Equation Chapter 1 Section 1**Title:** Modelling vocal anatomy's significant effect on speech

Running Title: Modelling anatomy's effect on speech

Author: B. G. de Boer

Institute: Amsterdam Center for Language and Communication, Univesiteit van Amsterdam

Address: Spuistraat 210, 1012 VT, Amsterdam, the Netherlands

Email: b.g.deboer@uva.nl

Abstract:

This paper investigates the effect of larynx position on the articulatory abilities of a human-like vocal tract. Previous work has investigated models that were built to resemble the anatomy of existing species or fossil ancestors. This has led to conflicting conclusions about the relation between the evolution of anatomy and the evolution of speech. Here a model is proposed to systematically investigate the relation between larynx height and articulatory abilities. It is a simplified model of primate vocal anatomy that nevertheless preserves the essential articulatory constraints due to limitations of muscular control. It is found 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. This has been a conjecture for a long time by those researchers of the evolution of speech who propose that the human vocal tract has evolved for speech. A short recapitulation of acoustic theory of speech production is presented to explain the reason for why this configuration is optimal.

The optimal configuration corresponds closely to human female anatomy, while in the human male the larynx is slightly lower than optimal. These results agree with the hypothesis that modern human vocal anatomy has evolved because of speech, and that male larynx position might have been lowered further for reasons of size exaggeration.

Keywords: Descended Larynx, Articulatory Model, Vocal Tract Evolution, Evolution of Speech

1 Introduction

The debate about the relation between the anatomy of the human vocal tract (specifically the position of the larynx) and the evolution of language is an old one (Negus 1938). It has even been the subject of what was possibly the first use of a computer model in the study of the evolution of language (Lieberman & Crelin 1971, Lieberman, Crelin, & Klatt 1972). Ever since Lieberman *et al.'s* work there has been a lively debate about whether the modern human larynx position has evolved for speech or not. Recently results of computer simulations have been published (Boë *et al.* 2002, Boë *et al.* 2007) that were used to argue that anatomy is in fact less important than was previously thought, and that neural control over the vocal tract is what differentiates our vocal abilities from those of the great apes, and by extension, our latest common ancestor (but see also Lieberman 2007). This paper presents simulation results that have been achieved with a simplified articulatory model. These results show that anatomy does matter, and that human anatomy and human vocal abilities agree with a maximization of articulatory abilities.

The focus of the debate is the shape of the human vocal tract. There are several other physical adaptations for speech such as fine control over breathing (MacLarnon & Hewitt 1999, 2004) and the anatomy of the vocal cords (e. g. Demolin & Delvaux 2006) as well as cognitive adaptations such as our ability for vocal imitation and our ability to control vocalizations consciously. However, the anatomy of the human vocal tract, and especially the permanently lowered position of the human larynx are perhaps most intriguing as they introduce a distinct evolutionary disadvantage. Whereas apes (and most other mammals) are able to link the larynx into the velum and can therefore breathe through their nose and swallow food at the same time, adult humans are unable to do this and therefore run an increased risk of choking on their food (Heimlich 1975). It has been argued that such a disadvantage must be offset by an advantage, and it is proposed that this is an ability to produce a larger range of speech sounds (probably first formulated by Negus 1938).

It must be stressed that a lowered larynx is not unique to humans, but exists in other mammals. This has been shown through many examples by Fitch and colleagues (Fitch & Reby 2001, Fitch & Hauser 2002). However, in all the cases that Fitch presents, the animals in question are still able to connect the larynx to the velum and are (presumably) able to swallow and breathe at the same time. Furthermore, their tongues are ordinary flat mammalian tongues. The human tongue, on the other hand, remains unique with its rounded shape. Although the uniqueness of human anatomy is therefore often summarized as consisting of a lowered larynx, in reality there are a number of anatomical changes in the vocal tract that make humans unique: a rounded tongue, a 90 degree angle between mouth and pharynx, as well as a lowered larynx combined with a raised velum that allow the rounded tongue freedom of movement.

An important alternative theory for explaining lowered mammalian larynges is that of size exaggeration (Fitch & Hauser 2002). Larger size of an acoustic system results in lower-frequency sounds. The frequency of sounds could therefore be used as an indication for body size. Ohala (1984) originally proposed that the fundamental frequency of the voice could be used to indicate body size, but as it is relatively easy to evolve larger vocal cords, this is not an honest signal (Fitch & Hauser 2002). The resonances of the vocal tract correlate much better to body size and therefore form a more honest signal. However, these resonances can be lowered and body size therefore exaggerated by lowering the larynx, and Fitch and Reby (Fitch & Reby 2001) show some truly spectacular examples of lowered larynges in animals. That the frequency of vocal tract resonances is also perceived as an indication of size and dominance by humans has been shown by Puts *et al.* (2006). It is therefore conceivable that the lowered larynx in humans can be explained as the result of size exaggeration. Interestingly Puts et al. found that men are more impressed by perceived body size than women, and any adaptations for size exaggeration are therefore most likely the result of male-male competition for dominance, rather than the result of men trying to impress women.

The physics of speech production lends itself well to modelling, but the resulting equations cannot by solved analytically in the general case, and must therefore be approximated numerically. For this reason computer models have been used extensively to investigate the articulatory abilities of the human vocal tract, the articulatory abilities of hypothetical evolutionary ancestors and the requirements of an ideal vocal tract. Lieberman and colleagues (Lieberman & Crelin 1971, Lieberman, Crelin, & Klatt 1972) made three-dimensional measurements of modern human, infant and monkey vocal tracts and proposed a shape of the Neanderthal vocal tract. They then calculated the acoustic properties of the different tracts with a computer model and compared the different ranges of signals that could be produced. They concluded that human vocal anatomy is necessary to produce the largest range of speech sounds, and because their Neanderthal model had ape-like anatomy, it could only produce a limited range of speech sounds.

Boë *et al.* (2002) recently have made a more sophisticated model of the Neanderthal vocal tract and conclude that it is much more similar to the modern human vocal tract. They also use a computer model to test its articulatory abilities, and find that it can produce the same range of speech sounds as modern humans can. However, their model can simulate a *range* of larynx heights (also those of infants) and they find that the range of speech sounds that can be produced does not depend strongly on larynx position. They therefore conclude that even vocal tracts with high larynges can produce the same range of speech sounds as modern humans can, and that therefore the bottleneck is in neural control of articulations (Boë *et al.* 2007).

A third series of computer models and mathematical models has been investigated by Carré and colleagues (Carré, Lindblom, & MacNeilage 1995, Carré & Mrayati 1995, Carré 2004, 2009). The aim of these models was not so much to investigate any particular vocal tract, but to derive from acoustic principles what a vocal tract would look like that has to produce as large a range of signals as possible (they therefore do not implement any articulatory or anatomic constraints). Carré and colleagues conclude that a human-like anatomy is essential for producing the maximal range of speech sounds.

Rather similar computer models have apparently resulted in contradictory conclusions. The present paper investigates the role of vocal anatomy with a simplified articulatory model in order to resolve the question whether the range of possible speech sounds is limited by anatomy or by control. In the next section, the basics of the acoustics of the vocal tract as well as the conditions for producing a maximal range of speech sounds are recapitulated. In section 3, the model is described and the design decisions are discussed in reference to existing models. In section 4 the results of the simulations with this model are presented. In the final two sections, the conclusions that can be drawn from this model are presented, and its impact on research into the evolution of speech and on the other existing models is discussed.

2 The acoustics

The basic acoustics that determine the range of signals that can be produced with a vocal tract is uncontroversial. However, in order to clarify the discussion, it helps to restate the basics. Speech sounds are perceived in terms of their constituent frequencies. Generally, the acoustic energy of a signal is concentrated in a number of peaks, which are called the formants. The perceived timbre, or quality, of a signal is determined mainly by the frequencies at which these peaks occur. On the production side, the position of the peaks in speech sounds

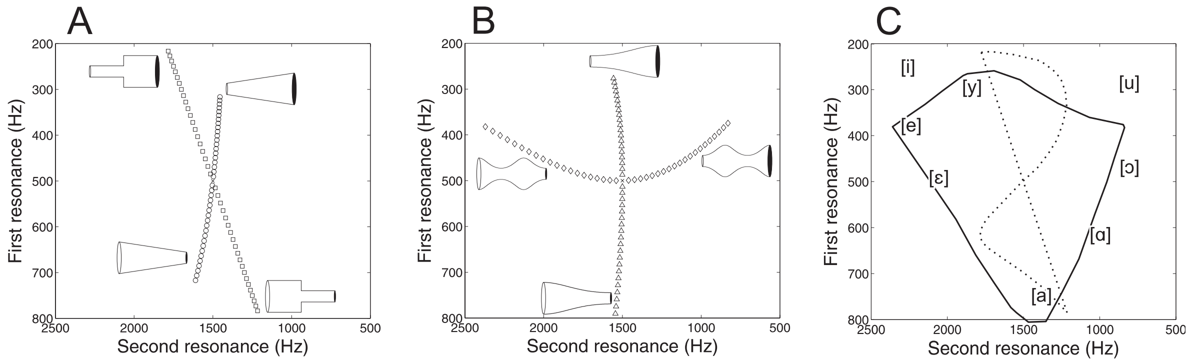

Figure 1: Effects in acoustic space of different articulatory systems. Diagram A shows the effects of simple articulators with one degree of freedom: a two-tube model (squares) and a conical tract (circles). Diagram B shows Schroeder's first cosinoidal deformation (triangles) and Schroeder's second cosinoidal deformation (diamonds). Drawings of the tubes are shown with the open end to the left. Diagram C shows the areas that can be covered by systems with two degrees of freedom: a combination of Schroeder's deformations (solid line) and a two-tube model (dotted line). For reference, positions of Dutch vowels as pronounced by the author are also given.

(especially vowels) is determined almost exclusively by the resonances of the vocal tract – in fact the vocal tract serves as the filter of the acoustic energy produced by the vocal cords or other possible sources of energy (Fant 1960).

Therefore, it is important to be able to produce a range of different resonance patterns if one wants to produce a range of different speech signals. In human speech, the first and the second resonances are most important (in distinguishing most vowels, for example) and the third resonance is also important for some sounds (for example in distinguishing /l/ from /r/ in English).

For the frequencies that are relevant for speech (lower than 3000 Hz) the details of the three-dimensional shape of the vocal tract are not important. The acoustics can be approximated very satisfactorily by considering the vocal tract to be a straight tube whose cross-sectional area can be changed along its length. What kind of control is then required to produce a large range of speech sounds? Lieberman et al. (1972) have stated that the human vocal tract is a two-tube system, whereas the chimpanzee (and human infant) vocal tract is a one-tube system. Although the acoustic analysis they present is sound, this formulation is confusing. In a prototypical two-tube system, only the cross-sectional areas of equally long front and back cavities can be controlled. However, in such a system the only acoustically relevant parameter is the ratio between the front and back cavity's area. Such a system can therefore only produce a one-dimensional manifold of possible resonance patterns (illustrated as the squares in figure 1A). Another one-dimensional manifold – but with different resonance patterns – could be obtained by only varying the degree of opening of the mouth (illustrated as the circles in figure 1A). In order to produce a large range of articulations, different ways of deforming the vocal tract (degrees of freedom) must be combined.

Not all ways of combining different deformations are equal, however. Ideally, each degree of freedom of the vocal tract would influence the resonance frequencies orthogonally: the curves that represent the effect of each degree of freedom performed in isolation would make right angles with each other. Schroeder (1967) provides a derivation of deformations of a cylindrical tube that are orthogonal (for small deformations) because they only influence one resonance at a time. The shapes are based on cosines. Their effects are illustrated in figure 1B, as the triangles (for the first resonance) and the diamonds (for the second resonance). Although for larger deformations, both resonances are influenced simultaneously, it can be observed that for small deformations (near the point where all lines cross) the lines are exactly horizontal and vertical, thus confirming that they only influence one resonance at a time.

The effects of different combinations of degrees of freedom are illustrated in figure 1C (for reference some Dutch vowels as pronounced by the author are also given). The solid line shows the area of the space of the first and second resonance that can be covered by combining Schroeder's first two orthogonal deformations, under the restriction that the ratio between maximal and minimal cross sectional area is less than 8 (a value that gives acoustic areas that are approximately in the human range). The dotted line shows the area that is covered by a two-tube model where the ratio of the lengths of the tubes can be varied in addition to the ratio of their cross-sectional areas. The same restriction of the ratio of maximal and minimal area applies. Different regions of the acoustic space are covered, but more importantly, the area covered by the two-tube model is substantially smaller than that covered by the model based on Schroeder's deformations.

Figure 1C illustrates the observation that some articulatory systems are better suited to produce a large range of articulations than others. Boë *et al.* (1989) have already observed that different models of the vocal tract produce different ranges of vowel articulations, and Carré *et al.* (1995) have investigated what kinds of vocal tracts result in the largest range of possible speech sounds. This paper focuses on the question what the effect is of larynx position on the range of signals that can be produced by a system that has articulatory constraints similar to those of the human vocal tract.

3 Model

The articulatory synthesizer that is used in this paper is a stripped-down version of Mermelstein's (1973) articulatory synthesizer. The articulatory part of Mermelstein's synthesizer is a geometrical model of the mid-sagittal cross section of the vocal tract (the mid-sagittal plane is the symmetry plane of the human body, and it is the plane that contains most information about an articulation). A geometrical model is a model that approximates the shape of the vocal tract using geometrical shapes (lines, circular arcs, and in the case of Mermelstein's original model also more complicated geometrical objects). The position and orientation of the elements of which the model consists are directly determined by articulatory actions which correspond closely to muscle actions. Because the deformations of a geometrical model are based directly on muscle actions, it is well-suited to investigate articulatory constraints of a given vocal tract anatomy. Other articulatory models, such as

Maeda's (1990) model (or the model based on Schroeder's basis functions, as illustrated in figure 1C) are based on a (linear) superposition of basic deformations of the whole vocal tract. As will be argued in more detail elsewhere (de Boer & Fitch in press) such models can potentially result in articulations that are impossible to produce with muscle actions.

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. As mentioned above, these include the limitation on possible deformations of the vocal tract caused by the limited possible actions of the musculature of the tongue, but they also include an almost constant volume of the tongue and a constraint on producing arbitrarily abrupt transitions in the vocal tract area function. This latter constraint is due to the fact that vocal tracts consist of continuous soft slightly elastic tissue, and that there are only a limited number of muscles that can deform the tongue.

These universal constraints on the kinds of deformations that can be made by mammalian vocal tracts make it impossible to produce exactly the deformations that Carré (2004, 2009) has found to be optimal for communication (even though the modern human vocal tract comes surprisingly close). Therefore it is necessary to use a range of geometric articulatory models to investigate the question of the evolution of the vocal tract.

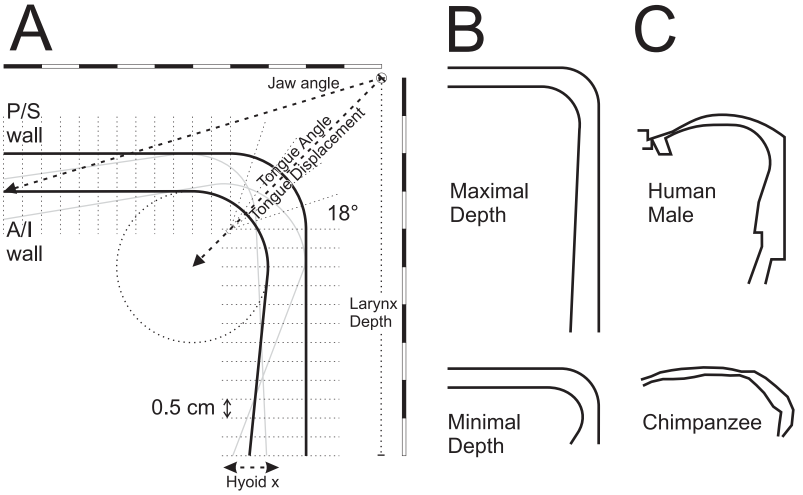

Figure 2: The articulatory model. Diagram A shows the articulatory model (solid outlines) as well as the articulatory parameters, two possible articulations (grey lines), the grid that is used for calculating cross-sections (dotted lines) and scale bars (10 cm). Diagram B shows the system with maximal and minimal larynx depth (with all articulatory parameters set to zero). Diagram C shows mid-sagittal outlines of the human (based on Mermelstein's 1973 model) and chimpanzee vocal tract (Fitch 2000) for comparison. In Diagrams B and C, all tracts are scaled to have the same horizontal extent.

The synthesizer used here does not model the exact anatomy in the region of the lips or in the region of the larynx. This is done to keep the synthesizer as simple as possible, and also to prevent the hard structures of the larynx and epiglottis to interfere with the movement of the tongue. This allows for a larger range of larynx positions than would be possible when using Mermelstein's original synthesizer, or when using Goldstein's (1980) model (that also has variable anatomy) where these hard structures *are* modeled.

The only articulatory motions that were modeled were the motion of the jaw (caused by the mylohyoid and masseter muscles) the motion of the tongue body (both tongue displacement and tongue angle, caused by the styloglossus, genioglossus and hyoglossus muscles) and the horizontal motion of the hyoid (caused by the pharyngeal constrictors, all muscle data from Boersma 1998). Ranges of these parameters are given in table 1. The tract was terminated at the mouth by a vertical plane at a constant position, instead of the more complicated termination model that is used in the Mermelstein model.

The model is illustrated in figure 2A. It consists of a posterior/superior (P/S) wall and an anterior/inferior (A/I) wall. Both consist of two straight lines that are tangent to a circle of 2 cm radius. The P/S wall is static, while the A/I wall can move. The oral termination of the A/I wall has a fixed horizontal position, while its laryngeal termination has a fixed vertical position. The jaw angle determines the vertical position of the oral termination of the A/I wall, while the hyoid horizontal position determines the horizontal position of the laryngeal termination. Jaw angle, tongue body angle and tongue displacement determine the position of the circular arc that describes the tongue body. Its position is calculated in the same way as in Mermelstein's model, by a controllable distance between the tongue center and the jaw joint, and a controllable angle that is added to the angle of jaw rotation. Larynx depth is not an articulatory parameter, but an anatomical parameter (it is fixed once for each instance of the model) and it is measured with respect to the jaw joint (as are all measurements in the Mermelstein model). Therefore the actual length of the horizontal tube is 8 cm, while the

Table 1: Parameter ranges (in cm for displacements and in radians for angles) of the articulatory model.

	min	max
Hyoid horizontal	–1	1
Hyoid vertical	–1	1
Jaw angle	–0.25	0.25
Tongue displacement	–1.5	1.5
Tongue body angle	–0.2	0.2

length of the vertical tube is equal to the larynx depth minus 2 cm. Exact dimensions can be determined from figure 2A, which is to scale.

The cross-sectional area at each point of the vocal tract is approximated by a large number of uniform tubes. Initially, cross-sectional diameters and lengths of each tube are determined by a grid (shown in figure 2A as dotted lines). The diameters are determined by the distance between the P/S and the A/I wall along each grid line, and the length of each tube is determined by the distance between the points halfway the P/S and the A/I walls. Cross-sectional diameters are converted to cross sectional areas in the same way everywhere by squaring the value of the diameter (corresponding to an elongated elliptical cross section).

The frequency response is determined by approximating these as lossless tubes (e. g. Flanagan 1965, section 3.2). The tract is assumed to be ideally closed (maximal pressure, zero flow) at the glottis and ideally open (zero pressure, maximal flow) at the lips. The peaks of the frequency response of this system are reasonable approximations of the resonances of the vocal tract. Frequencies were measured in two ways: one was a straightforward measurement, the other was a normalized measurement. In the normalized measurement, all vocal tracts were scaled to a mean length of 17.5 cm, in order to exclude effects of the absolute frequency range (vocal tracts with lower larynges have larger lengths and therefore lower resonances on average). As there was some length variation of the acoustic pathway for different articulations, scaling was done on the basis of the mean acoustic path length over all articulations for each larynx depth.

The range of possible articulations was explored by generating random articulations that had parameter values in the range given in table 1, and whose resulting acoustic tubes had minimal area of 0.3 cm^2 (with smaller areas turbulence would ensue, resulting in consonant-like signals). The first and second resonances were then determined. A random, rather than a systematic exploration was used for two reasons. First of all, the mapping from articulations to acoustic effects is highly non-linear. Systematic explorations with equally spaced parameter values tend to result in artifacts. The acoustic space is effectively covered by a number of 1-dimensional manifolds (comparable to those of figure 1 A and B) and these do not necessarily reach the most extreme parts of the acoustic space, or give a very representative covering. Secondly, different random subsets give an idea of how spread out the measurements are. A systematic exploration would give only one data point, and it is therefore impossible to get an idea of whether the acoustic ranges of different anatomies are sampled equally well with the same procedure.

In this paper, only the range of formant patterns is measured. Although this is standard practice in measuring abilities of vocal tracts (Lieberman & Crelin 1971, Lieberman, Crelin, & Klatt 1972, Boë *et al.* 2002, de Boer 2009) it should be noted that it might not be the only factor that determines fitness of a given anatomy. Carré (2009) notes that economy of gesture is possibly important for explaining complete systems of vowels, and Lindblom MacNeilage and Studdert-Kennedy (1984) have noted the same for syllable systems. However, what constitutes economy of gesture depends on the precise anatomy of the vocal tract, and also on cognitive factors such as what is easy to perceive, produce and learn. As these are all unknowns, this paper only focuses on the more easily measured extent of the acoustic space. Moreover, it will be argued below that individuals with a larger signaling space will always have the advantage over individuals with a smaller signaling space.

In order to estimate the extent of the vowel spaces for different anatomies, for each larynx depth in the range of 6 cm to 16 cm (in increments of 1 cm) 25 sets of 4000 articulations were calculated. The frequencies in Hertz of the resonances were converted into Bark using the following formula:

$$F_{Bark} = 7\sinh^{-1}\left(F_{Hertz}/650\right) \quad (1.1)$$

(originally proposed by Schroeder, Atal, & Hall 1979). The acoustic space was then divided up into squares of 0.5×0.5 Bark, and the number of squares in which at least one articulation fell was counted. This was used as a measure of the acoustic space covered by each articulatory model. Two other measures were also used, and these are the ranges of the first and the second resonances that can be produced. These were calculated as the difference (in Barks) between the maxima and minima of the first and second resonance, respectively.

4 Results

For each larynx depth from 6 cm (highest) to 16 cm (lowest) with increments of 1 cm (11 depths in total) 25 sets of 4000 random articulations were generated. The distributions of areas, and the extents in the first and second resonances are shown in figure 3, both calculated directly and after scaling to a uniform (average) length of 17.5 cm.

A number of observations can be made from these figures. The most important one is that there is indeed a larynx depth that results in the largest possible acoustic area covered. This is a depth of 9 cm, and this corresponds to an almost equal length of the horizontal and vertical parts of the vocal tract. From a comparison of the graphs of the extents in the first and second formant, it can be observed that the difference in area is caused by different factors for smaller and greater larynx depths. 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. For larger larynx depths both the extents of the first and of the second resonance decrease, but less dramatically so than for smaller larynx depths.

The measures do not have a very large spread for the different samples that were calculated for the same value of larynx depth. This indicates that the trends are real and not due to random fluctuations of the data sampling procedure. To illustrate this, according to

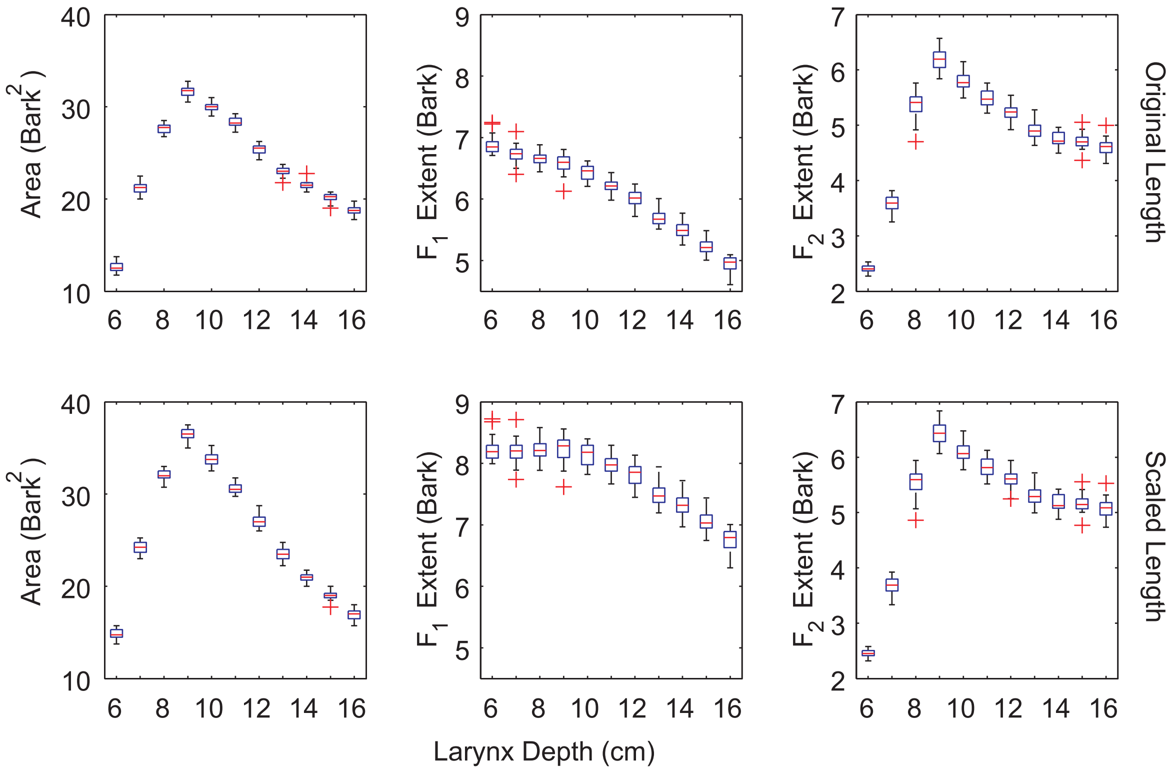

Figure 3: Articulatory extent the model with different larynx depths. The top row presents measurements that have not been corrected for length, the bottom row measurements that have been corrected for length. The left column shows total area covered by the articulations, the middle column the range of the first resonance, and the second column the range of the second resonance. Boxplots show the range from the first to the third quartile, horizontal lines in the boxes indicate the median and whiskers indicate the range of the whole data set. Possible outliers are indicated with crosses.

Wilcoxon's rank sum test, the difference between the larynx depth of 9 cm and larynx depths of 8 cm and 10 cm is significant with $p < 10^{-8}$ for both the original and scaled length cases.

A final observation is that there is no qualitative difference between the measurements of tracts with the original length, and those of tracts with scaled length. In both cases the way the acoustic abilities change with larynx depth is the same. Also, in both cases 9 cm appears to be the optimal larynx depth.

5 Conclusion

From the results it can be concluded that anatomy does matter in the range of speech signals that can be produced. Articulatory control cannot always compensate for the differences in anatomy of the models (and as the essential anatomical constraints are modeled, the same must be true for real primate vocal tracts). Vocal tracts with approximately human anatomy and control over the tongue and the jaw can produce the largest range of speech sounds if the vertical part is approximately as long as the horizontal part. This confirms the conjecture by Lieberman et al. (1972). In the model presented here, the vertical part was slightly shorter than the horizontal part for the optimal tract, but this might be an artifact of the details of the model. Incorporating more anatomical detail in the model is likely to result in small shifts in the ideal larynx position, but is not expected to alter the qualitative finding that there is an optimal position and that at this optimal position the vertical and horizontal parts of the tract are approximately equally long. This is because the anatomical differences will only result in changes in the absolute values for cross-sectional areas at different points in the tract, but will not change the possible deformations that can be made, and it is the qualitative shape of the deformations that determine the range of signals that can be produced.

The optimal tract configuration can approximately generate and combine Schroeder's (1967) first two basic deformations (illustrated in figure 1B). The deformation influencing the first resonance is implemented mainly by the movement of the jaw, while the deformation influencing the second formant is implemented mainly by movement of the tongue body. A higher or a lower larynx position results in deformations that are too asymmetrical to be good approximations of Schroeder's second basic deformation and therefore result in a smaller range of possible second resonance frequencies. The importance of being able to generate such deformations has already been stressed in the work of Carré (2004).

It is interesting to note that the optimal tract configuration that was found with the model corresponds quite closely to the configuration of the female vocal tract (see e. g. Story, Titze, & Hoffman 1998 for images). This lends support to the hypothesis that the human vocal tract has evolved to produce as large a range of speech sounds as possible, given pre-existing anatomy and musculature of the tongue, jaw and larynx. The (adult) male larynx (see e. g. Story, Titze, & Hoffman 1996 for images) on the other hand, which is situated about 2–3 cm lower (e. g. Goldstein 1980, Fitch & Giedd 1999) appears to be somewhat suboptimal, and this would fit with the hypothesis of size exaggeration as put forward by Fitch and Hauser (2002). The reason that the larynx is not as low as found in some animals could then be that with increasing larynx depth, the articulatory range decreases too much, and the adult male human larynx position is therefore a compromise between size exaggeration and articulatory range.

The anatomy of the real human vocal tract is quite a lot more complicated than the model used here (Story, Titze, & Hoffman 1996, 1998). For example, the oral part of the real vocal tract is curved rather than flat. Also, the part right above the vocal tracts (the epilaryngeal tube) is of almost constant cross-sectional area, and there is a sharp transition to the pharynx. In adition, acoustic end effects at the lips are not modeled. However, the curvature of the vocal tract can be ignored at the relevant frequencies, while the epilaryngeal tube and the end effect at the lips are expected to lower formants in real humans, but to have only am insignificant effect on the relative sizes of the ranges of speech sounds that can be produced with the different larynx depths.

In any case, the observations about the optimal larynx position and the relation between male and female vocal tracts agree with the outcome of a study using more realistic articulatory synthesizers (de Boer 2009) as well as with observations of human vocalizations (e. g. Fant

1975, e. g. Diehl *et al.* 1996, de Boer 2009) in which it was found that females appear to have a larger range of vowels than males.

6 Discussion

Apparently, anatomy does matter for producing a large range of speech sounds. However, this does not mean that without a lowered larynx speech is impossible, or that cognitive innovations (such as voluntary control and the ability to do vocal immitation) are not crucial as well. Given sufficient voluntary control, the vocal tract with the highest larynx position could still produce a range of articulations that would be sufficient for a modern language with a vertical vowel system (see the conclusion of Choi 1995 for a short overview). It is an interesting open question what the effect of vocal tract anatomy would be on the range of consonants that could be produced with a vocal tract, but it seems likely that even the most chimpanzee-like anatomy could produce a range of at least 10 different phonemes, which appears to be the lower range of human phoneme inventories.

Evidence of absence of a lowered larynx is therefore no evidence for the absence of speech in an ancestral hominin. However, the argument that modern human language is possible with a limited set of speech sounds is also not an argument against the hypothesis that modern human vocal anatomy evolved for speech. Because modern human anatomy is in a sense optimal for speech, it can be argued on the basis of the results presented here and elsewhere (de Boer 2009) that the modern human vocal tract owes its unique anatomy to the evolutionary pressures derived from vocal communication.

The disadvantage of the human anatomy – the risk of choking – already exists for vocal tracts with higher than modern larynx positions. As soon as there is a gap between the velum and the larynx – something which is necessary for enhanced freedom of motion of the tongue – there is a danger of food falling in the lungs. But once this risk exists, the evolutionary cost has been paid, so to speak, and the larynx position is expected to evolve towards the position that produces the largest range of signals. This happens because a population of communicators can always be invaded by mutants that can produce a slightly larger range of signals (Nowak, Krakauer, & Dress 1999, Zuidema & de Boer 2009). These mutants will be able to produce and perceive all signals that the existing population can produce. In addition, if they produce signals that fall (slightly) outside the range of the signals that the existing population can produce, these will still be perceived correctly by the existing population because of categorical perception. However, these signals will be less easily confused with other signals in the repertoire, because they have a larger acoustic distance to each other. The mutants can therefore produce more distinctive signals, or produce the same signals with less effort, and therefore have an advantage in communication. If communication confers fitness (as it is usually assumed when discussing the evolution of language) mutants with a larger signaling space will invade populations of individuals with a smaller signaling space. In addition, signals will tend to fill the available signaling space through (cultural) self-organization in the population of speakers, based on the pressure to be most easily understood (de Boer 2000).

This provides a path of ever increasing fitness from a chimpanzee-like anatomy to the human (female) anatomy. As has been argued above, if size exaggeration also plays a role, one expects the larynx to lower even more, until a compromise between articulatory range and size exaggeration is reached.

The findings presented here are in contradiction with the findings presented by Boë et al. (2002, 2007) who argued that larynx position is not important for the range of sounds that can be produced. As will be argued elsewhere in more detail (de Boer & Fitch in press) this is likely the artifact of the principal component based model that Boë et al. have used. Their model is based on Maeda's (1990) articulatory synthesizer. This model creates articulations by adding a number of basic deformations that have been derived from modern human articulations. These deformations are similar to Schroeder's (Schroeder 1967) basic deformations for influencing the first and second resonances independently. Boë et al.'s operations for scaling the pharynx and the larynx independently to approximate different vocal tracts do not alter the basic shape of these deformations, and thus allow for qualitatively similar changes to the first and second resonances as the ones that can be achieved by the human vocal tract. However, given the anatomy and musculature of the tongue, the necessary deformations do not correspond to articulations that can be achieved in reality with a higher larynx position and the correspondingly flatter tongue.

Anatomical constraints are important in determining the range of articulations that can be made and it appears that evolution has caused the human female vocal tract to be optimal for producing as large a range of articulations as possible, given the constraints of moving a mammalian tongue. The range of possible articulations for different anatomical configurations is something that is impossible to investigate experimentally or to solve mathematically. Therefore computer simulations are the tool of choice. However, these computer models should implement the basic articulatory and anatomical constraints correctly.

Acknowledgement

Bart de Boer is funded by the Netherlands Organisation for Scientific Research (Geesteswetenschappen), project number 016.074.324.

References

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., HEIM, J.-L., HONDA, K., MAEDA, S., BADIN, P., & ABRY, C. (2007): The vocal tract of newborn humans and Neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. A reply to Lieberman (2007a). *Journal of Phonetics, 35*(4), 564–581.

BOË, L.-J., PERRIER, P., GUERIN, B., & SCHWARTZ, J.-L. (1989). *Maximal vowel space.* Paper presented at the Eurospeech, Paris, France.

BOERSMA, P. (1998): *Functional phonology*. The Hague: Holland Academic Graphics.

CARRÉ, R. (2004): From an acoustic tube to speech production. *Speech Communication, 42*(2), 227–240.

CARRÉ, R. (2009): Dynamic properties of an acoustic tube: Prediction of vowel systems. *Speech Communication, 51*(1), 26–41.

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 II, 320*(série IIb), 471–476.

CARRÉ, R., & MRAYATI, M. (1995): Vowel transitions, vowel systems, and the distinctive region model. In C. e. a. Sorin (ed.), *Levels in speech communication: Relations and interactions* (Amsterdam: Elsevier, pp. 73–89

CHOI, J. D. (1995): An acoustic-phonetic underspecification account of marshallese vowel allophony. *Journal of Phonetics, 23*, 323–347.

DE BOER, B. (2000): Self organization in vowel systems. *Journal of Phonetics, 28*(4), 441–465.

DE BOER, B. (2009): Why women speak better than men (and its significance for evolution). In R. Botha & C. Knight (eds.): *The prehistory of language* (Oxford: Oxford University Press, pp. 255–265

DE BOER, B., & FITCH, W. T. (in press): Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*.

DEMOLIN, D., & DELVAUX, V. (2006): A comparison of the articulatory parameters involved in the production of sounds of bonobos and modern humans. In A. Cangelosi, A. D. M. Smith & K. Smith (eds.): *The evolution of language: Proceedings of the 6th international conference (evolang6)* (New Jersey: World Scientific, pp. 67–74

DIEHL, R. L., LINDBLOM, B., HOEMEKE, K. A., & FAHEY, R. P. (1996): On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics, 24*(2), 187–208.

FANT, G. (1960): *Acoustic theory of speech production*. 'sGravenhage: Mouton.

FANT, G. (1975): Non-uniform vowel normalization. *Speech Transmission Laboratory Quarterly Progress and Status Report, 16*(2–3), 1–19.

FITCH, W. T. (2000): The evolution of speech: A comparative review. *Trends in cognitive sciences, 4*(7), 258–267.

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*(3, Pt. 1), 1511–1522.

FITCH, W. T., & HAUSER, M. D. (2002): Unpacking "honesty": Vertebrate vocal production and the evolution of acoustic signals. In A. M. Simmons, R. R. Fay & A. N. Popper (eds.): *Acoustic communication* (New York: Springer, pp. 65–137

FITCH, W. T., & REBY, D. (2001): The descended larynx is not uniquely human. *Proceedings of the Royal Society of London Series B - Biological Sciences, 268*, 1669–1675.

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

GOLDSTEIN, U. G. (1980). *An articulatory model for the vocal tracts of growing children.* Unpublished PhD, Massachusetts Institute of Technology, Cambridge (MA).

HEIMLICH, H. J. (1975): A life-saving maneuver to prevent food-choking. *Journal of the American Medical Association, 234*(4), 398–401.

LIEBERMAN, P. H. (2007): Current views on Neanderthal speech capabilities: A reply to Boë et al. (2002). *Journal of Phonetics, 35*(4), 552–563.

LIEBERMAN, P. H., & CRELIN, E. S. (1971): On the speech of Neanderthal man. *Linguistic Inquiry, 2*, 203–222.

LIEBERMAN, P. H., CRELIN, E. S., & KLATT, D. H. (1972): Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist, 74*, 287–307.

LINDBLOM, B., MACNEILAGE, P. F., & STUDDERT-KENNEDY, M. (1984): Self-organizing processes and the explanation of language universals. In M. Butterworth, B. Comrie & Ö. Dahl (eds.): *Explanations for language universals* (Berlin: Walter de Gruyter & Co., pp. 181-203

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.

MACLARNON, A., & HEWITT, G. P. (2004): Increased breathing control: Another factor in the evolution of human language. *Evolutionary Anthropology, 13*, 181–197.

MAEDA, S. (1990): 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* (Dordrecht: Kluwer Academic Publishers, pp. 131–149

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

NEGUS, V. E. (1938): Evolution of the speech organs of man. *Archives of Otolaryngology, 28*, 313–328.

NOWAK, M. A., KRAKAUER, D., & DRESS, A. (1999): An error limit for the evolution of language. *Proceedings of the Royal Society of London, 266*, 2131–2136.

OHALA, J. J. (1984): An ethological perspective on common cross-language utilization of f0 of voice. *Phonetica, 41*(1), 1–16.

PUTS, D. A., GAULIN, S. J. C., & VERDOLINI, K. (2006): Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior, 27*(4), 283–296.

SCHROEDER, M. R. (1967): Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of the Acoustical Society of America, 41*(4 Pt. 2), 1002–1010.

SCHROEDER, M. R., ATAL, B. S., & HALL, J. L. (1979): Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In B. Lindblom & S. Öhman (eds.): *Frontiers of speech communication research* (London: Academic Press, pp. 217–229

STORY, B. H., TITZE, I. R., & HOFFMAN, E. A. (1996): Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America, 100*(1), 537–554.

STORY, B. H., TITZE, I. R., & HOFFMAN, E. A. (1998): Vocal tract area functions for an adult female speaker based on volumetric imaging. *Journal of the Acoustical Society of America, 104*(1), 471–487.

ZUIDEMA, W., & DE BOER, B. (2009): The evolution of combinatorial phonology. *Journal of Phonetics, 37*(2), 125–144.