Investigating the acoustic effect of the descended larynx with articulatory models[*]

Bart de Boer
ACLC, Universiteit van Amsterdam

It has been proposed that the low position of the human larynx (compared to other apes) is necessary for producing distinctive articulations, and that it therefore evolved for the purpose of speech. This idea, however, is controversial. Other animals with low larynges have been described, and speech is also possible without using the full range of possible articulations. The role of the descended larynx has been previously investigated with computer models, but these have produced contradictory results. Here it is proposed that for investigating the role of the descended larynx, articulatory constraints must be taken into account, and therefore computer models must be articulatory models. In this paper a strongly simplified model, as well as three more realistic models are investigated for the effect of larynx height. A short study of human data has also been done. It is found in all cases that a vocal tract with a vertical section that is approximately equally long or slightly shorter than the horizontal section performs best. This corresponds to the anatomy of the female vocal tract. An evolutionary interpretation of these observations could be that the female vocal tract has evolved to be optimal for speech, while the male vocal tract has also evolved under another pressure, most likely size exaggeration.

1 Introduction

Did the human larynx descend because of increased articulatory flexibility? This question has long been debated in the study of the evolution of language (DuBrul, 1958; Fitch, 2000; Lieberman & Crelin, 1971; Negus, 1938). Some researchers suggest that the human descended larynx is an evolutionary adaptation to speech (Lieberman & Crelin, 1971), while others suggest that it is

[*] The author was funded by the NWO vidi project “modeling the evolution of speech”. The author wishes to thank Rob van Son and Louis Pols for critical evaluation of a former version of this manuscript.

ACLC Working Papers 2:2, 2007: 61 – 86

the result of other factors, such as bipedal upright walking (Aiello, 1996; DuBrul, 1958) or size exaggeration (Fitch, 2000; Ohala, 1984). Advocates of the adaptive role of the descended larynx in speech point out that it creates room in the pharyngeal region, thus turning the single (oral) cavity vocal tract of other primates into a vocal tract with two controllable cavities: the oral cavity and the pharyngeal cavity. Opponents, apart from proposing other reasons for the descended larynx, put forward that there are many species with descended larynges which do not have enhanced articulatory abilities (Fitch & Reby, 2001), that there are modern human languages that do not use the full range of possible articulations (e. g. Choi, 1991; Ladefoged & Maddieson, 1996, pp. 286-288), and that a descended larynx is not necessary for producing the range of sounds that humans can make (e. g. Boë *et al.*, 2002).

Attempts have been made to build computer models of the acoustic abilities of vocal tracts without a lowered larynx. This is often done together with attempts to reconstruct an ancestral vocal tract, although these are really two logically independent questions. The original example of such work is that by Crelin and Lieberman (Lieberman & Crelin, 1971; Lieberman *et al.*, 1972; Lieberman *et al.*, 1969) who modeled Neanderthal, chimpanzee and rhesus monkey vocal tracts, while more recently Boë and colleagues have made a Neanderthal model (Boë et al., 2002). Similarly, Carré and colleagues (Carré *et al.*, 1995) have made a more theoretical model investigating what vocal tract configuration are needed for producing as distinctive signals as possible. The outcomes of these efforts are contradictory: Crelin and Lieberman find that a lowered larynx is needed and Neanderthals did not have it, and conclude that they were not capable of modern speech. Boë *et al.* find that a lowered larynx is not needed and conclude that Neanderthals were capable of modern speech. Carré *et al.* find that a lowered larynx *is* needed, but they do not take a position on whether Neanderthals had one or not.

Part of the discussion focuses on the reconstruction of the Neanderthal vocal tract. Lieberman’s reconstruction has been criticized and more recent reconstructions tend to propose a more human-like shape of the Neanderthal vocal tract (Arensburg *et al.*, 1989; Houghton, 1993; Schepartz, 1993). Still, as no fossilized Neanderthal vocal tract has been found, no definitive conclusions can be drawn. In any case, the question of whether Neanderthals had a vocal tract similar to that of modern *Homo sapiens* is independent from the question what the function of a lowered larynx is.

In the debate about the function of a lowered larynx, there is no disagreement about the basic acoustics: every researcher agrees that a vocal tract with two independently controllable cavities is needed to produce a maximally distinctive set of speech sounds. Rather, the debate focuses on whether a vocal tract that is anatomically like a modern human vocal tract is needed or whether, with sufficient (voluntarily) articulatory control, an ordinary primate vocal tract can produce the same range of articulations as a modern human vocal tract. Simplifying the debate, one could say that Lieberman et al.'s papers stress the importance of anatomy (most recently made by Lieberman (2006) in reference to Riede et al.'s (2005) model of a Diana monkey), whereas Boë et al.'s papers stress the importance of articulatory control, and propose that sufficient control can overcome the limitations of anatomy. Carré et al.'s contributions do not take a strong position in this debate, but stress the necessity to have two independently controllable cavities, but leave in the middle whether this is achieved through better articulatory control or through specialized anatomy.

In this paper the role of anatomy (in this case, the larynx position) versus articulatory control is investigated with a number of simplified models of the human vocal tract. The question underlying the research is: do differences in anatomy influence the ability of a vocal tract to produce different speech sounds, or are any differences in anatomy offset by articulatory control? First, the effect of larynx height on the range of acoustic signals that can be produced is studied in a highly simplified model of the vocal tract. Second, the articulatory abilities of more realistic models of the human vocal tract are studied. These models are based on the male and female vocal tracts, and it is investigated whether they show the same differences as were found in the simplified model. Finally, it is investigated whether a similar effect can be detected in real human data.

2 Basic Methods

As the most important aim of this paper is the theoretical investigation of the role of larynx position on the range of signals that can be produced, and as we cannot independently control larynx height in human subjects, computer models of the vocal tract were used. Another reason to use a computer model is that analytic approaches can only provide limited insight in the case under study. The acoustic effects of certain changes in configuration can be calculated analytically (such as the effect that the resonance frequency of a tube increases if the length is reduced). However, here it is investigated what the effect is on the *range* of signals that can be produced when the way a tract can be *deformed* is changed. This is a complex manipulation of boundary conditions and outside the abilities of ordinary mathematical analysis. Computer modeling is therefore the right tool to study this question.

The experiments presented in this paper are based on the use of geometric articulatory computer models. Such models represent the vocal tract as a number of geometric shapes that can be manipulated with articulatory parameters. The geometric shapes correspond directly to parts of the vocal tract, such as the pharynx, the tongue body, the palate, etc. The articulatory parameters can be mapped straightforwardly to muscle actions. Such models are the closest approximations one can make to the actions of real vocal tracts. The models used in this paper are based on Mermelstein's (1973) model. His model is a 2-dimensional model of the mid sagittal cross section of the vocal tract, as well as a model of the area of the three-dimensional cross sections at different positions in the vocal tract. Another geometrical model is Goldstein's (1980) model. Although this model has been designed to be able to model male and female vocal tracts, it is not possible to keep the upper vocal tract constant while changing the position of the larynx. Therefore the Mermelstein model was used and modified for the research presented here.

Generating a signal on the basis of an articulatory model involves three steps. The first step is to calculate the 2-dimensional mid sagittal outline of the vocal tract for a given set of articulatory parameters. The second step consists of converting this two-dimensional outline in a function giving the cross-sectional area of the vocal tract at each point along its length. The third step consists of calculating the acoustic properties of a tube with this particular area function. When the vocal tract is modeled as a series of concatenated lossless tubes, this last step can be performed by straightforward application of standard acoustics (Fant, 1960; Flanagan, 1965, section 3.2). The first two steps, however, depend on the anatomy of the vocal tract. Details of the different models will be given in the sections below.

Given an articulatory model, the question rises of how to explore and measure its articulatory abilities. The signal that would be produced by an individual articulation of a given model can be calculated from the lossless tube approximation mentioned above. From this articulation, the position of the resonance peaks can be calculated. Only the first and the second resonances (formants) were used in the analysis. It is well established that although three formants are needed to make all differences between possible vowel signals, the first two formants are the most important cues for establishing vowel quality. Furthermore, the first two formants have been widely used as the basis of perceptual acoustic space. This has not only been the case in research into the evolution of speech (Boë et al., 2002; Carré et al., 1995; Lieberman & Crelin, 1971), but also in research into perception (e. g. Peterson & Barney, 1952) or acquisition of speech (e. g. Kuhl *et al.*, 1997; Kuhl & Meltzoff, 1996).

Every articulation can therefore be considered to result in a point in a two-dimensional acoustic space. The comparison between two different articulatory models then boils down to a comparison of the extent of the areas in acoustic space covered by these two models. Potential measures of this extent are the total area, or measures of its maximal diameter. Here, the measures that are used are the area of acoustic space covered by the articulations a model can make, and the difference between the maximal and minimal formants generated by the model. The area covered by the articulatory model was calculated by dividing the acoustic space in a grid of squares of 0.5×0.5 Barks, and counting the number of tiles that had at least one acoustic signal in them. The number of tiles can then be converted to acoustic area by multiplying with the area of a tile (0.25 Bark^2).

The procedure for exploring the range of possible articulations is inspired by, but not exactly equal to the idea of Maximal Vowel Space as defined by (Boë *et al.*, 1989). It consists of generating a large number of articulations, and calculating what area of acoustic space they cover. It might appear that a systematic exploration of every possible combination of articulatory parameters would be most straightforward. The continuous ranges of articulatory parameter values could be divided into a number (say 10) of equally spaced values, and all possible combinations explored. This approach suffers from two problems, however. The first is that the number of articulations to be explored rises exponentially with the number of articulatory parameters. With 3 articulatory parameters, a thousand articulations must be explored, but with 6 parameters, a million need to be explored. This is infeasible. The second problem is that, due to the discretization of the parameter range, a bias might be introduced. It is quite possible (even likely) that articulations resulting in extreme values of the signal are missed.

A better approach, but at first counterintuitive one, is therefore to generate a large number of random articulations. This is called a Monte Carlo approach in computer science (Metropolos & Ulam, 1949). This approach does not suffer from sampling biases. An added advantage is that the procedure can be repeated a number of times, and the spread of the results be used to get an indication of how well the space is sampled. The "systematic" approach, on the other hand, would always give the same value, and therefore no idea could be obtained of how well the space is sampled.

Only valid articulations were used for calculating the acoustic area. Valid articulations are articulations were the articulators do not intersect. The ranges of the articulatory parameters were selected to be physiologically plausible, but it still remains possible that a combination of articulatory values results in parts of the vocal tract intersecting with each other. This was automatically detected when calculating the value of the cross section areas of the vocal tract. When the diameter was less than zero, intersections occurred, and such articulations were discarded. Only articulations where all cross sections have an area of at least 0.3 cm^2 were considered. Given typical airflow rate and a constriction length of 5 cm, smaller areas would cause turbulence (the Reynolds number would be over 2000), and therefore would result in fricatives or fricative vowels. By using this minimal area, the acoustic simulations were therefore limited to airflow without turbulence. Repeating the experiments with 0.1 cm^2 and 0.5 cm^2 did not change the qualitative results (although the absolute size of the acoustic space obviously changed).

Finally, the values obtained for the extent of the acoustic space depend on the representation of the formant values. Basic acoustic theory shows that shorter vocal tracts result in higher formants. All else being equal, shorter tracts would therefore result in apparently larger extents in acoustic space. In order to compare the abilities of vocal tracts with different geometries, they would therefore have to be normalized to the same length. It turns out that taking the logarithm of the formant frequencies gives the same result. Furthermore, it turns out that human perception is to good approximation logarithmic as well. Weber’s law of perception (Weber, 1834) states that the just noticeable difference between two signals is proportional to the value of these signals. This means that the size of the just noticeable difference is constant when taking the logarithm of signals. In other words, differences between the logarithms of signals give a reasonable measure of their perceptual distance, independent of the actual value of these signals.

A problem occurs at low frequencies, for which Weber’s law does not hold completely. A scale that takes this into account is the Bark scale. This scale is also used by other researchers in the field (e.g. Boë et al., 2002). It is also prudent from an evolutionary point of view to use a perceptually accurate scale, as there are indications that perception of speech was already similar for Neanderthals and *Homo sapiens* (Martínez *et al.*, 2004) and it appears as if the basics of perception are much older than any differences in vocal anatomy (Smith & Lewicki, 2006). The exact relation between Hertz and Bark was adopted from (Schroeder *et al.*, 1979; Schwartz *et al.*, 1997) and is as follows:

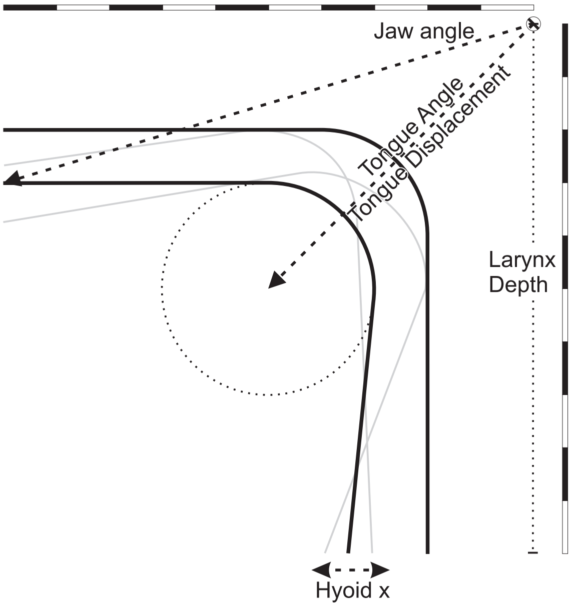
$$F_{Bark} = 7\sinh^{-1}\left(F_{Hertz}/650\right)$$

Repeating the experiments and making measurements in either the Mel using the ordinary logarithm of frequency did not result in any qualitative differences (although the exact numerical value and the significances did change).

3 The Simplified Model

The simplified model is a stripped-down version of Mermelstein’s model. The details of Mermelstein’s model are given in appendix A, so only the differences between the simplified and the original model will be discussed. The simplified model does not model the exact anatomy in the region of the lips, nor does it model the exact anatomy in the region of the larynx. This is done to keep the model as simple and symmetric as possible, so that the influence of larynx height is the determining factor of the model's behavior. Furthermore, 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. The tract was terminated at the mouth by a vertical plane at a constant position, instead of the more complicated plane that is used in the Mermelstein model. Also, some of the dimensions of the model were simplified somewhat. Finally, cross-sectional diameters are converted to cross sectional areas in the same way everywhere by squaring the value of the diameter.

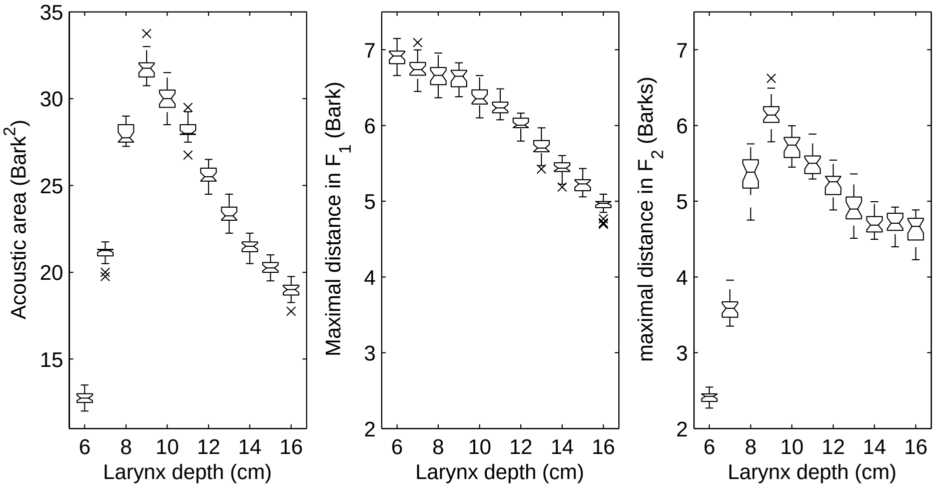
Figure 1: The simplified model. The outline of the model is shown in bold black lines. The articulatory parameters are shown as dashed arrows. The circle that is the basis of the tongue contour and the larynx depth are given as dotted lines. Two potential articulations (different from the rest position) are shown as thin grey lines. For scale, horizontal and vertical bars of 10 cm length are given.



The model is illustrated in figure 1, which can be compared to the more realistic model in figure 3. The ranges of the articulatory parameters are the same as for the Mermelstein model, and are given in table 3 in appendix A. The ranges were determined on the basis of what appeared anatomically plausible, and on the basis of what values did not result in impossible configurations (e.g. ones that intersected themselves). It should be noted that larynx depth 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 length of the vertical tube is equal to the larynx depth minus 2 cm (this is done in deference to Mermelstein's original coordinate system, that was relative to the jaw joint).

Models were investigated with larynx depths ranging from 6 cm to 16 cm in increments of 1 cm. For every larynx depth, 100 000 random articulations were generated (where articulatory parameters were uniformly distributed over their range). These were divided into 25 groups of 4000 articulations. For each of these groups the acoustic area and the ranges of F_1 and F_2 were calculated. The result is shown in figure 2.

Figure 2: The relation between acoustic area (left plot) and maximal distance in F_1 and F_2 (middle and right plots) and larynx depth in the simplified model. The box plot shows the median (horizontal line) as well as the first and third quartiles (top and bottom of the boxes). The total extent of the data set is indicated by the whiskers, while points that can be considered outliers are shown as crosses. Notches in the boxes indicate statistical significance; if the vertical range of the notches of two boxes do not overlap, their difference is significant at the 5% level.



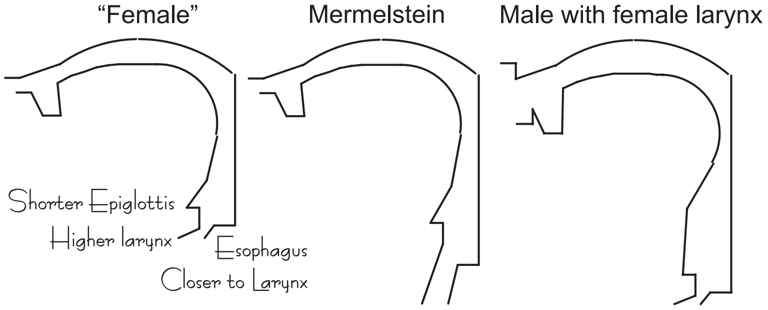
Note that a *smaller* value for larynx depth means a *higher* larynx. It is clear from this figure that the vocal tract that covers the largest acoustic range is the one with a larynx depth of 9 cm. It can also be observed that both higher and lower larynges result in significantly (using the Wilcoxon rank sum test at 5% confidence) smaller reachable areas of acoustic space. There appears to be an optimal larynx depth that is approximately equal to (but in the case of this model slightly smaller than) the horizontal dimension of the vocal tract. It is interesting to note that the difference is mainly due to the inability of models with a lower larynx to produce distinctions in the second formant. As for producing distinctions in the first formants, higher larynges actually appear to be very slightly better than models with medium or low larynges.

4 The Realistic Model

Having established that larynx depth influences the signal range of a simplified articulatory model, one can wonder whether this is an artifact of the model, or whether a similar effect obtains in human vocal tracts. Here, no attempt is made to reconstruct a fossil vocal tract. Instead, it is attempted to model the differences in articulatory ability that result from the different larynx positions in human male and female vocal tracts[1].

Mermelstein's model is of the male vocal tract. A reimplementation of his model was therefore used for modeling the male vocal tract. For the female vocal tract, however, his model needed to be modified. As the primary reason for building a female model was to investigate the role of a lowered larynx, as few changes as possible were made to the original model. Using data by Fitch and Giedd (Fitch & Giedd, 1999) as well as Story's data (Story *et al.*, 1996, 1998) it was estimated that the female larynx lays approximately 2.2 cm higher than the male larynx. This corresponds well with the 2.8 cm difference in Goldstein's (Goldstein, 1980) model. It should be noted that when the position of the larynx is mentioned, this is in reference to the larynx at rest. In the Mermelstein model (and in contrast to the simplified model), the larynx *can* move vertically as a result of the motion of the hyoid (caused by the sternohyoid and stylohyoid muscles). This is restricted to ± 0.5 cm.

Figure 3: Comparison of the original Mermelstein (1973) model (middle) to the model with the raised, "female" larynx (left) and the mixed model with male position and female shape (right). The differences between the models are indicated in the female model.



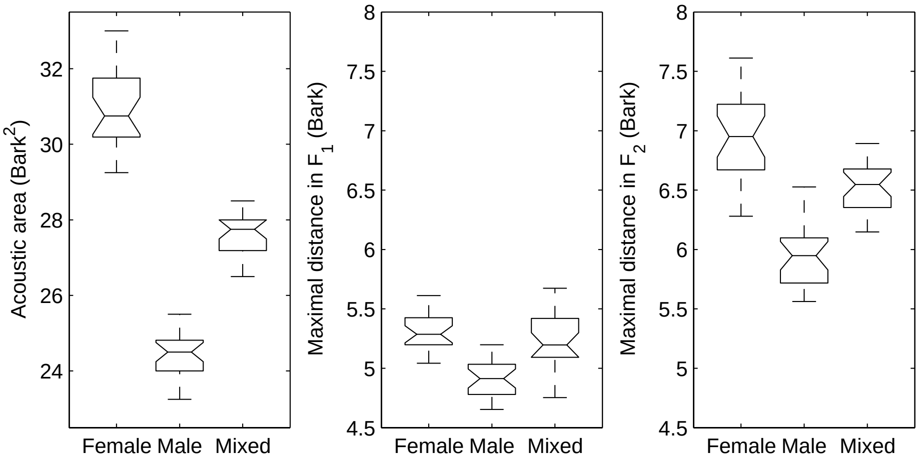
[1] Of course, there is significant individual variation in larynx position and exact vocal tract area function. As the aim of this paper is to investigate the effect of larynx height, only two typical models are investigated, based on data of typical males and females as published in the literature.

There are some other, smaller differences in anatomy as well. Most importantly, the epiglottis is smaller and the esophagus is closer to the larynx in females than in males (Negus, 1949, chapter 11). The larger male epiglottis would not fit in the Mermelstein model with the higher larynx position. In the model used here, the female epiglottis extends upwards 1.7 cm less than the male epiglottis. Finally, also based on Negus's drawings of dissected human larynges, (Negus, 1949, figure 189) the esophagus is modeled to extend 1.3 cm less above the larynx in the female model than in the male model. All these differences are illustrated in figure 3. These differences result in not just a length difference, but also in a different area function, and therefore different volumes of the female and male pharynx. Although this is realistic, it is nevertheless interesting to compare the effect of the lowered larynx alone. A third model was therefore built with a female larynx/epiglottis/esophagus anatomy at the position of the larynx in the male model. This is called the mixed model.

In order to convert the 2-dimensional cross section into a 3-dimensional area function, the same conversion functions are used for the male and the female models. A comparison of area functions derived from MRI-scans of the supralaryngeal vocal tracts of a male subject (Story et al., 1996) and a female subject (Story et al., 1998) articulating the same vowels has shown that most of the difference occurs in the pharyngeal part of the vocal tract. Given that there is considerable inter- and intrasubject variation when producing vowel articulations and that it is unclear whether there are systematic differences between male and female area functions (Soquet *et al.*, 2002) no attempt was made to model the differences in oral vocal tract area function between the male- and female subject. This also minimizes the differences between the models and allows for a purer comparison of the role of the position of the larynx. Details of the model are given in appendix A.

With all three articulatory models, 25 sets of 4000 articulations each were generated. For these articulations, the first and second formant were calculated in Barks as described above. The results are presented in figure 4. All differences are significant with $p < 0.01$, according to the Wilcoxon rank sum test, except for the first formant of the mixed and female models, where there is no significant difference. Apparently the male model, with its lower larynx is able to cover a somewhat smaller range in acoustic space than the female model. This space is even smaller than the space covered by the mixed model, although this model also covers a smaller part of acoustic space than the model with the higher larynx. This is in agreement with the findings of the simplified model.

Figure 4: Acoustic area and extent of the first and second formant of the more realistic models. Note that the female model is significantly better in all respects than the male model.



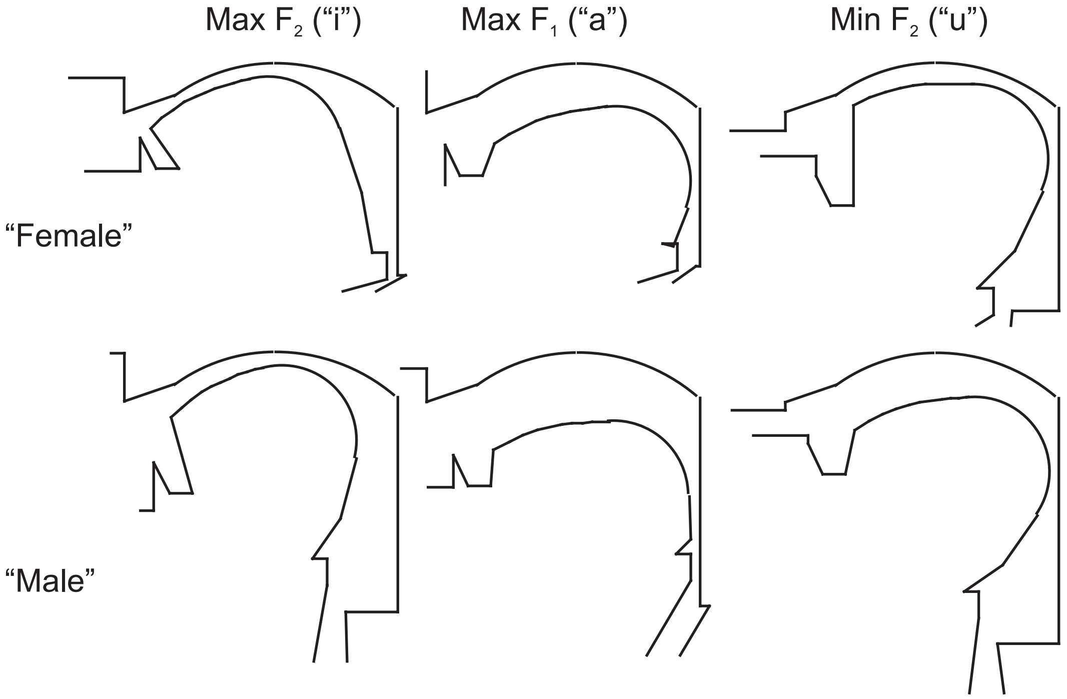
It is important to note that the numbers in figure 4 do not represent a good estimate of the total extent of the acoustic capabilities of the models. Due to the nature of the sampling procedure, the values are always undersampled. However, this has the same bias for every model, and therefore comparisons between models that have been sampled in the same way are valid. In order to get an idea of the total extent of the acoustic space of all models, the values for the complete data set of 100 000 articulations per model can be calculated. These values are given in table 1.

Table 1 Values of area and maximal extents for the complete data sets.

	Female	**Male**	**Mixed**
Area (Bark^2)	40	30	36
Max. F_1 size (Bark)	5.8	5.4	5.9
Max. F_2 size (Bark)	7.8	6.9	7.5

In order to check the validity of the calculated data points, it is necessary to verify whether they correspond to realistic articulations. Of course, this is impossible to do for all 100 000 data points of the data sets. It was therefore only done for the points with the highest and lowest second formant and for the point with the highest first formant for all 100 000 data points. These points correspond roughly to [i], [u] and [a], respectively, and were expected to have the most extreme articulations. They were therefore the articulations that were most likely to be unrealistic. Images of the vocal tract configuration of the male and female models are given in figure 5. The articulations appear to correspond well with the articulations humans make when producing these vowels, with the possible exception of the female articulation with maximal F_2. This articulation appears to have lips that are protruding more than would be the case in a human articulation of [i]. This is most likely an artifact of the undersampling of the available articulatory space by the Monte Carlo method. The articulation with less protruding lips would probably have even higher F_2, but it was not sampled. articulations are also plausible[2].

Figure 5: Comparison of articulations with maximal F_2 and F_1 and minimal F_2. Note that for both the male and female models, the articulations correspond well with the articulations for [i], [a] and [u]. There appear to be discontinuities in the outline drawn, but this is an artifact of the requirement of the programming environment to use integer values for plotting.



5 Real Human Data

All results so far have been obtained with highly stylized models of the vocal tract. It is therefore instructive to check what happens when applying the same measurements to real human data. The effect of lowering of the larynx on the reachable acoustic space can be estimated by comparing the acoustic spread of vowels produced by men with those produced by women.

[2] Hundreds of random articulations were also inspected visually, and none was impossible, although some articulations are at the extreme of what is comfortable, and thus are unlikely to be used in ordinary speech.

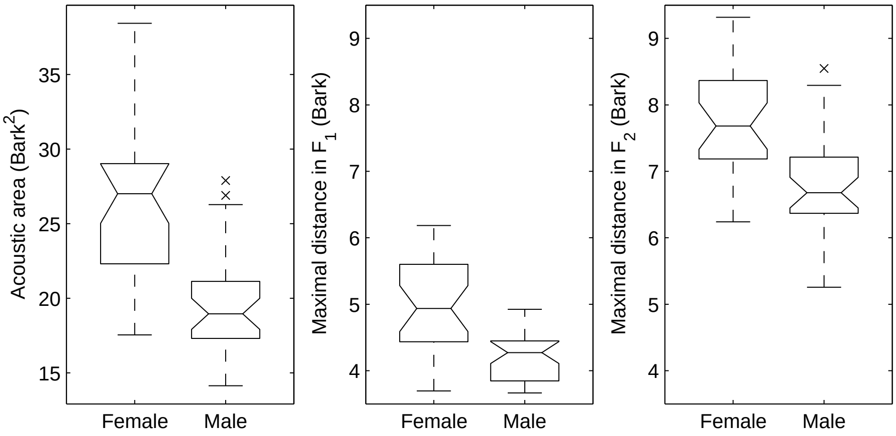
Studies of different languages show consistently larger vowel spaces for female articulations than for male articulations (e. g. the data presented in Fant, 1975). Although there is a debate to what extent these results can be explained by behavioral or by anatomical factors (e. g. Diehl, 1996; Goldstein, 1980) the consensus seems to be that anatomy is at least partly the cause.

In order to show that the methods used in this paper also work on real data, the classic data set from Peterson and Barney (1952) as reconstructed and made publicly available by Watrous (1991) has been used. This data set is ideally suited for this research, as it contains all data points of all speakers. Comparing acoustic ranges that are assumed to be caused by individual differences in anatomy, cannot reliably be done using averages. The surfaces of the acoustic space (consisting of the first and second formant frequencies in Bark) and the maximal distances in the first and second formant were calculated.

It must be noted that this comparison is different in nature than the one performed on the model-generated data. In the case of the model-generated data, different data sets generated by the same model were compared, whereas in this case, different data sets generated by different speakers are compared. Also, the Peterson and Barney data set only contains 20 vowels per speaker. This makes it impossible to use the same procedure to calculate area as was used in the model study. It was therefore decided to use the *convex hull* to calculate the areas of the vowel spaces of the human speakers. The convex hull is the area that is covered by all linear interpolations between all data points (Cormen *et al.*, 1993, section 35.3). It can also be imagined in the following way: represent every data point by a nail in a flat board, and then stretch a rubber band around the collection of nails. The area inside the rubber band is the convex hull. It was found that in the model-generated data, the area of the convex hull correlates almost perfectly with the area calculated with the grid-based method, although it is systematically higher. This is because the convex hull always includes at least as much or more area than is really covered by a data set. However, when comparing data that is calculated with the same method, this should not be a problem. Data generated by the two different methods should not be compared directly, however.

The different measures of the human data were compared using the Wilcoxon rank sum test. It was found that the female vowel spaces were larger than the male vowel spaces with $p < 0.001$ for all measurements. The measures of male and female vowel spaces are given in figure 6. From this figure it is directly clear that female vowel spaces tend to be somewhat larger than male vowel spaces. Although there might be other reasons for the difference between male and female performance, at least the human data does not contradict the observation from the modeling study that vocal tracts with a larynx in the female position have greater articulatory abilities than vocal tracts with the larynx in the male position. It should be noted, however, that the difference is small, that there is considerable overlap and that it is not expected that men would have practical difficulties with producing distinctive speech because of this difference.

Figure 6: Statistics of the comparison of human female and male data, taken from Peterson and Barney's dataset. Note that absolute values should not be compared with those of the modeling experiments, as the test conditions and the way of measuring acoustic area are different.



6 Discussion

In this paper the influence of larynx position on the reachable acoustic space has been investigated. In a strongly simplified computer model of the human vocal tract, it was found that there is an optimal vertical position of the larynx, for which the area in acoustic space covered by the signals that such a model can generate is maximized. In the model the optimal position occurred when the vertical part of the vocal tract was slightly shorter than the horizontal part. The same results were found when more realistic models (based on Mermelstein's model) of the male and female vocal tracts were compared. Here too, the model in which the vertical part of the vocal tract was slightly shorter than the horizontal part could generate a larger repertoire of signals than the model in which the vertical part was longer. A similar difference was found in the Peterson and Barney data set of vowels articulated by male and female speakers.

The modeling results show that articulatory constraints caused by differences in anatomy can influence the range of articulations that can be produced. The simplified models all had the exact same articulatory control, but differed only in the position of the larynx. The more realistic models also all had the same articulatory control, and the mixed model and the female model only differed in the position of the larynx. The male model also had slightly different anatomy in the laryngeal region.

The reason for this difference is that certain configurations of the modeled vocal tract allow for a range of deformations that results in a larger range of acoustic signals, given the way in which the different articulators (most importantly tongue, lips, pharynx and larynx) can be controlled. Apparently, given human-like abilities to control articulation, tracts in which the vertical part of the vocal tract is about equally long as the horizontal part, allow for the greatest range of signals. This agrees with Lieberman et al.(1969)'s analysis.

The obvious question is of course, whether this is also true for the case of human articulation. The results from the comparison of the male and female speakers of the Peterson and Barney data set show a difference that is very similar to the one found between the models of the male and female vocal tracts. Still, it remains possible that this difference is caused by other factors, such as sociolinguistic ones, or biases in measurement. One way to test this is to investigate the way articulatory capabilities change over puberty. The anatomy-is-important hypothesis would predict that boys and girls would be indistinguishable before puberty, but different after puberty. Although research into the effect of development on formant frequencies exists (Lee *et al.*, 1999), their paper only presents averages over age groups, and not individual data points. As the hypothesis presented here is that for males the articulatory range changes over puberty, these averages would be over a multimodal distribution in the critical age group. Therefore the data in (Lee et al., 1999) cannot be used directly. However, given the correct data, the influence of the descent of the larynx in puberty it is empirically testable.

The results do make perfect sense from an evolutionary perspective, however. The female larynx position appears to be very close to the one found to be optimal in the simplified model. This is an argument in favor of the hypothesis (Lieberman et al., 1969) that the human larynx position is optimized for producing as distinctive articulations as possible. The fact that the male larynx is slightly lower than the optimal position could be explained evolutionarily by the fact that this helps to exaggerate size (Fitch, 2000; Ohala, 1984). It has been found that this is important for animals, and it has also recently been found that lower formants help human males to impress other human males (Puts *et al.*, 2006). That the human male larynx is not as low as found in certain animals (Fitch & Reby, 2001) can then be explained by the fact that the male vocal tract needs to remain able to produce a sufficient repertoire of distinctive speech sounds.

Of course this does not mean that human males are necessarily worse at communicating through speech. Indeed, humans can still speak even with deformations of the vocal tract, such as cleft palate, or for that matter, with their mouth full of food. Also, the fact that not all languages use the full extent of possible speech sounds (Choi, 1991; Ladefoged & Maddieson, 1996) indicates that maximal articulatory capabilities are not essential for modern languages. However, evolution (both cultural and biological) is very good at fine-tuning. Most languages do use the maximally distinctive vowels, for example, and this can be explained as the result of cultural and functional evolution (e. g. de Boer, 2000; Schwartz et al., 1997). As for biological evolution: it is easier to produce distinctive speech when one has the right anatomy. If communication is an important factor in survival, then the larynx position that has the best position for communication will therefore be selected for. The findings of the simplified model indicate that for extremely high larynges (comparable to the chimpanzee vocal tract) small differences in larynx height already make an important difference in useable acoustic space.

The findings of the more complex model and the human data seem to indicate that in human females, the evolutionarily optimal position is near the position that results in the largest range of possible speech sounds, while in human males the evolutionarily optimal position is slightly lower, resulting in lower formants and a more impressive voice.

These results, although certainly not the last word on the role of the descended human larynx, contribute to the debate on whether anatomical constraints are important in understanding the evolution of the vocal tract. They provide an argument that anatomy does matter and that the human vocal tract has a shape that facilitates speech production. It also provides an argument that the female vocal tract is probably the best point of reference when investigating the link between anatomy and distinctive speech.

7 Appendix A: Details of the articulatory model.

The articulatory models used in this paper are based on a reimplementation of Mermelstein's (Mermelstein, 1973) model. The reimplementation was based on the description provided in that paper, as well as on the reimplementation by Boersma (Boersma, 1998). However, as the description in Mermelstein's paper was not quite complete, some of the details of the implementation had to be measured from his figures. As the model used here can therefore be somewhat different in its details from Mermelstein's original model, a short description is given. The male and female models are identical, except for a number of parameters describing the shape of the larynx. These can be found in table 2.

The vocal tract outline in the midsagittal plane is approximated by two composite curves consisting of straight lines, circular arcs and a more complex curve to approximate the tongue. The exact shape of these curves is determined by eight articulatory parameters. These are the *x* and *y* position of the hyoid (x_H and y_H) the angle of the jaw (α_J) the angle and displacement of the tongue (α_T

and d_T) the angle of the tongue blade (α_B) and the protrusion and spread of the lips (x_L and z_L). The geometry and the relation of the parameters to the model are illustrated in the left part of figure 7. The minimal and maximal values of the parameters are given in table 3.

Table 2 The constants used for calculating the outline of the male and female vocal tracts (all lengths in cm, all angles in radians).

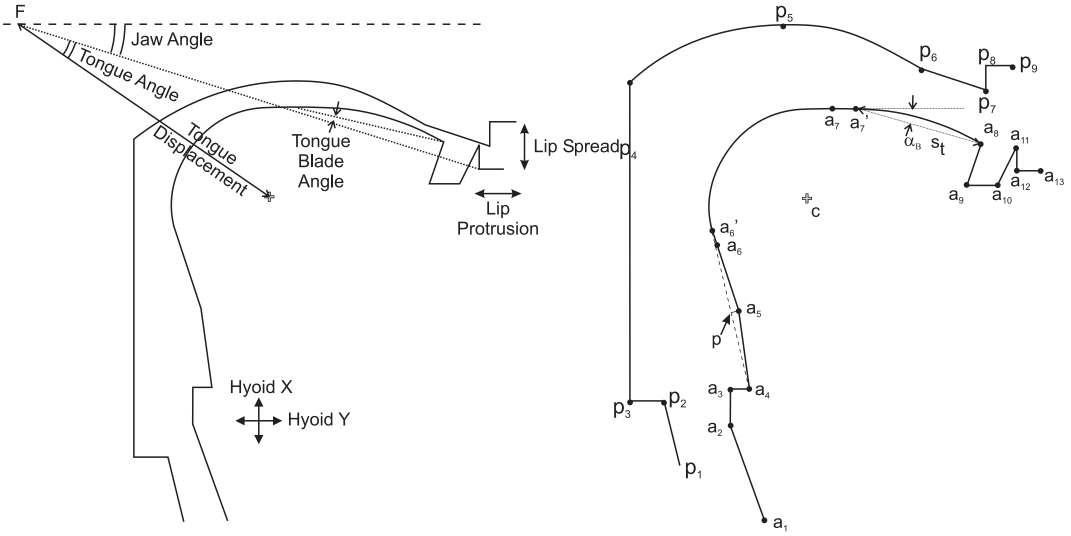
	shared	**male**	**female**		**shared**
H_x	5.6			$H1_x$	–0.4
H_v		–7.6	–7.1	$H1_v$	–0.7
HK_x	0.3			S	11.3
HK_v		–2.7	–1	A_i	–0.237
W	0.9			A_c	0.29
G_x	4.7			A_h	1.73
G_v		–9	–7.7	R_t	2
R_x	4			S_t	3.4
V_v	–2.6			$LT2_x$	–1
M_x	7.2			LT_x	–0.4
M_v	–1.4			LT_v	–0.8
N_x	9.87				
N_v	–2.27				
U_x	11.2				
U_v	–2.7				

Table 3 The minimal and maximal values of the articulatory parameters.

	min	max
x_H	–1	1
y_H	–1	1
α_J	–0.25	0.25
α_T	–0.25	0.25
d_T	–1.5	1.5
α_B	–0.2	0.2
x_L	0	1.5
y_L	–1.5	1.5

The posterior superior outline is defined by nine points $p_1 \ldots p_9$. The anterior inferior outline is defined by thirteen points $a_1 \ldots a_{13}$. The shape of the outlines and the positions of the points are illustrated in the right part of figure 7. The origin of the coordinate system for locating the points is the point around which the jaw rotates, indicated in figure 7 as point F.

Figure 7: The articulatory parameters (left) and the calculated points on the outlines (right). Note that the posterior/superior outline (with $p_1 \ldots p_9$) and the anterior/inferior outline (with $a_1 \ldots a_{13}$) are only shown in the correct relative positions in the left image.



The points on the posterior/superior outline are calculated as follows:

$$
\begin{aligned}
p_1 &= \left(x_H/2 + H_x + HK_x - W, y_H + H_y + HK_y\right)\\
p_2 &= \left(x_H + G_x, y_H + G_y\right)\\
p_3 &= \left(R_x, y_H + G_y\right)\\
p_4 &= \left(R_x, V_y\right)\\
p_5 &= \left(M_x, M_y\right)\\
p_6 &= \left(N_x, N_y\right)\\
p_7 &= \left(U_x, U_y\right)\\
p_8 &= \left(U_x, U_y + z_L\right)\\
p_8 &= \left(U_x + x_L, U_y + z_L\right)
\end{aligned}
$$

where the constants are given in table 2. All segments are straight lines, except for the segments p_4–p_5 and p_5–p_6. These are circular arcs that are horizontal in point p_5. The points for the anterior/inferior outline are slightly more complicated to calculate. The first four points are straightforward:

$$
\begin{aligned}
a_1 &= \left(x_H/2 + H_x + HK_x, y_H + H_y + HK_y\right)\\
a_2 &= \left(x_H + H_x + H1_x, y_H + H_y + H1_y\right)\\
a_3 &= \left(x_H + H_x + H1_x, y_H + H_y\right)\\
a_4 &= \left(x_H + H_x, y_H + H_y\right)
\end{aligned}
$$

points a_5 and a_6 are calculated with the aid of point a_6'. This point is the point where a line from a_4 is tangent to the circle describing the tongue body. The tongue body is a circle with radius R_t and center point (c_x,c_y). The position of the tongue body depends on the motion of the jaw and the tongue as follows:

$$\begin{aligned} c_x &= (S + d_T)\cos\left(A_j - A_c - \alpha_J - \alpha_T\right) \\ c_y &= (S + d_T)\sin\left(A_j - A_c - \alpha_J - \alpha_T\right) \end{aligned}$$

The point p is midway between a_4 and a_6'. Point a_5 is calculated using a line perpendicular to the line a_4–a_6'. The length of this line is based on the length of the line a_4–a_6' as follows: $0.57\cdot\left(\left|a_6' - a_4\right| - 3.48\right)$. Note that when the distance is short, a_5 moves to the left and when it is long, it moves to the right. Point a_6 is now the point on the line tangent to the tongue body, through a_5.

Point a_7' is the point around which the tongue blade rotates, while point a_7 is the point where the tongue blade connects to the tongue body. Point a_7' is oriented at a fixed angle with respect to the tongue. As the tongue is rotated by jaw movement, the value of this angle in the absolute coordinate system is changed by jaw movement. The position of point a_7' is calculated as follows:

$$a_7' = \left(c_x + R_t\cos\left(A_j + A_b + \alpha_J\right), c_y + R_t\sin\left(A_j + A_b + \alpha_J\right)\right).$$

The tip of the tongue, point a_8 is at a fixed distance from this point, and its position is calculated as follows:

$$a_8 = a_7' + \left(S_t\cos\left(A_j + \alpha_J + \alpha_T - \alpha_B\right), S_t\sin\left(A_j + \alpha_J + \alpha_T - \alpha_B\right)\right)$$

the starting point of the tongue blade is point a_7. This is calculated such that the tongue blade starts tangent to the tongue body, as follows:

$$a_7 = \left(c_x + R_t\cos\left(A_j + A_b + \alpha_J + \alpha_T - \alpha_B\right), c_y + R_t\sin\left(A_j + A_b + \alpha_J + \alpha_T - \alpha_B\right)\right).$$

It is possible that in this respect the model described here differs slightly from Mermelstein's model, as the starting point of the tongue blade is somewhat unclear in his description.

In order to calculate the final points on the anterior-inferior outline, the motion of the jaw needs to be taken into account. In Mermelstein's model, it is assumed that there is a fixed distance between the top of the lower teeth and the point around which the jaw rotates. The top of the lower teeth is point a_{11} in our model. Its position is calculated as follows:

$$a_{11} = \left(S_j\cos\left(A_j - \alpha_J\right), S_j\sin\left(A_j - \alpha_J\right)\right)$$

Points a$_9$ and a$_{10}$ are relative to this point:

$$a_9 = a_{11} + \left(LT2_x, LT_y\right)$$
$$a_{10} = a_{11} + \left(LT_x, LT_y\right).$$

Points a_{12} and a_{13} depend on both the position of the jaw and on the spread and protrusion of the lips, as follows:

$$a_{12} = a_{11} + \left(0, -z_L\right)$$
$$a_{13} = a_{11} + \left(x_L, z_L\right).$$

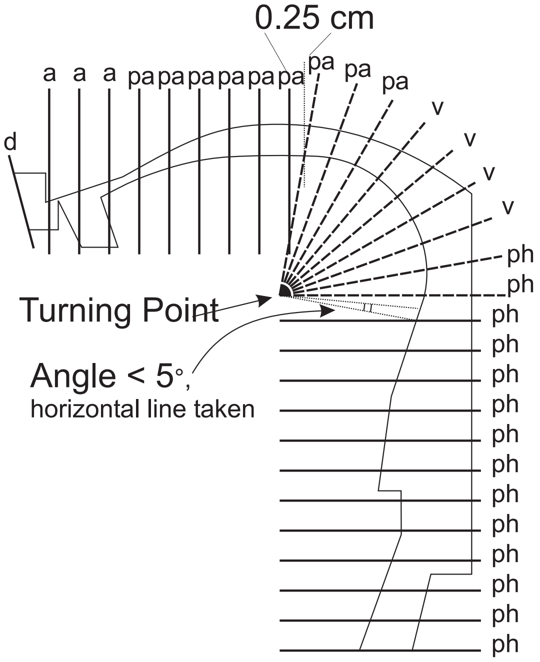
All points are connected by straight lines, except a_6 and a_7, which are connected by a circular arc with radius R_t, and a_7 and a_8, which are connected by a curve that is quadratic in polar coordinates. The starting radius is the tongue radius, and the ending radius is the distance between the tongue center and point a_8. The starting angle is the angle of the line from the tongue center to a_7, and the ending angle is the angle between the tongue center and a_8. The angle changes linearly, while the radius increases quadratically, according to the following equation:

$$r_t = (1-t)^2 r_{start} + t^2 r_{end}.$$

These equations and constants give the complete 2-dimensional midsagittal section of the model. On the basis of this section, the area function of the vocal tract can be estimated. In order to do this, a number of cross sections is calculated. In accordance with Mermelstein’s model, the cross sections in the laryngeal/pharyngeal section of the vocal tract are horizontal, the cross sections in the uvular/velar area are radial, and in the front part of the vocal tract, they are vertical. This is illustrated in figure 8. The horizontal sections start at the lowest possible point of the larynx, and continue until they reach the vertical coordinate –4.3 cm (relative to the jaw turning point). They are spaced 0.5 cm apart. The radial sections all pass through the *turning point* (7.2, –4.3). The first radial section starts at an angle that is 5° counterclockwise of the line that passes through the turning point and through the intersection between the anterior/inferior outline and the last horizontal section, or, if this would result in an intersection pointing downwards, is taken to be horizontal (still passing through the turning point). Subsequent intersections are taken at intervals of 10°, until they pass the vertical. The vertical sections start at a point that is 0.25cm to the left of the last anterior/inferior intersection of a radial line. They are spaced 0.5 cm apart, and continue until an intersection with either the posterior/superior outline or the anterior/inferior outline is no longer possible.

The last section is taken between point p_8 and a_{13}. These represent the horizontal extremities of the lips.

Figure 8: Sections as used for calculating the cross sectional area in the model. Note that precise positions of the sections depend on the position of the larynx.



As these sections do not accurately follow the orientation of vocal tract, it is necessary to scale them and interpolate them in order to obtain accurate uniform tubes. For this, first the centerline of the vocal tract is calculated. Then for each section, the angle with the centerline is calculated. For all sections except the first and the last this is done by calculating the two angles between the line perpendicular to the section and the two lines connecting the center of the section with the centers of the preceding and following sections. The angle with the centerline is then taken as the average of these two angles. For the first and the last sections, only the angles with the lines connecting to the following or the preceding sections are used. The length of the section is then multiplied with the sine of this angle in order to obtain the cross-sectional length perpendicular to the center line.

With these scaled cross sectional lengths, the areas of the vocal tract at the sections are estimated using Mermelstein's equations. Mermelstein uses different functions converting cross section into area for different regions of the vocal tract. As the model described here is slightly different in its calculation of cross-sections and as it must work for vocal tracts of different lengths, it is possible that the exact extent of the different regions is slightly different for our model than for Mermelstein's. It does follow his model as closely as possible, however. Unfortunately, it is not exactly defined in Mermelstein's paper where the different regions start and end. In the reimplementation the following criteria were used: for the pharyngeal region, the y-coordinate of the posterior-superior intersection must be less than –3.3, for the velar region the y-coordinate must be less than –1.8. For the palatal region, the x-coordinate of the posterior-superior intersection must be less than 9.7 and for the alveolar region the x-coordinate must be less than 11.2. The rest of the sections are considered to be part of the labial region.

For the pharyngeal region the area is that of an ellipse with one axis increasing from 1.5 cm at the larynx to 3 cm at the upper end. The other axis has the length of the cross section. For the velar region, the area is $2c^{1.5}$ where c is the length of the cross section. For the palatal region, it is $1.6c^{1.5}$. For the alveolar region, the following scheme is used:

$$\begin{array}{rl} 1.5c & c < 0.5 \\ 0.75 + 3(c - 0.5) & 0.5 \le c < 2 \\ 5.25 + 5(c - 2) & 2 \le c \end{array}.$$

Finally, the labial region, is again assumed to be elliptical, with one axis equal to the cross sectional length and the other equal to: $2 + 1.5(z_L - x_L)$.
Areas must minimally be 0.3 cm^2 (or the other values specified in the paper), while maximal areas for the different regions are given in table 4.

Table 4: Maximal areas in the different regions of the vocal tract (in cm^2).

Region	**Max. area**
Pharyngeal	6
Velar	5
Palatal	7
Alveolar	8
Labial	15

The center points of the cross sections are not equally far apart. In order to approximate the vocal tract with uniformly spaced tubes, a linear interpolation of the calculated areas is done at intervals of 0.5 cm. At the end of the vocal tract, two more tubes of 12 cm^2 and 30 cm^2 are added in order to model the way sound is radiated at the lips. Acoustic properties are calculated as if all sound waves are reflected at the glottis and all waves are radiated at the final tube.

8 List of references

Aiello, Leslie C. (1996). Terrestriality, bipedalism and the origin of language. In W. G. Runciman, J. Maynard Smith & R. I. M. Dunbar (Eds.), *Evolution of social behaviour patterns in primates and man* (pp. 269–289). Oxford: Oxford Academy Press.

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, Louis-Jean, Heim, Jean-Louis, Honda, Kiyoshi, & Maeda, Shinji. (2002). The potential neandertal vowel space was as large as that of modern humans. *Journal of Phonetics, 30*(3), 465–484.

Boë, Louis-Jean, Perrier, Pascal, Guerin, Bernard, & Schwartz, Jean-Luc. (1989). *Maximal vowel space.* In Eurospeech-1989, 2281–2284

Boersma, Paul. (1998). *Functional phonology*. The Hague: Holland Academic Graphics.

Carré, René, Lindblom, Björn, & MacNeilage, Peter 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.

Choi, John D. (1991). Kabardian vowels revisited. *Journal of the International Phonetic Association, 21*, 4–12.

Cormen, Thomas H., Leiserson, Charles E., & Rivest, Ronald L. (1993). *Introduction to algorithms*. Cambridge (MA): The MIT Press.

de Boer, Bart. (2000). Emergence of vowel systems through self-organisation. *AI Communications*, *13*, 27–39.

Diehl, Randy L. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics*, *24*(2), 187-208.

DuBrul, E LLoyd. (1958). *Evolution of the speech apparatus*. Springfield (IL): Charles C. Thomas.

Fant, Gunnar. (1960). *Acoustic theory of speech production*. 's Gravenhage: Mouton.

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

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

Fitch, W. Tecumseh, & Giedd, Jay. (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. Tecumseh, & Reby, David. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society of London Series B - Biological Sciences, 268,* 1669–1675.

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

Goldstein, Ursula Gisela. (1980). *An articulatory model for the vocal tracts of growing children.* Unpublished PhD, Massachusetts Institute of Technology, Cambridge (MA).

Houghton, Philip. (1993). Neandertal supralaryngeal vocal tract. *American Journal of Physical Anthropology, 90,* 139-146.

Kuhl, Patricia K., Andruski, Jean E., Chistovich, Inna A., Chistovich, Ludmilla A., Kozhevikova, E. V., Rysinka, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277,* 684–686.

Kuhl, Patricia K., & Meltzoff, Andrew N. (1996). Infant vocalization in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America, 100*(4), 2425–2438.

Ladefoged, Peter, & Maddieson, Ian. (1996). *The sounds of the world's languages*. Oxford: Blackwell.

Lee, Sungbok, Potamianos, Alexandros, & Narayanan, Shrikanth. (1999). Acoustics of children’s speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America, 105*(3), 1455–1468.

Lieberman, Philip H. (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, Philip H., & Crelin, Edmund S. (1971). On the speech of neanderthal man. *Linguistic Inquiry, 2,* 203–222.

Lieberman, Philip H., Crelin, Edmund S., & Klatt, Dennis H. (1972). Phonetic ability and related anatomy of the newborn and adult human, neanderthal man, and the chimpanzee. *American Anthropologist, 74*, 287–307.

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

Martínez, I., Rosa, M., Arsuaga, J.-L., Jarabo, P., Quam, R., Lorenzo, C., Gracia, A., Carretero, J.-M., Bermúdez de Castro, J.-M., & Carbonell, E. (2004). Auditory capacities in middle pleistocene humans from the sierra de atapuerca in spain. *Proceedings of the National Academy of Sciences, 101*(27), 9976–9981.

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

Metropolos, Nicholas, & Ulam, Stanislaw. (1949). The monte carlo method. *Journal of the American Statistical Association, 44*(247), 335–341.

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

Negus, V E. (1949). *The comparative anatomy and physiology of the larynx*. London: William Heinemann Medical Books Ltd.

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

Peterson, Gordon E., & Barney, Harold L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175–184.

Puts, David Andrew, Gaulin, Steven J. C., & Verdolini, Katherine. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior, 27*(4), 283–296.

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.

Schepartz, L. A. (1993). Language and modern human origins. *Yearbook of Physical Anthropology, 36*, 91–126.

Schroeder, Manfred 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* (pp. 217–229). London: Academic Press.

Schwartz, Jean-Luc, Boë, Louis-Jean, Vallée, Nathalie, & Abry, Christian. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, *25*, 255–286.

Smith, Evan C., & Lewicki, Michael S. (2006). Efficient auditory coding. *Nature, 439*(7079), 978–982.

Soquet, Alain, Lecuit, V., Metens, Thierry, & Demolin, Didier. (2002). Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with mri. *Speech Communication, 36*, 169–180.

Story, Brad H., Titze, Ingo R., & Hoffman, Eric A. (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America*, *100*(1), 537-554.

Story, Brad H., Titze, Ingo R., & Hoffman, Eric 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.

Watrous, Raymond L. (1991). Currrent status of Peterson-Barney vowel formant data. *Journal of the Acoustical Society of America, 89*(5), 2459–2460.

Weber, Ernst Heinrich. (1834). *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae*. Leipzig: Koehler.

Contact information:
Bart de Boer
ACLC
Universiteit van Amsterdam
Spuistraat 210
1012 VE Amsterdam
tel. +31 20 525 2182
b.g.deboer@uva.nl