each of the four tubes resonates at the same value (since the two equally long outside tubes with quarter-wave resonance are half as long as the inside tubes with half-wave resonance). With the low Helmholtz for F1 and the other resonances near F2 maximum, this is the model's [i] (see Fig. 9). (3) The two-tube min–max configuration discussed above is retained for the model's [a], since it gives the appropriate high F1 and low F2 and since, as we have already seen, adding further tubes produces negligible improvement.

In the same way described above, adding tubes beyond a four-tube configuration results in shapes generating at least two Helmholtz resonances with values tending toward (0, 0), but it cannot extend the (F1, F2) space where its own two Helmholtz resonances have already established [u]. Hence, four tubes are sufficient to generate the classic vowel triangle in the (F1, F2) space.

3.2.2. *Constriction control is the key*

The following summarizes the basic properties of the Min–Max Area Model:

1. The F1 extrema are approximated by two-tube configurations.
2. The maximum (F1, F2) space is approximated by the set of four-tube configurations.
3. In general terms, the maximum (F1, F2, …, Fn) space is approximated by the set of 2n-tube configurations.
   Alternating minimum and maximum area tubes this way actually generates *n* minimum-area tubes, each of which constitutes a constriction inside the model. Thus we observe that a given number *n* of constrictions ensures generation of the corresponding maximum (F1, F2,.., F*n*) space.
   Fig. 9 confirms this conceptualization, by showing real simulations of the (F1, F2) space with n-tube models for various values of *n*. For each *n* value, *n*-tube configurations are explored as exhaustively as possible that is, by abandoning the "min–max area" constraint. As the number of tubes rises from 2 to 6, tube lengths and areas vary randomly, to explore the resulting (F1, F2) space generated. Specifically, tube lengths are drawn randomly under the sole constraint that the total length sum to 17.5 cm, and tube areas are drawn randomly from 13 values distributed logarithmically: $A=\{0.125, 0.177, 0.25, 0.353, 0.5, 0.707, 1, 1.414, 2, 2.83, 4, 5.66, 8\}$. Formant computation takes into account wall vibration and lip radiation (Badin & Fant, 1984).
   It appears clear that the (F1, F2) space is essentially complete and stable for $n\geq4$, in agreement with property (2).
   Then, applying this reasoning to the human vocal tract, we can propose some further properties.
4. The human vocal tract is typically a 2-constriction system, with one inside the vocal tract and the other at the lips, which should thereby be able to produce the maximum (F1, F2) space. In this space with [i] [a] [u] at the corners, point vowels can be reached provided that an adequate constriction can indeed be generated at any place and with any length inside the tract. Systems with more than the three point vowels (i.e., with other vowels intervening in the (F1, F2) space) can then be efficiently predicted by dispersion theories (e.g. Liljencrants & Lindblom, 1972; Lindblom, 1984; Schwartz et al., 1997).
5. On the other hand, the (F2, F3) space generated by the human vocal tract does NOT cover the whole range of possible values for a given acoustic tube, since in general a third constriction cannot be generated (except in rare cases such as dorsum-apex double constrictions inside the vocal tract). In a recent paper, though, we showed (Schwartz, Boë, Badin, & Sawallis, 2012) through VLAM simulations how human vocal tract anatomy and articulatory degrees of freedom shape the (F2, F3) space, and also how the FC framework allows us to explain the universality of [b d g] as an optimally contrastive set in this space.

Thus, the key to optimally exploiting acoustic capacities of a given acoustic tube turns out to be the *control of constrictions* inside the tube. The number of constrictions that a given articulatory system is able to create inside the vocal tract, and the precision of control over the constrictions' position and size, determine the resulting acoustic resonance modulation. We next discuss how these requirements might set limits on human vocal tract anatomy or control in the context of theories of speech emergence in phylogeny.

4. Back to the link between vocal tract anatomy and orofacial control: proposal for a new framework

At this stage, we have shown in Section 2 that children are able to produce a large range of formant variations by their first birthday, and that they reach the range predicted by VLAM simulations quite soon thereafter (actually before 1 year for F1, and between 15 and 18 months for F2). We have also shown in

Section 3 that, while FC provides a possible basis for the first dimension of efficient modulation – giving birth to syllables, constriction control provides a basis for the second dimension – giving birth to the modulation needed for contrasts within vowel and consonant systems in human languages.

In this discussion, the key question for phylogeny and ontogeny deals with control in relation with anatomy. FC is a theory for the emergence of one basic control mechanism for communication, exapted from ingestive mechanisms. This emerging control capitalizes on the anatomical and dynamical properties of the jaw, but it embeds these factors in a general framework about neurocognitive control of serial order (MacNeilage, 1998, 2011).

Lieberman's paper, like much of his published work on the larynx hypothesis, deals with constriction control. While larynx height does not directly control formant range, vocal tract anatomy could play a role in the ease of vocal tract control (e.g. Perkell, 1996). In this section, we address the question, implicit in Lieberman's work, of possible anatomical requisites for constriction control.

4.1. Potential anatomical constraints on phonetic control

Lieberman's answer has focused for over 30 years on larynx position. However, it has become increasingly clear over those decades that nothing convincing has been proposed yet about larynx position. We have shown previously (Boë et al., 2002a, 2002b, 2007a, 2007b) and expand further in this paper on the counter argument that a number of claims alleged to validate the role of larynx descent in vocal tract control are erroneous. Some such errors involve inaccurate claims about larynx position in Neanderthals, others are related to acoustic simulations. We do not replay these arguments here; readers should come to their own conclusions. We would point out that that arguments involving animal vocalizations (Fitch, 2010, Chapter 8) reinforce our claim that larynx height is peripheral to the understanding of anatomical requisites for communication.

The question of vocal tract shape is more relevant. The only solid evidence at our disposal comes from comparative morphological analyses and 3D modeling of the tongue musculature of humans and chimpanzees by Takemoto (2001, 2008), who concludes: “Applying the muscular hydrostat theory to the external shape of the tongue suggests that the primary actions of the chimpanzee tongue are protrusion and retrusion [sic], whereas the human tongue can be deformed in the oral cavity with a high degree of freedom. It is hypothesized that the evolution of the external shape of the tongue is one of the factors that led to the development of human speech” (Takemoto, 2008, p. 966). However, Takemoto's comment is qualitative and holistic rather than an attempt at a thoroughly reasoned answer based on actual speech data.

To shed some light on this question, let us consider some data on newborns. As shown in Fig. 10, tongue fibers for newborns are quite similar to those of adults. This similarity allows the comparison of the intrinsic muscle fibers of the tongue for a 39-weeks fetus (near full term) with those of an adult. Even though the tongue occupies a comparatively larger volume in an infant's oral cavity, the configuration of the muscle fibers makes it likely that adequate coordination between the mandible and the tongue would enable infants to produce constrictions all along the vocal tract. This hypothesis is confirmed by the acoustic data of Section 2 showing that before 12 months for F1 and soon thereafter for F2, infants produce extreme acoustic configurations congruent with VLAM simulations, which thereby indicates that all requisite constrictions are articulatorily achievable. Of course, this does not mean that constriction *control* is achieved. The infant's uncontrolled articulatory exploration does suggest however that the acoustic capacities to be generated by the constriction control are essentially complete, and that control is the final missing step.

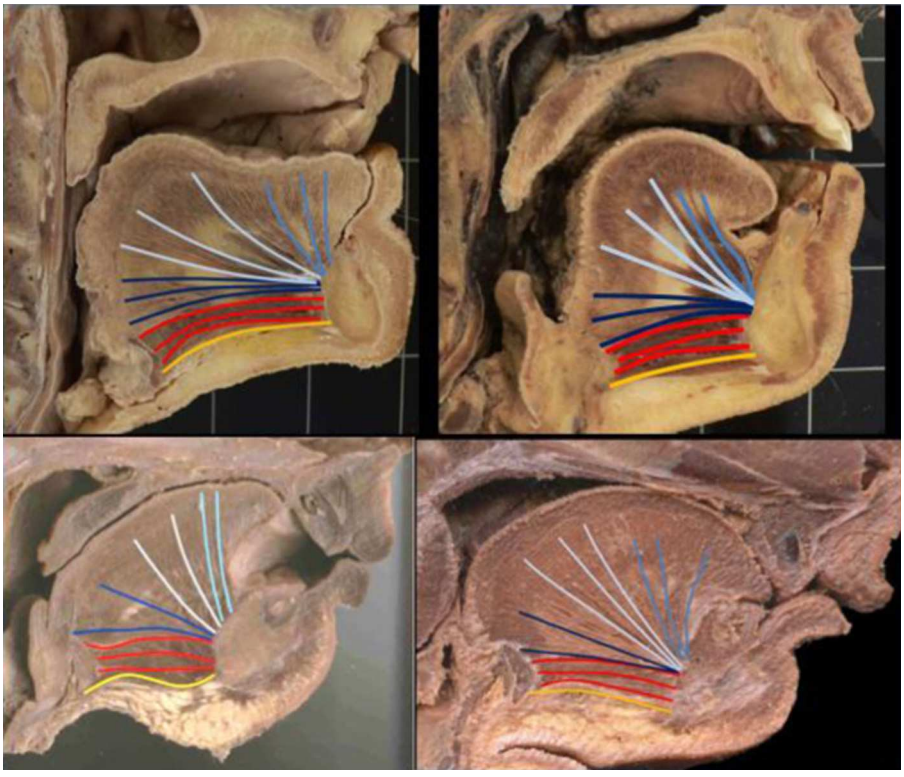

Fig. 10. The intrinsic muscle fibers of the tongue. On the bottom row, two fetuses shortly preterm (31–32 weeks amenorrhea on the left, 35–36 weeks amenorrhea on the right), and on the top row two adults. (Illustrations courtesy of Prof. Guillaume Captier, of the Laboratoire d’anatomie de la Faculté de Médecine de Montpellier).

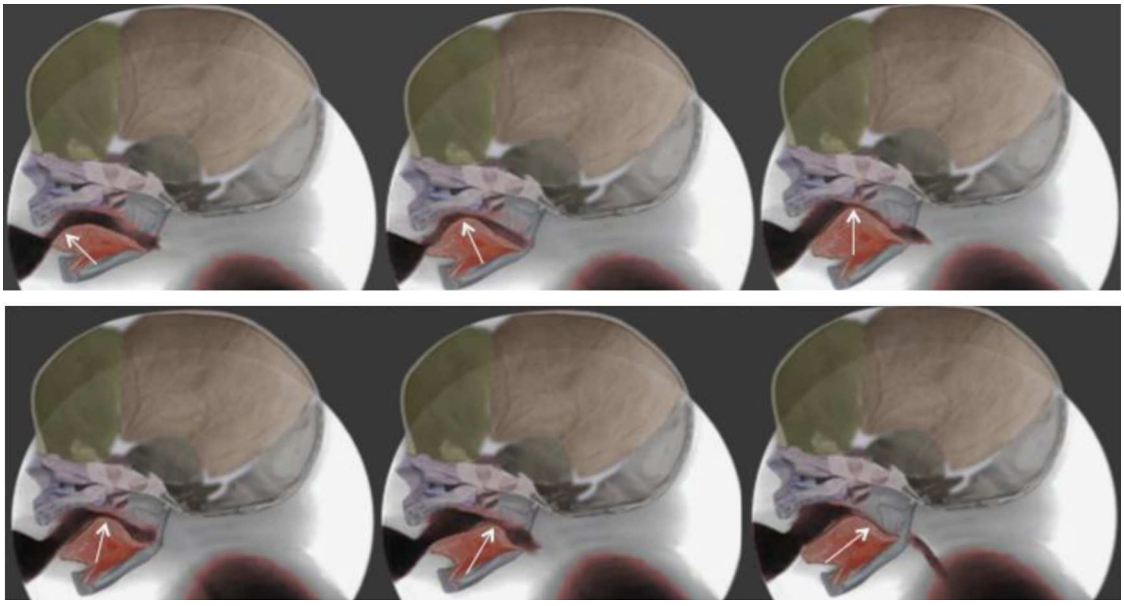

Fig. 11. Swallowing in the newborn. These images were selected – and colorized for clarity purposes – by Nicolas Kielwasser (OsteoGraph) from cineradiography of a newborn doing a barium swallow test (courtesy Dr. Danièle Robert-Rochet, Dept. of Pediatric Otorhinolaryngology, Centre Hospitalier Universitaire Marseille la Timone – the original film can be viewed at http://www.orl-marseille.com/soins/deglut-anat.htm). First, the bone structure was reconstructed (Boë et al., 2008) from Fenart's 3D data base (Fenart, 2003). Then the original gray contrast of the barium liquid was automatically extracted (Adobe After Effect software) and colored in black and red (giving as a side effect a red tinge to lung tissue which had the same level of gray). Next, for the hard structures – the skull and mandible, initial anatomical drawings of skull and mandible (Boë et al., 2008) were fitted to one specific image extracted from the film by non-linear deformation of landmarks identified on both the drawing and the image. Then, for each further image of the film, these two references were aligned to the moving skull and mandible by translation and rotation, and then superimposed back on the target images. The tongue was processed similarly, except that, as a deformable organ, its initial anatomical drawing was fitted using the contours as rendered more visible by the barium contrast agent. Although the nipple blocks full closure anteriorly, the newborn clearly displaces an occlusion from lips through palate and velum down to pharynx (as indicated by the white arrow), thereby exploring the full range of possible vocal tract constrictions, including the three universal stop places (Schwartz et al., 2012). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Expanding frame-content: A new proposal for A possible precursor of constriction control

So, how could control have evolved? We would like to introduce a new hypothesis, which expands in some sense on the fundamental exaptation hypothesis of the Frame-Content Theory, that speech evolved from the ingestion-related cycles of mandibular oscillation associated with mastication. Ingestion mechanisms naturally involve swallowing as well as chewing, and we propose that swallowing provided the basis for a second exaptation process as a precursor to constriction control.

Swallowing (of amniotic fluid) is first seen in human embryos at about 10–12 weeks of gestation. By birth a fetus is swallowing almost half the remaining fluid each day (Wall & Smith, 2001). To achieve this vital physiological function, the newborn displaces one or two areas of contact between tongue and the opposing oral and pharyngeal cavity walls from front of the mouth back to pharynx passing by hard palate and velum (see Fig. 11). Of course, the nature of contact between tongue and palate is quite different in speech and in swallowing (where actions are forceful to prevent dysphagia, Konaka et al., 2010), but the gestures are similar. Evidence for the similarity between swallowing and the articulatory gestures of speech was first provided by Hiiemae et al. (2002, 2003). In a recent study (Serrurier, Badin, Barney, Boë, & Savariaux, 2012), two articulatory models were elaborated, respectively for ingestion and speech. The mastication-swallowing model is more general than that for speech production in that it reproduces the gestures of both speech and ingestion with good precision (to about 1.5 mm).

The control of these actions is of course quite different. Swallowing is largely a reflex behavior, though its onset may be voluntary, while speech production is completely under voluntary control. However, fMRI studies comparing swallowing and voluntary tongue elevation display similar cortical networks. Both involve the left lateral pericentral cortex, the anterior parietal cortex, the anterior cingulate cortex and the adjacent supplementary motor area. The overlap of these networks suggests that these brain regions mediate processes shared by swallowing and tongue movement (Martin et al., 2004; Grabski et al., 2011).

Ingestion in mammals is distinguished from that of other vertebrates not only by mastication, but also by suckling, and by complexity of food transport and swallowing, both in terms of mature behavior (Wall & Smith, 2001) and of developmental processes (Campbell-Malone, German, Crompton, & Thexton, 2011). Associated evolutionary processes from non-anthropoid mammals to macaques have been described by Franks, Crompton, and German (1984). We propose that swallowing provides the necessary second dimension of the Frame-Content Theory. Cyclic mandibular gestures derived from mastication provide a precursor of vertical control and underlie the emergence of manner of articulation. Swallowing provides a precursor of horizontal control and underlies the emergence of place of articulation. Altogether, speech would derive, during both phylogeny and ontogeny, from a specialization of mastication and swallowing for ingestion, by capitalizing on a progressive reorganization of their control.

5. Conclusion

We willingly acknowledge Lieberman's role in raising and pursuing the question of the adaptation of human anatomy to increase capacities for speech modulation and the efficacy of speech control. However, more than 40 years after his first paper on the “larynx assumption” (Lieberman & Crelin, 1971), it is now entirely clear that Lieberman's laryngeal descent hypothesis is incorrect. It is neither anatomically valid (e.g., Boë et al., 2002b; Fitch, 2010), nor acoustically accurate (e.g., Boë et al., 2002b, 2007a, 2007b).

In this paper, we advance several proposals to contribute to this important question. Anchoring the whole theme within a framework of the emergence of communication, we have proposed a "principle of efficient modulation". We have related efficient modulation to the Frame-Content Theory (MacNeilage & Davis, 2000; MacNeilage, 1998). We have shown that the main articulatory question is constriction location and control rather than larynx position. We have also proposed an expansion of the Frame-Content Theory whereby, with mastication still providing the "first pillar," swallowing would provide a precursor of the "second pillar" of the "principle of efficient modulation." In our view, mandible cycles provide the evolutionary bootstrap for control of the first (i.e., *vertical*) dimension, and swallowing introduces the bootstrap for control of the second (*horizontal*) dimension.

Of course, much remains to be done to evaluate this theory. Whatever the future of this "swallowing" assumption, Philip Lieberman has made a great contribution since 1971 in stressing the importance of the question of the acoustic abilities of the vocal tract. Despite the failure of the larynx hypothesis, his persistent pursuit of the topic has long animated research in this area and must be acknowledged.

Acknowledgments

The authors are grateful to: Dr. Danièle Robert, for cineradiography; Houri Vorperian for the child MRI images; Marie-Josèphe Deshayes and Djilali Hadjouis for the X-Rays; Philippe Mennecier and Jean-Louis Heim for the Neanderthal fossil; Nicolas Kielwasser for the swallowing graphics; and Louis Bonder, Bernard Guérin and Frédéric Berthommier for useful discussions and information on acoustic spaces. This project was funded by ANR SkullSpeech.

References

Badin, P., Boë, L. J., Sawallis, T., & Schwartz, J. L. (in preparation). Low larynx or long lips?.
Badin, P., & Fant, G. (1984). Notes on vocal tract computation. *STL-QPSR*, 53–108.
Barbier, G., Perrier, P., Ménard, L., & Boë, L. -J. (2012). Contrôle lingual en production de parole chez l'enfant de 4 ans: une méthodologie associant étude articulatoire et modélisation biomécanique. In L. Besacier, B. Lecouteux, & G. Sérasset (Eds.), *Proceedings of the 29e Journées d'Études sur la Parole* (pp. 393–400). Vol. 1, Grenoble, France.
Boë, L. –J. (1999). Vowel spaces of newly-born infants and adults consequences for ontogenesis and phylogenesis. In *Proceedings of the 14th international congress phonetic sciences* (pp. 2501–2504). San Francisco.
Boë, L.-J., Captier, G., Granat, J., Deshayes, M. -J., Heim, J. -L., Birkholz, P., Badin, P., Kielwasser, N., & Sawallis, T. (2008). Skull and vocal tract growth from fetus to 2 years. In R. Sock, S. Fuchs, & Y. Laprie (Eds.) *Proceedings of the Eighth International Seminar on Speech Production, ISSP8* (pp. 157–160). Strasbourg, France.
Boë, L.-J., Granat, J., Badin, P., Autesserre, D., Pochic, D., Zga, N., Henrich, N., & Ménard, L. (2007a). Skull and vocal tract growth: From newborn to adult. In J. Trouvain, & W.J. Barry (Eds.), *Proceedings of 16th international congress of phonetic sciences* (pp. 533–536). Saarbrücken, Germany.
Boë, L. -J, Heim, J. C., & Abry, C. (2002a). The size of the pharynx: An irrelevant parameter for speech emergence and acquisition. In *Proceedings of fourth international conference on the evolution of language*. Harvard University, March 23–28.
Boë, L.-J., Heim, J. L., Honda, K., & Maeda, S. (2002b). The potential of Neandertal vowel space was as large as that of modern humans. *Journal of Phonetics*, *30*, 465–484.
Boë, L.-J., Heim, J.-L., Honda, K., Maeda, S., Badin, P., & Abry, C. (2007b). The vocal tract of primates, newborn human and Neanderthal: Acoustic capabilities and consequences for the debate on the origin of language. A reply to Philip Lieberman (2007a). *Journal of Phonetics*, *35*, 564–581.
Boë L. -J. & Maeda S. (1997). Modélisation de la croissance du conduit vocal. Espace vocalique des nouveau-nés et des adultes. Conséquences pour l'ontogenèse et la phylogenèse. In *Journées d'Études Linguistiques: «La voyelle dans tous ses états»*, Nantes, pp. (98–105).
Boë, L.-J., Ménard, L., Serkhane, J., Birkholz, P., Kröger, B. J., Badin, P., Captier, G., Canault, M., & Kielwasser, N. (2008). La croissance de l'instrument vocal: Contrôle, modélisation, potentialités acoustiques et conséquences perceptives. *Revue Française de Linguistique Appliquée*, *13*, 59–80.
Boë, L.-J., Perrier, P., Guerin, B., & Schwartz, J.-L. (1989). Maximal vowel space. *Proceedings of Eurospeech*, *89*(2), 281–284.
Boysson-Bardies de, B., Halle, P., Sagart, L., & Durand, C. (1989). A cross-linguistic investigation of vowel formant in babbling. *Journal of Child Language*, *16*, 1–17.
Buhr, R. D. (1980). The emergence of vowels in an infant. *Journal of Speech and Hearing Research*, *23*, 73–94.
Campbell-Malone, R., German, R. Z., Crompton, A. W., & Thexton, A. J. (2011). Ontogenetic changes in mammalian feeding: insights from electromyographic data. *Integrative and Comparative Biology*, http://dx.doi.org/10.1093/icb/icr026.
Captier, G., Boë, L.-J., & Barbier, G. (2010). Anatomie et croissance du conduit vocal du fœtus à l'enfant de 5 ans. *Biométrie Humaine et Anthropologie*, *28*, 65–73.
Carré, R. (2004). From an acoustic tube to speech production. *Speech Communication*, *42*, 227–240.
de Boer, B. (2008). Acoustic tubes with maximal and minimal resonance frequencies. In *Proceedings of Acoustics'08*, (pp. 5063–5068).
de Boer, B. (2010). Modelling vocal anatomy's significant effect on speech. *Journal of Evolutionary Psychology*, *8*, 351–366.
de Boer, B., & Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive behavior*, *18*, 36–47.
Fant, G. (1960). *Acoustic theory of speech production*. Mouton: The Hague.
Fenart, R. (2003). Crâniographie vestibulaire. Analyse morphométrique positionnelle. *Biométrie Humaine et Anthropologie*, *21*, 231–284.
Fitch, W. T. (2010). *The evolution of language*. New York: Cambridge University Press.
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.
Franks, H. A., Crompton, A. W., & German, R. Z. (1984). Mechanism of intraoral transport in macaques. *American Journal of Physical Anthropology*, *65*, 275–282.
Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech production and perception. *Neuron*, *56*, 1127–1134.
Goldstein, U. G. (1980). *An articulatory model for the vocal tracts of growing children*. Cambridge, MA: MIT [Thesis of Doctor of Science].
Gould, S. J., & Vrba, E. S. (1982). Exaptation; a missing term in the science of form. *Paleobiology*, *81*, 4–15.
Grabski, K., Lamalle, L., Vilain, C., Schwartz, J.-L., Vallée, N., Tropres, I., Baciu, M, Le Bas, J. F., & Sato, M. (2011). Functional MRI assessment of orofacial articulators: neural correlates of lip, jaw, larynx and tongue movements. *Human Brain Mapping*, *33*, 2306–2321.
Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in Cognitive Science*, *16*, 114–121.
Hiiemae, K. M., & Palmer, J. B. (2003). Tongue Movements in feeding and speech. *Critical Reviews in Oral Biology and Medicine*, *14*, 413–429.
Hiiemae, K. M., Palmer, J. B., Medicis, S. W., Hegener, J., Jackson, B. S., & Lieberman, D. E. (2002). Hyoid and tongue surface movements in speaking and eating. *Archives of Oral Biology*, *47*, 11–27.
Kent, R. D., & Forner, L. L. (1979). Developmental study of vowel formant frequencies in an imitation task. *Journal of the Acoustical Society of America*, *65*, 208–217.
Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances. *Journal of the Acoustical Society of America*, *72*, 353–365.
Konaka, K., Kondo, J., Tamine, K., Hori, K., Ono, T., Maeda, Y., Sakoda, S., & Naritomi, H. (2010). Relationship between tongue pressure and dysphagia in stroke patients. *European Neurology*, *64*, 101–107.
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.
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. (1984). *The biology and evolution of language*. Cambridge, MA: Harvard University Press.
Lieberman, P. (2012). Vocal tract anatomy and the neural bases of talking. *Journal of Phonetics*, *40*, 608–622.
Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge, UK: Cambridge University Press.
Lieberman, P., & Crelin, E. S. (1971). On the speech of the Neanderthal man. *Linguistic Inquiry*, *2*(2), 203–222.
Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, *48*, 839–862.
Lindblom, B. (1984). Can the models of evolutionary biology be applied to phonetic problems? In *Proceedings of tenth international congress phonetic sciences* (pp. 67–81).
MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*, 499–511.
MacNeilage, P. F. (2011). Lashley's serial order problem and the acquisition/evolution of speech. *Cognitive Critique*, *3*, 49–84.

MacNeilage, P. F., & Davis, B. L. (2000). Deriving speech from non-speech: A view from ontogeny. *Phonetica*, *57*, 284–296.
Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In: W. Hardcastle, & A. et Marchal (Eds.), *Speech production and speech modeling* 131–149. Dordrecht, Netherlands: Kluwer Academic.
Martin, R. E., MacIntosh, B. J., Smith, R. C., Barr, A. M., Stevens, T. K., Gati, J. S., & Menon, R. S. (2004). Cerebral areas processing swallowing and tongue movement are overlapping but distinct: a functional magnetic resonance imaging study. *Journal of Neurophysiology*, *92*, 428–443.
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.
Ménard, L. (2002). *Production et perception des voyelles au cours de la croissance du conduit vocal: variabilité, invariance et normalisation*. Grenoble III: Université Stendhal [Thèse de doctorat].
Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2004). Role of vocal tract morphology in speech development: Perceptual targets and sensori-motor maps for synthesized French vowels from birth to adulthood. *Journal of Speech, Language and Hearing Research*, *47*, 1059–1080.
Ménard, L., Schwartz, J.-L., Boë, L.-J., & Aubin, J. (2007). Articulatory-acoustic relationships during vocal tract growth for French vowels: Analysis of real data and simulations with an articulatory model. *Journal of Phonetics*, *35*, 1–19.
Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, *53*(4), 1070–1082.
Morrill, R. J., Paukner, A., Ferrari, P. F., & Ghazanfar, A. A. (2012). Monkey lipsmacking develops like speech. *Developmental Science*, *15*, 557–568.
Perkell, J. S. (1996). Properties of the tongue help to define vowel categories: Hypotheses based on physiologically-oriented modeling. *Journal of Phonetics*, *24*, 3–22.
Schroder, 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. Öhman (Eds.), *Frontiers of speech communication research* 217–229. London: Academic Press.
Schwartz, J.-L., Boë, L.-J., Badin, P., & Sawallis, R. T. (2012). Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial-coronal-velar stop series. *Journal of Phonetics*, *40*, 20–36.
Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion–focalization theory of vowel systems. *Journal of Phonetics*, *25*, 255–286.
Serkhane, J., Schwartz, J. -L., & Bessière, P. (2003). Simulating vocal imitation in infants, using a Growth articulatory model and speech robotics. In *Proceedings of the 15th international congress of phonetic sciences*, Barcelona.
Serkhane, J., Schwartz, J.-L., Boë, L.-J., Davis, B. L., & Matyear, C. (2007). Infants' vocalizations analyzed with an articulatory model: A preliminary report. *Journal of Phonetics*, *35*, 321–340.
Serrurier, A., Badin, P., Barney, A., Boë, L.-J., & Savariaux, C. (2012). The tongue in speech and feeding: Comparative articulatory modelling. *Journal of Phonetics*, *40*, 745–763.
Stevens, K. N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In: E. E. Davis, E. E., Jr., & P. B. Denes (Eds.), *Human communication: A unified view* New-York: McGraw-Hill.
Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3–45.
Sussman, H. M., Duder, C., Dalso, E., & Cacciatore, A. (1999). An acoustic analysis of the development of CV coordination: A case study. *Journal of Speech Language and Hearing Research*, *42*, 1080–1096.
Sussman, H. M., Minifie, F. D., Buder, E. H., Stoel-Gammon, & Smith, J. (1996). Consonant-vowel interdependencies in babbling and early words: Preliminary examination of a locus equation approach. *Journal of Speech and Hearing Research*, *39*, 424–433.
Takemoto, H. (2001). Morphological analyses of the human tongue musculature for three-dimensional modeling. *Journal of Speech, Language and Hearing Research*, *44*, 95–107.
Takemoto, H. (2008). Morphological analyses and 3D modeling of the tongue musculature of the chimpanzee (Pan troglodytes). *American Journal of Primatology*, *70*, 966–975.
Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, *50*, 1510–1545.
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 early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America*, *117*(1), 338–350.
Wall, C., & Smith, K. K. (2001). *Ingestion in mammals.. Encyclopedia of Life Sciences. (1–6), London, UK:* Macmillan publishers.

Louis-Jean Boë, Pierre Badin
GIPSA-lab Speech and Cognition Department (ICP), UMR 5216 CNRS – Grenoble University,
961 rue de la Houille Blanche—Domaine universitaire —BP 46, 38402 Saint Martin d'Hères Cedex, France

Lucie Ménard
Département de linguistique, Laboratoire de phonétique, CHU Ste-Justine, Center for Research on Brain, Language, and Music,
Université du Québec à Montréal, Montréal, Québec, Canada H3C 3P8

Guillaume Captier
IURC—Laboratoire de Biostatistique d'Épidémiologie et de Recherche Clinique 641 avenue du Doyen G,
Giraud 34093 Montpellier Cedex 5, France

Barbara Davis
Department of Communication Sciences and Disorders, University of Texas, Austin, TX 78712, USA

Peter MacNeilage
Department of Psychology, University of Texas, Austin, TX 78712, USA

Thomas R. Sawallis
GIPSA-lab Speech and Cognition Department (ICP), UMR 5216 CNRS – Grenoble University,
961 rue de la Houille Blanche—Domaine universitaire —BP 46, 38402 Saint Martin d'Hères Cedex, France
New College, University of Alabama, Tuscaloosa, AL 35487, USA

Jean-Luc Schwartz *
GIPSA-lab Speech and Cognition Department (ICP), UMR 5216 CNRS – Grenoble University,
961 rue de la Houille Blanche—Domaine universitaire —BP 46, 38402 Saint Martin d'Hères Cedex, France
E-mail address: jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

* Corresponding author. Tel.: +33 4 76 57 47 12; fax: +33 4 76 57 47 10.