Title: Computer models of vocal tract evolution: An overview and critique

Running title: Models of Vocal Tract Evolution

Authors:

Bart de Boer

University of Amsterdam

Spuistraat 210, 1012 VT, Amsterdam, the Netherlands

Tel: +31 20 525 2182

Fax: +31 20 525 3021

b.g.deboer@uva.nl

and

W. Tecumseh Fitch

University of St. Andrews, St. Andrews, KY16 9JP, Scotland

Abstract

Human speech has been investigated with computer models since the invention of digital computers, and models of the evolution of speech first appeared in the late 1960's and early 1970's. Speech science and computer models have a long shared history because speech is a physical signal and can be modeled accurately. This paper gives a brief overview of the use of computer models in the study of the evolution of the vocal tract. We also present a critical case study of one model that has been used to study the vocal abilities of Neanderthals. We argue that this study contains subtle but fatal flaws which invalidate the conclusions drawn from the model, illustrating the dangers of applying computer models outside the area for which they have been developed. Future models need to make use of a broader database of anatomical and physiological data from other animals, especially nonhuman primates, to understand the path leading to modern *Homo sapiens*.

1. Introduction

The use of computer models in the study of speech has a long tradition, (Schroeder, 1993) and several researchers have employed such models in studies of the evolution of speech (Lieberman & Crelin 1971, Boë, Heim, Honda, & Maeda, 2002; de Boer, 2002). This is understandable as the processes underlying speech are mechanical and acoustic, and therefore lend themselves to physics-based modeling better than most aspects of language. It is possible to study the vocal tract and the vocal cords and all other anatomical structures relevant for the production of speech directly in humans and other species. It is also possible, using X-ray films and ultrasound, to study how the different articulators are used dynamically, in living humans and in other species (e.g., Perkell 1969, Fitch, 2000b). There are even some fossils (Arensburg *et al.*, 1989; MacLarnon & Hewitt, 1999; Martínez *et al.*, 2008, Alemseged et al. 2006) that are potentially relevant for the study of the evolution of speech. Researchers interested in the evolution of other components of language do not have such direct access to relevant data. Thus, the topic of using computer models to increase understanding of speech evolution is of considerable interest to language evolution researchers.

The major initial engineering effort to produce working speech synthesizers and speech processing systems was meant to improve techniques for transmitting voice over telephone lines, but this work also led to better understanding of how speech works in humans. We will not give a complete historical overview of all methods to synthesize speech. For a more complete overview, see (Klatt, 1987). Speech models, originally designed by engineers to reproduce modern human speech, have also been used to investigate the articulatory abilities of human ancestors, and of nonhuman species. This paper presents an overview of the history of modeling the human vocal tract and applying such models to modeling ancestral vocal tracts and vocal tracts of apes and monkeys, as well as a more detailed case study and critique of one prominent recent model. Our aim is to illustrate that computer modeling and phonetics have successfully interacted over many years to increase our understanding of speech and its evolution. However, such interaction is not without its potential problems and pitfalls. We will show that one must

pay careful attention when applying computer models originally developed for modeling modern human speech to animals, infants, or fossil humans. Subtle assumptions built into a model may invalidate its use outside its original intended domain. Such assumptions, present in any model, may only become apparent when using the model outside the area for which it was originally designed. It is hoped that the lessons from this case study will be transferable to other aspects of the evolution of language.

2. Historical Overview: Models of the Vocal Tract

Even before the advent of modern digital computers, vocal tract models were conceived and built. Chiba and Kajiyama (1942) were probably the first to propose electrical analogs of the vocal tract. Although these analogs were initially meant to facilitate mathematical analysis, they were also made into complete tunable models of the vocal tract (Dunn, 1950; Rosen, 1958; Stevens, Kasowski, & Fant, 1953). A famous example of these analog computers was the Swedish LEA (electrical line analogue). It was used (as well as Sweden's first digital computers BARK and BESK) in the development of the source-filter theory by Fant (1960). Modeling was used to test whether the proposed theories could in fact reproduce human speech sounds (the focus was mainly on vowels) given human data on articulation. Because the necessary calculations were complex, researchers became aware very early of the necessity of using computers (e. g. van de Berg, 1955, section II). From these modeling studies it became clear that the source of acoustic energy (the larynx in voiced speech) could essentially be studied separately from the upper vocal tract, which served to filter and thus modulate the source signal. It also became clear that a reasonably simple model was sufficient to describe the acoustic aspects of vowel production.

This research showed that, in order to calculate the acoustic properties of a vocal tract quite accurately, all that is needed is the *area function* (Fant 1960). The area function gives the cross sectional area at each point along the length of the vocal tract. The exact shape of the cross section is unimportant for the frequencies that are relevant for speech. Given an area function, there are several (more or less equivalent) ways of calculating the acoustic response. Usually, acoustic responses are represented using

formants, the resonant frequencies of the vocal tract. When vowels are described, the first three formants already give a good impression of the acoustic signal.

For complicated vocal tract shapes, solving the mathematical equations in closed form becomes intractable without computers. Therefore, very simple mathematical approximations have been developed to aid insight into basic issues of vocal acoustics (e.g., Lieberman & Blumstein 1988; Titze 1994). These heuristic approximations are usually based on two to four cylindrical tubes of given length and diameter. Although such models are very useful for investigating the effects of changes in vocal tract shape, they should not be mistaken for realistic models of the vocal tract. A real vocal tract cannot make all signals that can be produced by a system consisting of two cylindrical tubes, nor can a system consisting of two cylindrical tubes produce all signals that can be produced by a real vocal tract. The original investigators of such models (Chiba & Kajiyama, 1942; Dunn, 1950; Fant, 1960; Flanagan, 1965) were well aware of this distinction, but more recent researchers sometimes overlook it.

Modeling also played an important role in the early understanding of the dynamics of the vocal cords (e. g. Flanagan & Cherry, 1969) and in understanding the interactions between the vocal cords and the vocal tract (Flanagan & Meinhart, 1964). It turned out that the simplest possible model, based on a single mass-spring oscillator approximation of the vocal cords cannot adequately explain the behavior of the human vocal cords. A coupled system of two oscillators was minimally needed to explicate vocal cord motion (Dudgeon, 1970; Ishizaka & Flanagan, 1972). More elaborate models, able to reproduce more vocal phenomena, were developed by Titze (1973, 1974). Because of the complicated dynamics of the interaction between air flow, the vocal cords and the vocal tract, computers were essential in their investigation.

From the mid 1960's to the early 1970's a number of efforts were made to build articulatory synthesizers. These were both meant to be used for synthesizing artificial speech as well as for better understanding of speech production, especially the relation between the (discrete) phonemic level and the continuous speech signal. The models of Henke (1966) and Coker and Fujimura (1966) are probably the earliest examples. While Coker and Fujimura's model was a geometrical model, in which the geometry

and the articulators of the vocal tract were modeled directly, Henke's model was more data-oriented. It was based on pre-defined vocal tract shapes that were used as targets for the simulated vocal tract. Modeling efforts that were even more data-oriented tried to extract articulatory models from measured vocal tract shapes directly, based on either Fourier analysis (Liljencrants, 1971) or factor analysis (Lindblom & Sundberg, 1971). Coker and Fujmura's model formed the basis of Mermelstein's (1973) articulatory model as well as the models derived from it (Boersma, 1998; de Boer, 1999; Goldstein, 1980). Liljencrants, Lindblom and Sundberg's models formed the basis of Maeda's (1990) articulatory models and its derivates (Boë, 1999). Mermelstein's and Maeda's models are illustrated in figure 1. Henke's model, although not as influential as the other articulatory models, provided the basis for a totally new line of research: modeling primate and Neanderthal vocal tracts.

3. Reconstruction of Vocal Capacities of Nonhumans

The first attempt to model non-human vocal tracts was undertaken by Lieberman, Klatt and Wilson (1969) using Henke's computer model. Articulations and their corresponding area functions were determined by manipulating the vocal tract of an anaesthetized rhesus monkey, as well as by measuring a cast of the vocal tract of a recently dead monkey. The acoustic products of these estimated vocal tract shapes were determined using the computer model. This work was the first to systematically explore the potential acoustic output of a nonhuman species' vocal tract. In the study of human speech, such exploration had never been undertaken, as linguists had never been particularly interested in hypothetical speech sounds, and because human languages tend to use the available acoustic space maximally, so the available space is already fully explored when modeling actual vowels, as shown in another classical example of the use of a computer models in the investigation of speech, Liljencrants & Lindblom (1972).

Another imporant step was to model the vocal abilities of potential ancestors of *Homo sapiens* in order to get an idea of how speech could have evolved. This was done by Lieberman and Crelin (1971) using a

hypothetical reconstructed Neanderthal vocal tract. The Neanderthal vocal tract was based on a fossil skull of a male Neanderthal (the La Chapelle-aux-Saints skull) and the *Homo sapiens* infant vocal tract. By manipulating this hypothetical vocal tract, Lieberman and Crelin explored the phonetic abilities of Neanderthals concluding that they were significantly surpassed by those of adult modern humans. Although their reconstruction of the Neanderthal vocal tract has been challenged (e. g. Houghton, 1993, Boë et al. 1999, 2002), their methodology is still considered innovative and has inspired further attempts at investigating the speech of fossil hominids.

One such attempt investigated the prerequisites of speech (Carré, Lindblom and MacNeilage 1995). They followed a rather different method than that of Lieberman and Crelin. Instead of estimating the acoustic capabilities of a given articulatory model, Carré and colleagues made a model that determined what area functions were *optimal* for producing formants that have either minimal or maximal frequencies. They did this by starting out with a linearly increasing area function and then determining what perturbations would either raise (for maximization) or lower (for minimization) the formant frequencies. They then applied these perturbations to the area function and repeated the process 20 times, taking care that areas did not become smaller than 0.5 cm^2 or larger than 10cm^2. This procedure resulted in four different area functions (maxima and minima for both the first and the second formant). The formant frequencies associated with these area functions, assuming a length of the vocal tract of 18 cm, corresponded closely to those found in human speakers. The shapes of the area functions indicated the necessity of two independently controllable cavities of approximately equal length: an oral (mouth) cavity and a pharyngeal (throat) cavity. This implies that, to produce as distinctive speech sounds as possible, a vocal tract must have an independently controllable oral and pharyngeal cavities nearly equal in length. Because humans have such a vocal tract, but other primates do not, Carré and colleagues argued (like Lieberman and Crelin 1971) that the modern human vocal tract must have evolved for producing distinctive speech.

A different attempt at investigating the vocal abilities of Neanderthals (as well as newborn human infants) was made by Boë, Heim, Honda and Maeda (2002), who constructed an articulatory model of

the Neanderthal vocal tract, based on a more recent reconstruction of Neanderthal anatomy. Its computational base was Maeda's (Maeda, 1990) articulatory synthesizer, modified (Boë, 1999) such that it could model vocal tracts with different ratios of oral to pharyngeal cavity length. Boë and colleagues systematically explored the acoustic space that could be reached by this model to determine what they call the 'maximal vowel space'. The maximal vowel space corresponds relatively closely to the articulations that Carré and colleagues generated, but also contains the points in between the articulations with maximal and minimal formant frequencies, thus covering the acoustic space more thoroughly. The results suggested that Neanderthals could produce vowels that were as "distinct" as those of modern human, that newborn infants can produce the same range of vowels, and therefore that vocal tract anatomy does not determine the range of speech sounds that an organism can potentially produce. This model will be discussed in more detail in the case study below.

Another example of the use of computational models in studying non-human vocal abilities is presented in a paper by Riede, Bronson, Hatzikirou and Zuberbühler (2005). They investigate Diana monkey (*Cercopithecus diana*) formant frequencies, based on a combination of anatomical data from live and dead monkeys as well as on a computer simulation of the species' vocal tract. They conclude that Diana monkeys could, in principle at least, make relatively large changes to the formants of their calls. Although this work is an important step towards a better understanding of primate vocalizations, their computer simulation is rather simplistically based on a simple three- or four-tube model that does not model articulatory constraints. It also does not model air sacs, although these are likely to play an important role in Diana monkey vocalizations (see Fitch & Hauser 1995, Hewitt, MacLarnon & Jones 2002, Riede, Tokuda, Munger, & Thomson, 2008, de Boer, 2008, to appear). For these reasons the paper has been sharply criticized by Lieberman (2006). He points out that manipulations of a four tube model cannot possibly capture the constraints on tongue movement that exist in a real monkey's vocal tract, and that Riede and colleagues' model does not provide a realistic estimate of the acoustic diversity of signals that Diana monkeys could theoretically produce.

This controversy revolves around the question of how realistic articulatory models of non-human vocal tracts should be. Ideally, such models would be as accurate as possible, but in practice accuracy is limited by two factors. The first is computational: articulatory models require a lot of computation, and this limits the level of achievable accuracy. Although in the earlier work this played an important role, with ever increasing speed of computers it is becoming less of an issue. The second limiting factor is our relative ignorance concerning the anatomy and actual dynamics of non-human vocal tracts. Little is known about the vocal tracts and vocal abilities of living primates and other mammals. Many of the parameters that would be required to build accurate models remain unknown in nonhuman species, and what observations exist strongly suggest that mammalian vocal tracts are flexible, dynamic structures whose static anatomy provides a poor guide to vocal potential (Fitch 2000b). Obviously, even less is known about the vocal tracts of Neanderthals and other fossil hominins

For these reasons, at present, some simplifications need to be made, and the question becomes which simplifications are valid. The models of the human vocal tract described above all make a large number of simplifications, but nonetheless work well in the context for which the models were meant: reproducing human speech. However, problems can occur when extending such models to animals or fossil hominids, in two areas First, a non-human vocal tract may add different structures, capabilities or constraints than those present in a human. For example, laryngeal air sacs are large extensions of the glottal region which often have soft walls. When modeling vowel articulations, the human vocal tract can be reasonably approximated without damping but a vocal tract with air sacs (e.g. most apes and many monkeys, Hewitt, MacLarnon & Jones 2002) cannot be realistically modeled without damping.

Second, a crucial difference between recreating *actual* articulations, and exploring the range of *possible* articulations, occurs when using models based on factor analyses of the human tongue, such as the Liljencrants or Maeda models (Liljencrants, 1971; Maeda, 1990). Such models can recreate existing tongue shapes with few parameters reasonably well. However, this is no guarantee that they produce

realistic tongue shapes when using *random* articulatory parameter values, or when applied to animals with different anatomy. Both problems are illustrated by the use of lossless tube approximations of animal vocal tracts, such as (Riede, Bronson, Hatzikirou, & Zuberbühler, 2005). These examples highlight the need to consider and understand each animal's vocal tract in its own terms, taking into account articulatory and acoustic constraints applicable to that species, and checking the resulting articulations for acoustic and articulatory plausibility (e.g. Owren, Seyfarth, & Cheney, 1997).

4. The Boë/Maeda Study of Vocal Tract Potential: A Case Study

Several recent papers published by Boë and colleagues (Boë, 1999; Boë, Maeda, & Heim, 1999), including most recently (Boë, Heim, Honda, & Maeda, 2002), continue the debate concerning the vocal abilities of extinct hominids (cf. Lieberman 2007). Boë and colleagues report a series of computer simulations which, they claim, show that human newborns and Neanderthals have the same (or greater) vowel range as adult humans. If the simulation results are correct they would overturn a long-held belief in speech science: that vocal anatomy plays a crucial role in controlling the range of phonemes that adult humans can produce (Lieberman, Klatt & Wilson 1969, Carré, Lindblom, & MacNeilage, 1995; Fant, 1960; Stevens & House, 1955). We will argue that Boë and colleagues' results are an artifact of the articulatory synthesis model used in the studies.

At issue is whether the downward descent of the human larynx, hyoid and tongue root, which creates an elongated pharynx relative to most mammals, enlarges the phonetic repertoire. This vocal tract configuration has been argued to enable a novel "front/back" dimension of tongue movement, enlarging the producible repertoire of vowels to include the point vowels, /i/, /a/ and /u/, which are found in nearly every modern human language (Crothers, 1978; Maddieson, 1984). This core idea is based on data from comparative anatomy (Bowles, 1889; Laitman & Crelin, 1976; Negus, 1929, 1949), computer simulations (Lieberman, Klatt & Wilson 1969; Carré, Lindblom & MacNeilage 1995) and acoustic analyses of vocalizations from both children and animals (Buhr, 1980; Owren, Seyfarth, & Cheney, 1997).

Why would a lowered larynx enlarge the phonetic repertoire? The argument has both an acoustic and a physiological element. As far as the acoustics go, it has been relatively clear since (Fant, 1960) that the ability to produce sharp discontinuities in vocal tract area near the center of the vocal tract - often termed a "two tube" vocal tract – is a requirement to produce the point vowels (see also Carré, Lindblom, & MacNeilage, 1995; Lieberman, 1984 ; Nearey, 1978; Fitch 2000).

The main issue in dispute is therefore about the physiological capability of the vocal tract of any particular species to attain such discontinuous shapes. First, it is important to note that a lowered larynx *per se* is not the key phonetic factor. It is crucially the lowering of the *tongue root* that is critical in reconfiguring the human vocal tract. Both the intrinsic musculature of the tongue and the extrinsic muscles that attach it to the rest of the body are constant among primates and generally among mammals, and comparisons of chimpanzees and humans reveal no essential differences in this basic anatomy (Takemoto, 2001). Recent research shows that many mammals lower the larynx during vocalization (Fitch, 2000b), and some even have a permanently descended larynx (Fitch, 2002; Fitch & Reby, 2001; Weissengruber, Forstenpointner, Peters, Kübber-Heiss, & Fitch, 2002). What is different about the human vocal tract is its shape: it is distinctly bent, at almost a right angle, around its middle (Lieberman, 1984), and the descended human tongue root means that the tongue shape is similarly reconfigured.

Because of this reconfiguration of tongue shape, the extrinsic tongue muscles exert different vector forces, leading to different vocal tract configurations, in humans than they do in other mammals. When the human styloglossus muscle tenses, it bunches the middle of the tongue backwards and upwards to form a midpoint constriction in the vocal tract, a prerequisite for creating an /u/. The same muscle in a normal mammal (e.g., a dog or an opossum), simply moves the tongue body backward (caudally) in the oral cavity. Such a movement will have a much smaller phonetic effect on the lowest two formant frequencies. For a dog, opossum or chimpanzee to create a midpoint constriction, it would either need a muscle attaching at the tongue midpoint and pointed upward towards the hard palate (and no such muscle exists), or would need to use the intrinsic musculature of the tongue to distort the tongue's shape. That is, to create a comparable vocal tract transfer function (and thus formant frequencies), an animal without a

reconfigured vocal tract would have to use a different set of articulatory maneuvers and musculature actions than those use by humans to attain this configuration.

These and similar considerations, combined with physiologically-constrained computer models (Lieberman, Klatt & Wilson 1969), led to the view that a human-like vocal tract configuration is required to produce [u]. Similar considerations apply for [i], and true [a], and perhaps several consonants as well. Because such sounds are ubiquitous in all the world's languages, the reconfigured human vocal tract has long been seen as an important element in our capacity for speech, and thus in the evolution of human language (Fitch, 2000a; Lenneberg, 1967; Lieberman, 1984).

Although an ordinary mammalian vocal tract, properly controlled, might be unable to produce point vowels, any mammal could nonetheless produce enough phonetic distinctions to support a basic spoken language (Fitch, 2000b; Lieberman, 1968). The remarkably humanlike utterances of Hoover, a harbor seal which learned to produce some stereotyped English phrases (Ralls, Fiorelli, & Gish, 1985) nicely illustrates this point, since there is no evidence of a descended larynx (either permanent or temporary) in harbor seals. Thus, the earliest speech sounds, during human evolution, might well have been produced with a chimpanzee-like vocal tract (Lieberman, 1984, 2000; Lieberman & Crelin, 1971). At issue, then, is not the question of whether a Neanderthal could speak, or had language in general (which, as most commentators agree, is dependent mainly upon neural factors), but instead about the specific phonetic characteristics of the speech they could produce.

The simulations reported by Boë's group suggest that either Neanderthal, or human infant, vocal tracts could produce all the speech sounds available to adult modern humans. Do these simulations provide a justification for reconsidering the view sketched above? We suggest that the answer is no, because the model used mathematically incorporates the possibility of a full range of human tongue shapes. This is because tracings of adult human tongues, making all speech gestures, were used as the starting point for Maeda's (1990) vocal tract model. Because of this "anthropomorphic" nature, incorporating the assumption of a full range of vocal tract shapes into the model itself, we do not think that model is

appropriate for testing questions about the phonetic capabilities of newborn humans or nonhuman species.

Shinji Maeda's vocal tract model (Maeda, 1990) is an articulatory model based on cineradiographic and labiographic data from two adult female speakers. In the digitization procedure, a coordinate system based on a two-tube vocal tract (with straight oral and pharyngeal portions, connected at a fixed angle by a curved region) was used. A factor analysis was used to extract 4 empirically-derived factors: jaw position, tongue position, tongue shape and tongue tip. Note in particular that the model derives two key variables for the tongue, which explain most of the variance in vocal tract shape: tongue (dorsum) position (front/back), and tongue (dorsum) shape (arched/flat). Due to the choice of coordinate frame, both of these factors entail a two-tube vocal tract and correspondingly reshaped tongue; see Figure 3 from (Maeda, 1990). Crucially, once the tongue shape factors have been extracted, they will allow the corresponding tongue shapes regardless of any distortion or rearrangement of the vocal tract geometry via changes in the coordinate frame. We see this as a crucial flaw underlying any attempt to use this model to understand nonhuman vocal production. In Maeda's model the human capability for deforming the tongue, and thereby creating particular vocal tract shapes, is intrinsic to the model irrespective of changes in the nominal laryngeal position, or changes in oral/pharyngeal cavity length.

Boë's simulation approach involves linearly stretching the oral and pharyngeal portions of the coordinate system of the Maeda model, supposedly matching the static vocal anatomy of infants or other species. But such warping preserves both the oral and pharyngeal components, and the sharp angle between them, of the original adult human data. This is precisely the aspect of the anatomy that is in question. Most animals do not have an L-shaped vocal tract,: the tongue lies flat in the oral cavity, with the tongue root lying in the same oral plane. Thus, the sharp tongue shape changes seen in the human vocal tract, and the resultant abrupt changes in vocal tract shape (necessary to produce point vowels), are guaranteed to be present in any species modeled using Maeda's software. The model 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. The problems with the model are illustrated in figure 2.

Another potential problem when using an articulatory model based on a linear factor analysis emerges because the underlying system is not really a linear superposition of the different articulatory parameters. This makes it almost inevitable that certain combinations of parameter values result in articulations that are impossible with the original system. This is not a problem when using the model to mimic attested vocal tract shapes, or small perturbations around these. But when using such a model to explore the available articulatory space in a drastically re-configured vocal tract model, this assumption of linear superposition could easily lead to misleading results. An example of a simple system showing this problem is given in figure 3.

The problems with the Maeda model are perhaps best illustrated by their study of infant vocalizations. Independent studies have shown that a mature supralaryngeal vocal tract anatomy, with a rough match between oral and pharyngeal cavity length is not achieved until age 6–8 years (Fitch & Giedd 1999, Lieberman & McCarthy 1999; Vorperian, Kent, Lindstrom, Gentry & Yandell 2005). Hence the Maeda model will lead to misleading inferences concerning the speech motor capabilities of young children in studies such as (Serkhane, Schwartz, Boë, Davis, & Matyear, 2007) which attempt to compare the formant frequency ranges of children's vocalizations with the boundaries imposed by their supralaryngeal vocal tract anatomy. These problems are of course compounded in attempts to reconstruct vocal capabilities of extinct hominids for which the true vocal anatomy is not, and probably never will be, known.

We conclude that Boë and his colleagues simulations suffer from logical circularity, in the sense that the model they use implicitly builds in precisely the key capability – vocal tract flexibility – that is disputed, and thus nearly guarantees the results they have consistently reported. This is not a criticism of the Maeda model, when used for the purposes for which it was developed (namely, understanding human speech). But using the model to "test" speech capabilities of babies or animals appears to us inappropriate. A better understanding of other species' (including our evolutionary ancestors) vocal abilities, as well as those of infants, requires models more directly tied to the physiological and anatomical nature of the mammalian vocal tract (e.g. Wilhelms-Tricarico, 1995), updated with respect to

modern data (e.g. from MRI and cineradiography of babies and/or animals (Bosma & Lind, 1965; Fitch, 2000b)). Alternatively, a modification of Maeda's factor-analytic approach, incorporating tongue shape parameters extracted from the specific species being modeled, would be equally appropriate.

Until such further research is performed, a considerable body of well-established research indicates that the descent of the larynx, hyoid and tongue root (whether permanent or dynamic) is a prerequisite for producing the full phonetic range of modern human speech. Despite their strongly worded claims, the results of Boë and colleagues' simulations do not, and in principle cannot, demonstrate that vocal tract anatomy is "irrelevant" to human speech production.

5. Discussion

The long and productive history of interaction between computer models and the study of speech shows that computer modeling and experimental studies complement each other nicely. Application of such models to speech evolution provides a number of more general lessons for modeling the evolution of language. The first lesson is that one cannot always apply models and theoretical frameworks that have been developed for *modern* human language to the *evolution* of language and expect valid results. For example, classic models of vocal fold vibration (e. g. Alipour, Berry, & Titze, 2000; Rosenberg, 1971), combined with the assumption that the source of vocal energy and the filtering action of the vocal tract are independent, reproduce ordinary human speech quite successfully. However, these assumptions are not necessarily valid for producing other types of (human) phonation, let alone for the kinds of phonation that animals use (or ancestral hominids used). Animals may use other tissues for sound production, which vibrate in different ways, and these might be more strongly coupled to the upper vocal tract (Fitch, Neubauer, & Herzel, 2002), violating the assumption of source/tract independence.

Similar limitations on generality are inherent in all models of complex phenomena, which inevitably entail simplifications. In order to reproduce realistic behavior, parameters and algorithms in these models need to be tuned to the available data. This increases the likelihood that misleading results will be

obtained in other, unobserved, regimes of behavior. We suggest that the same is true for other linguistic phenomena, including syntactic and semantic processing, where models that describe modern human behavior well might not be appropriate for modeling behavior of ancestral hominids.

A related lesson is that maximal realism of every aspect of a model is not a goal in itself. Although it can sometimes be necessary to use highly realistic physical simulations, it should be kept in mind that the aim of many simulations is not to produce an extremely realistic model per se, but to understand how ancestral hominids produced sounds and to measure the approximate range of sounds they could hypothetically produce. For gaining such insights it is often better to use models at a slightly higher level of abstraction. More abstract models often provide more insight, because it is easier to understand why they show the behavior that they do, and they do not incorporate as many implicit and possibly incorrect assumptions. Furthermore, more abstract models are simpler computationally, and easier to reimplement and adapt by other researchers.

The final lesson we draw from the history of modeling the evolution of speech is that models should be based on *physical* simplifications, not on *computational* simplifications alone. Computational simplifications like curve fitting, factor analysis and interpolation can be used advantageously when trying to build usable systems to recreate a given set of utterances or articulations. However, as the case study illustrates, such models may not generalize well, and cannot be reliably extended far outside the range of utterances for which they were developed. When a model is intended to represent a range of animal or hypothetical ancestral vocal tracts, each simplification should be physically and physiologically justified. This makes it possible to demonstrate the physical validity of the model outside the range of utterances and anatomical measurements on which it was based. We hope that the observations presented in this paper will be of help to future researchers building models of the evolution of speech and language.

Acknowledgments

We thank René Carré and Shinji Maeda for copies of their computer code, as well as very useful discussions concerning the functioning of their respective models, and the late Peter Ladefoged, Philip Lieberman and Frank Guenther for their helpful comments on an earlier version of the case study presented in this paper. Bart de Boer is funded by the Netherlands Organisation for Scientific Research (Geesteswetenschappen), project number 016.074.324. Part of this work was funded by an EU FP7 Grant to TF (CHLASC).

References

Alemseged, Z., Spoor, F., Kimbel, W. H., Bobe, R., Geraads, D., Reed, D., et al. (2006). A juvenile early hominin skeleton from Dikika, Ethiopia. *Nature*, 443, 296–301.

Alipour, F., Berry, D. A., & Titze, I. R. (2000). A finite-element model of vocal-fold vibration. *Journal of the Acoustical Society of America, 108*(6), 3003–3012.

Arensburg, B., Tillier, A. M., Vandermeersch, B., Duday, H., Schepartz, L. A., & Rak, Y. (1989). A Middle Palaeolithic human hyoid bone. *Nature, 338*(6218), 758–760.

Boë, L.-J. (1999). Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults consequences for ontogenesis and phylogenesis. In *14th International Congress of Phonetic Sciences* (Vol. 3, pp. 2501–2504). San Francisco.

Boë, L.-J., Heim, J.-L., Honda, K., & Maeda, S. (2002). The potential Neandertal vowel space was as large as that of modern humans. *Journal of Phonetics, 30*(3), 465–484.

Boë, L.-J., Maeda, S., & Heim, J.-L. (1999). Neandertal man was not morphologically handicapped for speech. *Evolution of Communication, 3*(1), 49–77.

Boersma, P. (1998). *Functional Phonology*. The Hague: Holland Academic Graphics.

Bosma, J., & Lind, J. (1965). Cry motions of the newborn infant. *Acta Paediatrica Scandanavica, Suppl 163*, 61–92.

Bowles, R. L. (1889). Observations upon the mammalian pharynx, with especial reference to the epiglottis. *Journal of Anatomy and Physiology, London, 23*, 606–615.

Buhr, R. D. (1980). The emergence of vowels in an infant. *Journal of Speech and Hearing Research, 23*, 75–94.

Carré, R., Lindblom, B., & MacNeilage, P. F. (1995). Rôle de l'acoustique dans l'évolution du conduit vocal humain. *Comptes Rendus de l'Académie des Sciences, Série IIb, 320*, 471–476.

Chiba, T., & Kajiyama, M. (1942). *The Vowel: Its Nature and Structure.* Tokyo: Tokyo-Kaiseikan. Reprinted as: Chiba, T., & Kajiyama, M. (1958). *The Vowel: Its Nature and Structure.* Tokyo: Phonetic Society of Japan.

Coker, C. H., & Fujimura, O. (1966). Model for Specification of the Vocal-Tract Area Function. *Journal of the Acoustical Society of America, 40*(5), 1271.

Cook, P. R. (1993). SPASM, a real-time vocal tract physical model controller; and Singer, the companion software synthesis system. *Computer Music Journal, 17(1)*, 30–44.

Crothers, J. (1978). Typology and universals of vowel systems. In J. Greenberg, C. A. Ferguson & E. A. Moravcsik (Eds.), *Universals of Human Language* (Vol. 2, pp. 93–152). Stanford, CA: Stanford University Press.

de Boer, B. (1999). *Self-organisation in vowel systems*. Brussels: Ph. D. Thesis AI-lab Vrije Universiteit Brussel.

de Boer, B. (2002). Evolving Sound Systems. In A. Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language* (pp. 79–97). Berlin: Springer Verlag.

de Boer, B. (2008). The acoustic role of supralaryngeal air sacs. *Journal of the Acoustical Society of America, 123*(5 Pt 2), 3732–3733.

de Boer, B. (to appear). Analysis of air sac acoustics.

Dudgeon, D. E. (1970). Two-Mass Model of the Vocal Cords. *Journal of the Acoustical Society of America, 48*(1A), 118.

Dunn, H. K. (1950). The Calculation of Vowel Resonances, and an Electrical Vocal Tract. *Journal of the Acoustical Society of America, 22*(6), 740–753.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fitch, W. T., & Hauser, M. D. (1995). Vocal production in nonhuman primates: Acoustics, physiology, and functional constraints on "honest" advertisement. *American Journal of Primatology, 37*, 191–219.

Fitch, W. T. (2000a). The evolution of speech: a comparative review. *Trends in Cognitive Sciences, 4*(7), 258–267.

Fitch, W. T. (2000b). The phonetic potential of nonhuman vocal tracts: Comparative cineradiographic observations of vocalizing animals. *Phonetica, 57*, 205–218.

Fitch, W. T. (2002). Comparative Vocal Production and the Evolution of Speech: Reinterpreting the Descent of the Larynx. In A. Wray (Ed.), *The Transition to Language* (pp. 21–45). Oxford: Oxford University Press.

Fitch, W. T., Neubauer, J., & Herzel, H. (2002). Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour, 63*, 407–418.

Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society of London, B, 268*(1477), 1669–1675.

Flanagan, J. L. (1965). *Speech analysis, synthesis and perception*. Berlin: Springer.

Flanagan, J. L., & Cherry, L. (1969). Excitation of Vocal-Tract Synthesizers. *Journal of the Acoustical Society of America, 45*(3), 764–769.

Flanagan, J. L., & Meinhart, D. I. S. (1964). Source-System Interaction in the Vocal Tract. *Journal of the Acoustical Society of America, 36*(10), 2001–2002.

Goldstein, U. G. (1980). *An Articulatory Model for the Vocal Tracts of Growing Children.* Unpublished PhD, Massachusetts Institute of Technology, Cambridge (MA).

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

Henke, W. L. (1966). *Dynamic articulation model of speech production using computer simulation.* Massachusetts Institute of Technology, Cambridge, MA.

Hewitt, G., MacLarnon, A., & Jones, K. E. (2002). The functions of laryngeal air sacs in primates: a new hypothesis. *Folia Primatologica, 73*, 70–94.

Houghton, P. (1993). Neandertal Supralaryngeal Vocal Tract. *American Journal of Physical Anthropology, 90*, 139–146.

Ishizaka, K., & Flanagan, J. L. (1972). Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords. *The Bell system technical journal, 51*(6), 1233–1268.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America, 82*(3), 737–793.

Laitman, J. T., & Crelin, E. S. (1976). Postnatal development of the basicranium and vocal tract region in man. In J. F. Bosma (Ed.), *Symposium on Development of the Basicranium*. Washington, D.C.: U.S. Government Printing Office.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. New York, NY: Wiley.

Lieberman, D. E., & McCarthy, R. C. (1999). The ontogeny of cranial base angulation in humans and chimpanzees and its implications for reconstructing pharyngeal dimensions. *Journal of Human Evolution, 36*, 487–517.

Lieberman, D. E., McCarthy, R. C., Hiiemae, K., & Palmer, J. B. (2001). Ontogeny of postnatal hyoid and larynx descent in humans. *Archives of Oral Biology, 46*, 117–128.

Lieberman, P. (1968). Primate vocalization and human linguistic ability. *Journal of the Acoustical Society of America, 44*(6), 1574–1584.

Lieberman, P. (1984). *The Biology and Evolution of Language*. Cambridge, MA: Harvard University Press.

Lieberman, P. (2000). *Human language and our reptilian brain: the subcortical bases of speech, syntax and thought*. Cambridge, MA: Harvard University Press.

Lieberman, P. (2006). Limits on tongue deformation - Diana monkey formants and the impossible vocal tract shapes proposed by Riede et al. (2005). *Journal of Human Evolution, 50*(2), 219–221.

Lieberman, P. (2007). Current views on Neanderthal speech capabilities: A reply to Boë et al. (2002). *Journal of Phonetics, 35*, 552–563.

Lieberman, P., & Crelin, E. S. (1971). On the speech of Neanderthal man. *Linguistic Inquiry, 2*(2), 203–222.

Lieberman, P., Klatt, D. H., & Wilson, W. H. (1969). Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science, 164*, 1185–1187.

Lieberman, P., & Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge, UK: Cambridge University Press.

Liljencrants, J. (1971). A Fourier series description of the tongue profile. *Speech Transmission Laboratory Quarterly Progress and Status Report, 12*(4), 9–18.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems. *Language, 48*, 839–862.

Lindblom, B., & Sundberg, U. (1971). Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. *Journal of the Acoustical Society of America, 50*(4), 1166–1179.

MacLarnon, A., & Hewitt, G. P. (1999). The Evolution of Human Speech: the role of enhanced breathing control. *American Journal of Physical Anthropology, 109*(3), 341–343.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Maeda, S. (1989). Compensatrory 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 Speech Modelling* (pp. 131–149). Dordrecht: Kluwer Academic Publishers

Martínez, I., Arsuaga, J.-L., Quam, R., Carretero, J.-M., Gracia, A., & Rodríguez, L. (2008). Human hyoid bones from the middle Pleistocene site of the Sima de los Huesos (Sierra de Atapuerca, Spain). *Journal of Human Evolution, 54*, 118–124.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53*(4), 1070–1082.

Nearey, T. (1978). *Phonetic Features for Vowels*. Bloomington: Indiana University Linguistics Club.

Negus, V. E. (1929). *The Mechanism of the Larynx.* London: Heinemann.

Negus, V. E. (1949). *The Comparative Anatomy and Physiology of the Larynx*. New York: Hafner Publishing Company.

Owren, M. J., Seyfarth, R. M., & Cheney, D. L. (1997). The acoustic features of vowel-like *grunt* calls in chacma baboons (*Papio cyncephalus ursinus*): Implications for production processes and functions. *Journal of the Acoustical Society of America, 101*, 2951–2963.

Perkell, J. S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge, MA: MIT Press.

Ralls, K., Fiorelli, P., & Gish, S. (1985). Vocalizations and vocal mimicry in captive harbor seals, *Phoca vitulina*. *Canadian Journal of Zoology, 63*, 1050–1056.

Riede, T., Bronson, E., Hatzikirou, H., & Zuberbühler, K. (2005). Vocal production in a non-human primate: morphological data and a model. *Journal of Human Evolution, 48*(1), 85–96.

Riede, T., Tokuda, I. T., Munger, J. B., & Thomson, S. L. (2008). Mammalian laryngseal air sacs add variability to the vocal tract impedance: Physical and computational modeling. *Journal of the Acoustical Society of America, 124*(1), 634–647.

Rosen, G. (1958). Dynamic Analog Speech Synthesizer. *Journal of the Acoustical Society of America, 30*(3), 201–209.

Rosenberg, A. E. (1971). Effect of Glottal Pulse Shape on the Quality of Natural Vowels. *Journal of the Acoustical Society of America, 49*(2 (Part 2)), 583–590.

Schroeder, M. R. (1993). A brief history of synthetic speech. *Speech Communication, 13*(1–2), 231-237.

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*(3), 321–340.

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America, 27*(3), 484–493.

Stevens, K. N., Kasowski, S., & Fant, G. (1953). An Electrical Analog of the Vocal Tract. *Journal of the Acoustical Society of America, 25*(4), 734–742.

Takemoto, H. (2001). Morphological analyses of the human tongue musculature for three-dimensional modeling. *Journal of Speech, Language and Hearing Research, 44*, 95–107.

Titze, I. R. (1973). The Human Vocal Cords: A Mathematical Model Part I. *Phonetica, 28*(3), 129–170.

Titze, I. R. (1974). The Human Vocal Cords: A Mathematical Model Part II. *Phonetica, 29*(1), 1–21.

Titze, I. R. (1994). *Principles of voice production*. Englewood Cliffs, N.J.: Prentice Hall.

van de Berg, J. (1955). Calculations on a Model of the Vocal Tract for Vowel /i/ (Meat) and on the Larynx. *Journal of the Acoustical Society of America, 27*(2), 332–338.

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.

Weissengruber, G. E., Forstenpointner, G., Peters, G., Kübber-Heiss, A., & Fitch, W. T. (2002). Hyoid apparatus and pharynx in the lion (*Panthera leo*), jaguar (*Panthera onca*), tiger (*Panthera tigris*), cheetah (*Acinonyx jubatus*), and domestic cat (*Felis silvestris* f. *catus*). *Journal of Anatomy (London), 201*, 195–209.

Wilhelms-Tricarico, R. (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America, 97*(5), 3085–3098.

Figure Captions

Figure 1: Illustrations of a geometric articulatory model (A) based on Mermelstein's (1973) model and of a statistical model (B) based on the principal components of Maeda's (1990) model. For both models, the projections of the vocal tract in the symmetry plane of the head are shown. The mouth is to the right. In the geometric model, the outline is determined by direct manipulations of the outline through the action of the articulatory parameters. Two examples are shown in grey. In the statistical model, articulations are formed by adding factors. Effects of the four factors of Maeda's model (the variations caused by jaw motion, and the variations caused by the first three remaining principal components) are shown. A complete vocal tract outline is calculated by adding the effects of all factors.

Figure 2: Principle underlying of Boë's (1999) adaptation of Maeda's (1990) model. The vertical part and the horizontal part of Maeda's vocal tract can be scaled independently. Outline A shows Maeda's original model and the effect of manipulating his first principal component (see figure 1). The outline B shows what happens when the vertical part is scaled 50%. The tongue root is deformed unrealistically, bcause the motions of the original tongue are preserved despite drastic changes in actual muscle angles. The outline C shows a tracing of a chimpanzee vocal tract (Fitch 2000a, Fig. 1), and outline D the hypothesized actions (corresponding to Maeda's first principal component) it could perform with its tongue in the resting (high tongue-root and larynx) position.

Figure 3: Illustration of artifacts caused by using a independent components model to sample potential articulations. The extremely simplified articulatory model consists of a straight tube with a wedge-shaped tongue (A.1). The tongue has two degrees of freedom: front-back position (A.2) and height (A.3). By generating random articulations, two principal components can be derived (B.1 and B.2). The solid line shows the average of the tongue shape, while the dotted lines indicate displacements caused by the components. B.1 models height, while B.2 can be said to model position. The range of the two principal

components is –0.25 to 0.25 (indicated as the grey square in C) and possible combinations of the parameters are illustrated as black dots. Valid combinations result in realistic, albeit rounded tongue shapes (original is grey line in D, while reconstruction is the curved black line). Parameter combinations outside the valid range result in impossible tongue shapes (E shows combination 0.25, 0.25).

Figure 1

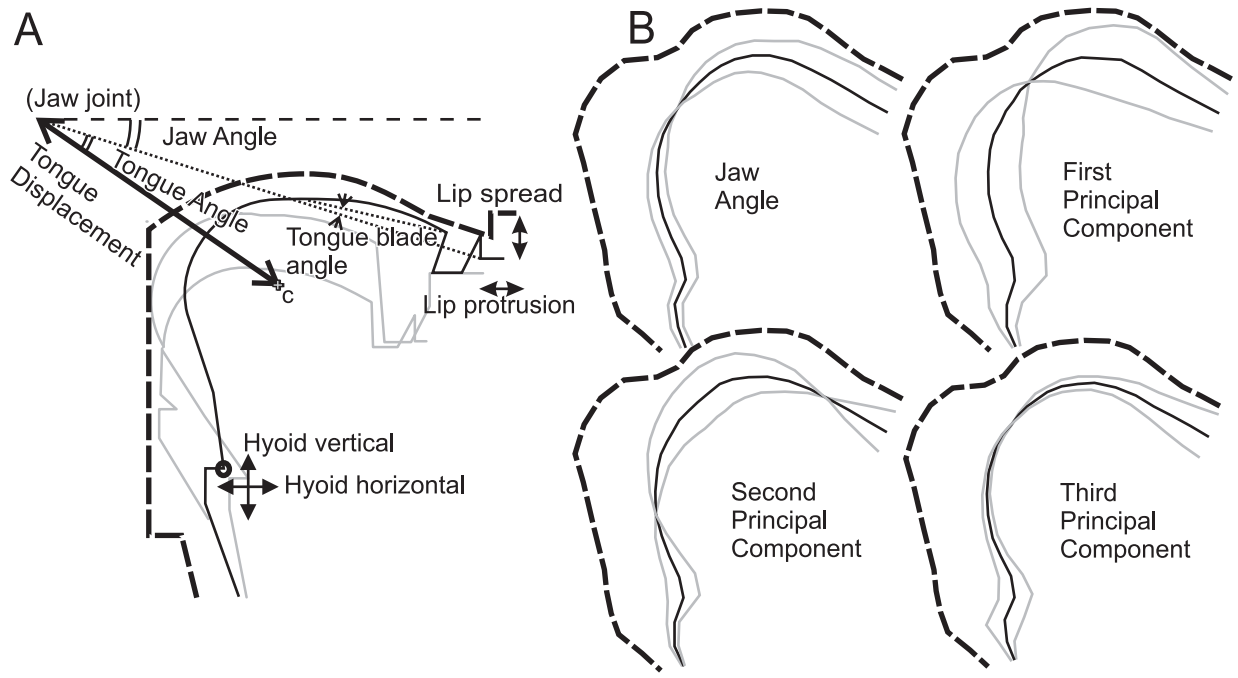
A
(Jaw joint)
Jaw Angle
Tongue Displacement
Tongue Angle
Tongue blade angle
c
Lip spread
Lip protrusion
Hyoid vertical
Hyoid horizontal
B
Jaw Angle
First Principal Component
Second Principal Component
Third Principal Component

Figure 2

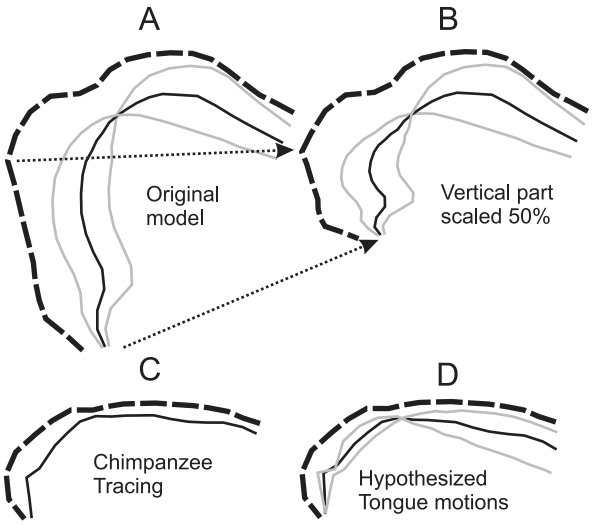
A
B
Original
model
Vertical part
scaled 50%
C
D
Chimpanzee
Tracing
Hypothesized
Tongue motions

Figure 3