Title: Investigating the acoustic effect of the descended larynx with articulatory models

Running title: Articulatory models of the descended larynx

Author: Bart de Boer

Affiliation:
Amsterdam Center for Language and Communication
Universiteit van Amsterdam
Spuistraat 210
1012 VT Amsterdam
the Netherlands
Electronic mail: b.de.boer@ai.rug.nl

Appeared in:
Journal of Phonetics, 2010

Abstract:

In this paper a strongly simplified articulatory model, as well as three more realistic models are investigated for the effect of larynx height on the extent of vowel signaling space. The models explore a larger range of larynx positions than previous models, and the use of the convex hull for measuring articulatory abilities is introduced. 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 most closely to the anatomy of the female vocal tract. These findings are consistent with the hypothesis that the female vocal tract has evolved to be optimal for speech, while the male vocal tract has also evolved under another pressure, possibly size exaggeration.

Keywords:

Descent of the larynx

Evolution

Vowel space

Articulatory distinctiveness

Male and female formants

1 Introduction

The anatomy of the human larynx, tongue and pharynx is different from that of other closely related primates: the tongue is round, the larynx is lower, and the pharynx is therefore longer (Negus, 1949; Fitch, 2006). Furthermore, the larynx cannot be locked into the nasal passage in adult humans, thus making it impossible to breathe and swallow at the same time. The evolutionary significance of this has been debated for a long time (e. g. Negus, 1938; e. g. DuBrul, 1958; Lieberman & Crelin, 1971; Fitch, 2000) and recently also in the Journal of Phonetics (e. g. Boë *et al.*, 2002; Boë *et al.*, 2007; Lieberman, 2007).

There are two independent questions in this debate: is it possible to have modern languages that do not make use of the full range of potential modern speech sounds, and did the modern human vocal tract evolve under pressure for producing as large a range of speech sounds as possible? From languages with small phoneme inventories, we know that it is possible to have modern languages that do not use the full potential of the human vocal tract (Maddieson, 1984; Choi, 1991; Ladefoged & Maddieson, 1996, pp. 286-288). 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. However, the fact that languages can flourish with a limited set of articulations does not necessarily lead to the conclusion that the vocal tract did not evolve for speech.

In order to establish whether a certain trait (the position of the human larynx) evolved for a certain purpose (producing as large a range of speech sounds as possible) evolutionary biology tells us that it needs to be established how close it is to optimal (Parker & Maynard Smith, 1990). In addition it needs to be established that there is a path of ever increasing fitness from the ancestral state (comparable to chimpanzee or gorilla vocal tracts) to the optimal state and that there is heritable variation of the trait. Given that the position of the larynx is no different than other anatomical traits it can be safely assumed

that there is such heritable variation. I first argue that individuals who can produce a larger range of signals always have an advantage over individuals with a smaller range, and I then present computer models showing that the human female vocal tract is near optimal for producing the largest possible range of vocalic signals.

As a first approximation, I assume that increase in fitness related to speech is mostly determined by the amount of information that can be communicated (when discussing the results for the male vocal tract, I discuss what happens if multiple factors determine fitness). Information theory (Shannon, 1948) shows that when communicating under noise, signals that are more distinct have the advantage. This result is completely independent from the nature of the signals involved. Furthermore, evolutionary game theory (Nowak *et al.*, 1999; Zuidema & de Boer, 2009) shows that individuals with a larger signaling space can always invade a population of individuals with a smaller signaling space (all other factors being equal). Intuitively, this is possible because the larger signaling space contains all the possible articulations of the smaller signaling space, while, because of categorical perception, individuals with the smaller signaling space will still correctly classify signals that fall slightly outside their space.

Furthermore, it has been shown that through cultural evolution and self-organization, sound systems tend to move towards maximal use of the available phonetic space (Liljencrants & Lindblom, 1972; Lindblom *et al.*, 1984; Schwartz *et al.*, 1997; de Boer, 2000b, 2000a). Although it is possible that, through historical processes and a pressure to conform to the speech community, systems can emerge that do not use the whole acoustic space, such systems tend to be exceptional. For example, of the 451 languages in UPSID (Maddieson, 1984; Maddieson & Precoda, 1990) 337 have /i/, /a/ and /u/ (either short or long), and thus use the complete vowel triangle. The advantage of individuals with a larger signaling space combined with a tendency of speech communities to (culturally) evolve towards maximal use of available acoustic space causes a higher fitness for individuals with a larger signaling space at any point in the evolution of the vocal tract.

Below, I propose a range of computer models of increasing complexity to investigate how optimal the modern human vocal tract is. The approach is based on the field of artificial life (Langton, 1989): a range of models is implemented to investigate the effect of a range of hypothetical vocal tracts. In this way insight is obtained into the potential effects of larynx position. Finally, the modeling results are compared with a classic set of data on human vowel production (Peterson & Barney, 1952) in order to check the extent to which the predictions of the model conform to reality.

2 Modeling and Measuring Articulatory Range

It is necessary to use a computer model that captures the constraints on articulation. The experiments described below are therefore based on the use of geometric articulatory computer models (for a caveat on the use of models based on factor analysis, see Fitch & de Boer, 2010). 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 relatively straightforwardly to muscle actions.

It was necessary to abstract away from certain anatomic details in order to investigate the effect of larynx position in as pure and understandable a way as possible. For instance, in the simplest model used here, no details of the laryngeal anatomy were modeled, nor were the lips. Such simplified models can help to answer the question as to which factors are most important. Experiments with a more realistic model and verification with real-world data are used to check whether the simplifications are valid.

An important methodological issue is how to compare the articulatory abilities of two different (hypothetical) anatomies. Articulatory abilities of different vocal tracts are usually measured by the range of formant patterns they can generate. Although for a complete description of a speech sound, the whole frequency response should be taken

into account, usually only the first two formants are represented. The first two formants do not represent the complete signal, but the range of possible two-formant positions correlates well with the complete range of frequency responses that can be generated with a given vocal tract. The frequencies of F_1 and F_2 have therefore been widely used to compare the use of acoustic space between different conditions, not only in research into the evolution of speech (Lieberman & Crelin, 1971; Carré *et al.*, 1995; Boë et al., 2002), but also in research into vowel perception (e. g. Peterson & Barney, 1952; Rosner & Pickering, 1994) or acquisition of speech (e. g. Kuhl & Meltzoff, 1996; e. g. Kuhl *et al.*, 1997).

The extent in F_1-F_2 space is therefore a useful measure of the articulatory abilities of a given vocal tract. However, it can be argued that one should also look at the kind of trajectories and transitions that can be made with a given vocal tract anatomy. Such transitions and the mapping between articulatory actions and their acoustic effects is especially important in the study of consonants and syllables, and is stressed by Stevens' quantal theory of speech (Stevens, 1972, 1989) and by Carré's work on the distinctive region model (Carré & Mrayati, 1995; Carré, 2004).

There are several reasons why basing an evaluation of articulatory abilities on the types of trajectories and transitions that can be generated is problematic. First of all, a larger F_1-F_2 space would also result in a larger range of possible trajectories. Still, it might be possible that in the larger space, certain transitions are less easily produced. Unfortunately, there is little consensus about the nature of transitions that are most salient in speech or how these are generated. Stevens (Stevens, 1972) proposes that points that are most stable are crucial, and predicts that systems of human speech avoid regions of abrupt transitions. Carré (Carré & Mrayati, 1995; Carré, 2004) on the other hand, argues that rapid transitions are crucial, and predicts that inventories of speech sounds make use of regions of rapid transition. It is possible that both points of view are valid; stability might be useful for vowels, whereas rapid transitions might be useful for consonants. Given our lack of knowledge, it is unfortunately impossible to calculate an evaluation

criterion based on transitions. In future experimental work, it would be valuable to investigate what transitions and trajectories are salient in human speech perception and production, and to investigate with articulatory models which anatomical arrangements produce them most reliably. However, such a research program falls outside the scope of the work presented here.

3 Basic Methods

The models used in this study 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. It was decided not to use Goldstein's (1980) model because while that model was designed to investigate 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 is to calculate the 2-dimensional mid-sagittal outline of the vocal tract for a given set of articulatory parameters. The second consists of converting this two-dimensional outline in an area function. The third consists of calculating the acoustic properties of this particular area function. This last step can be performed by application of standard acoustics theory (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 are given in the sections below.

The procedure for exploring the range of possible articulations is based on the idea of Maximal Vowel Space as defined by (Boë *et al.*, 1989). It consists of generating a large number of articulations, and calculating the area of acoustic space which they cover. A systematic exploration of every possible combination of articulatory parameters might

appear most straightforward. The continuous ranges of the articulatory parameters can be divided into a number of equally spaced values, and all possible combinations explored. However, this approach suffers from two problems. The first is that the number of articulations to be explored rises exponentially with the number of articulatory parameters. The second is that, due to the discretization of the parameter range, it is likely that articulations resulting in extreme values of the signal are missed.

A better approach is to generate a large number of random articulations (using a Monte Carlo approach, Metropolis & 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 always gives the same value, and therefore gives no idea of how well the space is sampled.

Vocalic signaling space was defined using the frequencies of F_1 and F_2 measured on a 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 that the basics of perception are much older than any differences in vocal anatomy (e. g. 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 scale or using the ordinary logarithm of frequency did not result in any qualitative differences.

The measures of the range of articulations that a model can produce are the total area and

its maximal extents in the first and second formant. The area covered by the articulatory model was calculated by dividing the acoustic space into a grid of tiles of 0.5×0.5 Barks, and counting the number of tiles that had at least one acoustic signal in them.

The ranges of the articulatory parameters were selected to be physiologically plausible, but a combination of articulatory values could still result in an articulation where parts of the vocal tract intersect each other. This was automatically detected, first identifying any intersections of each vocal tract wall with itself (for example, the tongue body with the epiglottis) and then, when calculating the value of the cross sectional areas of the vocal tract, by detecting intersections of one wall with the other. All articulations with intersections were discarded. Also, because the acoustic simulations were limited to airflow without turbulence, 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. Repeating the experiments with 0.1 cm^2 and 0.5 cm^2 did not change the qualitative results.

4 The Exploratory Model

The exploratory model is a simplified version of Mermelstein's (1973) model. It 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 as possible, and also to prevent the hard structures of the larynx and epiglottis from interfering with the movement of the tongue (this is a problem when changing the larynx position in Mermelstein's model).

The only articulatory motions that were modeled were the motion of the jaw the motion of the tongue body (both tongue displacement and tongue angle) and the horizontal motion of the hyoid. Ranges of these parameters (as well as those for the extra parameters

of the more realistic models described below) are given in table 1. The tract was terminated at the mouth by a vertical plane at a constant position. Finally, cross-sectional diameters are converted to cross sectional areas in the same way everywhere by squaring the value of the diameter.

Table 1: The minimal and maximal values of the articulatory parameters.

	min	max
Hyoid horizontal	–1	1
Hyoid vertical	–1	1
Jaw angle	–0.25	0.25
Tongue tip angle	–0.25	0.25
Tongue displacement	–1.5	1.5
Tongue body angle	–0.2	0.2
Lip protrusion	0	1.5
Lip spread	–1.5	1.5

The model is illustrated in figure 1, and can be compared to the more realistic model in figure 3. 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. 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 length of the vertical tube is equal to the

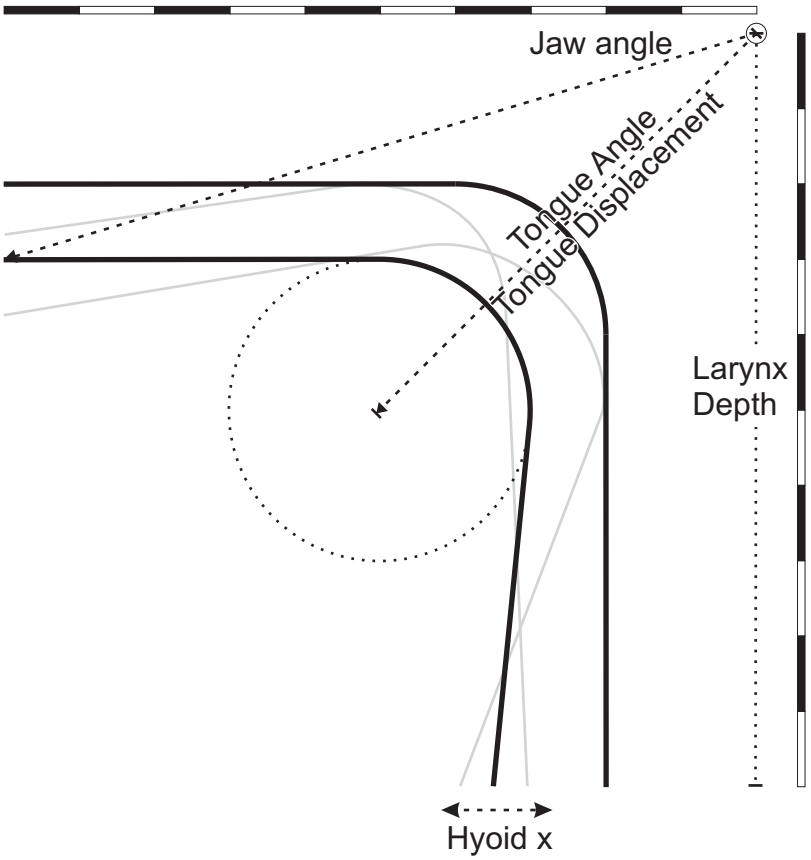

Figure 1: The simplified model. The outline of the model is shown in bold black lines. The articulatory parameters are shown as dashed arrows. The tongue contour and the larynx depth are given as dotted lines. Two potential articulations are shown as thin grey lines. For scale, horizontal and vertical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

larynx depth minus 2 cm. Exact dimensions can be determined from figure 1, which is to scale.

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. 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. It is clear from this figure that the (simplified) 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 smaller reachable areas of acoustic space (using the

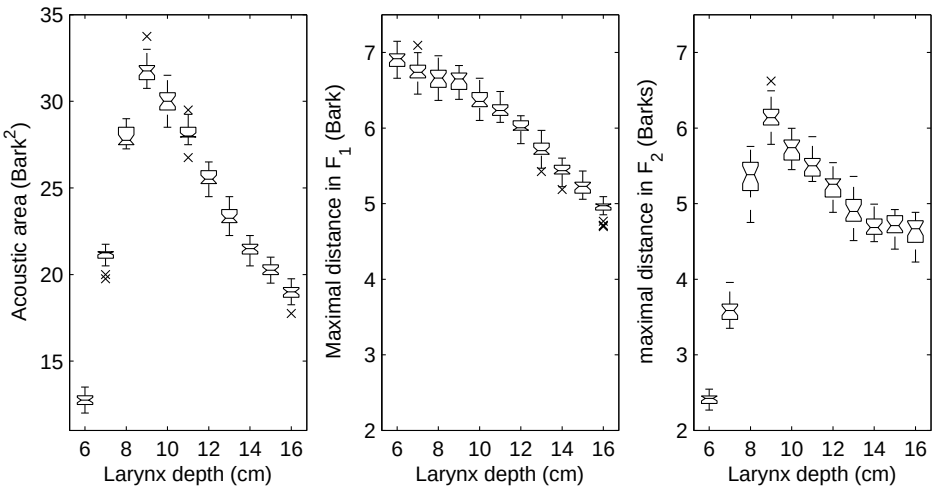

Figure 2: The relation between acoustic area and larynx depth in the simplified model. The box plot shows the median (horizontal line) as well as the first and third quartiles (top and bottom of the boxes). The total extent is indicated by whiskers, while outliers are shown as crosses. Notches in the 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.

Wilcoxon rank sum test to compare all neighboring larynx depths, $p < 0.05$). 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 formant, higher larynges actually appear to be very slightly better than models with medium or low larynges.

5 The Realistic Model

Is the effect found only an artifact of the simplified model, or is a similar effect also obtained with more realistic models? To answer this question, three articulatory synthesis models were compared, all based on Mermelstein's (1973) model.

Mermelstein's model is of the male vocal tract. An exact reimplementation of his model

was therefore used for modeling a male vocal tract. For a 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 lies approximately 2.2 cm higher than the male larynx. This corresponds well with the 2.8 cm difference in Goldstein's (Goldstein, 1980) model. The larynx depth, measured in an equivalent way to that of the simplified models, is then 8.8 cm for the female model and 11 cm for the male 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). In the model, this is restricted to 0.5 cm in both directions.

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). 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. All models are described in more detail in (de Boer, 2007).

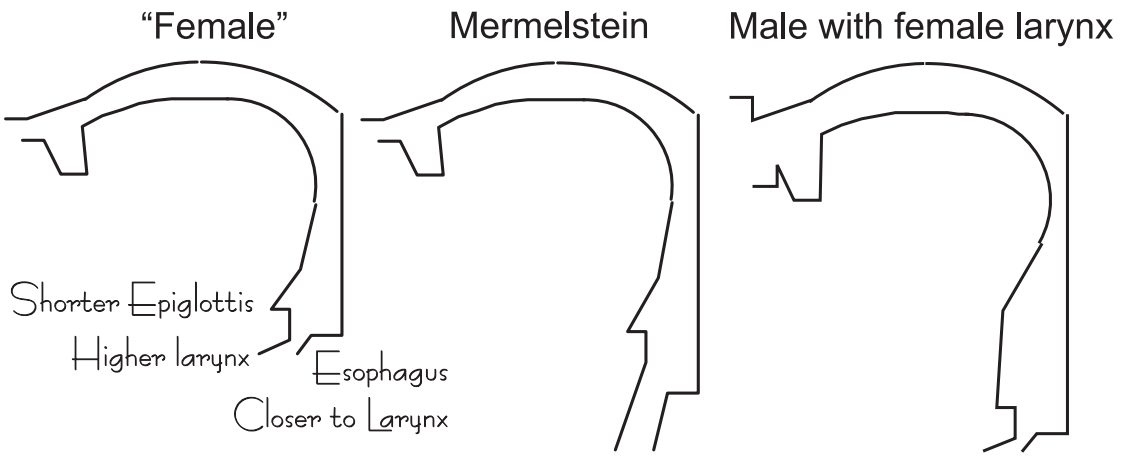

Figure 3: Comparison of the original Mermelstein (1973) model (middle) to the model with the raised, "female" larynx (left) and the model with male position and female shape (right). The differences between the models are indicated in the female model.

In order to convert the 2-dimensional cross section into a 3-dimensional area function, the same conversion functions are used for the male, female and mixed 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 clear focus on the role of the position of the larynx.

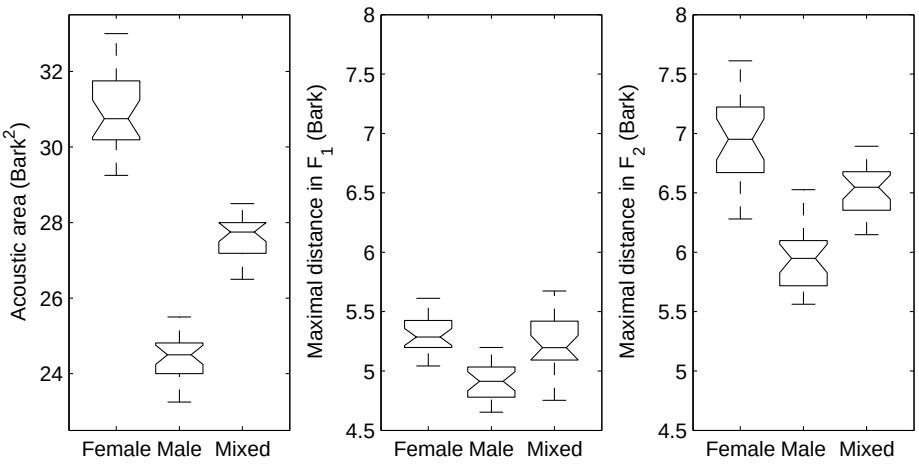

Figure 4: Acoustic area and extent of the first and second formant of the more realistic models. Note that the female model is significantly better in all respects than the male model.

With all three articulatory models, 25 sets of 4000 articulations each were generated. For these articulations, the first and second formants were calculated. 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, covers 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.

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 gauge 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 2.

Table 2: 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. 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 (e. g. Engwall & Badin, 1999) with the possible exception of the

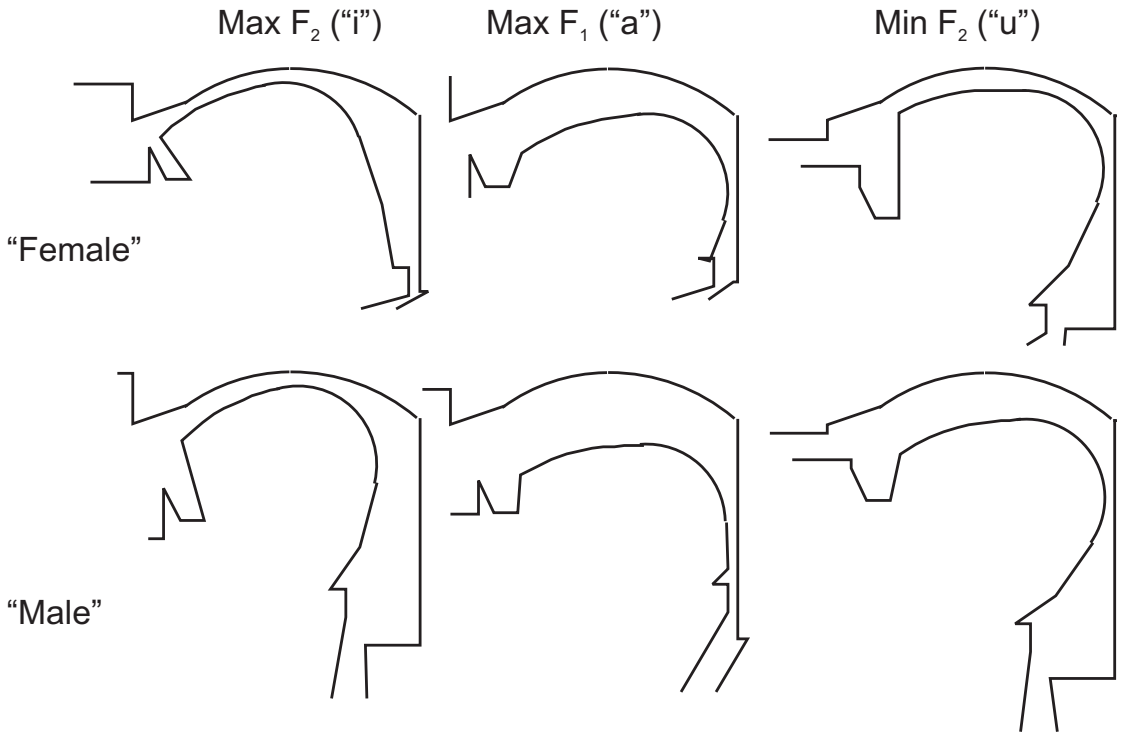

Figure 5: Comparison of articulations with maximal F_2 and F_1 and minimal F_2. Note that for both the male and female models, the articulations correspond well with the articulations for [i], [a] and [u]. Discontinuities in the outlines are artifacts of the requirement of the programming environment to use integer values for plotting.

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 selected by the sampling procedure. It should also be noted that the articulations that are depicted in figure 5 are by definition relatively extreme, as they are the ones with the most extreme formant values. However, all articulatory models can in principle produce equally extreme articulations, so the comparison between the different conditions remains reasonable.

6 Human data

Having established that larynx depth influences the signal range of simplified articulatory models, one can ask whether this is an artifact of the models, or whether a similar effect obtains in real vocal tracts. A potentially promising way is to look at human data, as human males and females have different larynx heights (and corresponding pharynx lengths and volumes).

Studies of different languages show consistently larger extents in both the first and the second formant of the vowel spaces associated with female articulations than with male articulations (e. g. the data presented in Fant, 1975). There is a debate to what extent these results can be explained by behavioral or by anatomical factors (Nordström, 1977; Goldstein, 1980; Traunmüller, 1984; Diehl *et al.*, 1996) but the consensus appears to be that both factors play a role. Looking at data from children can perhaps help to solve part of this puzzle. Fitch and Giedd (1999) do not find anatomical differences between pre-adolescent children, but Perry et al. (2001) find that eight year old boys have lower average formant values than girls of the same body weight. This might be due to boys protruding their lips more than girls (Fitch & Giedd 1999, citing Sachs *et al.*, 1973). In order to tease apart anatomy from behavioral effects, data from men, women, boys and girls are studied here. If differences are due to behavior, one would expect to find

differences between boys and girls. If differences are due to larynx descent, one would expect to find that women, boys and girls fall in the same group.

The classic American English vowel data set from Peterson and Barney (1952) has been used as reconstructed and made publicly available by Watrous (1991) and in the software package PRAAT (Boersma & Weenink, 2008). This data set is ideally suited for this research, as it contains data from men, women and children, as well as *all* data points of *all* speakers.

A methodological problem is that 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 for the model-generated data, the area of the convex hull correlates almost perfectly (correlation coefficient 0.98) 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 area as is really covered by a data set. However, when comparing areas that are calculated with the same method, this should not be a problem. Areas generated by the two different methods should not be compared directly, however.

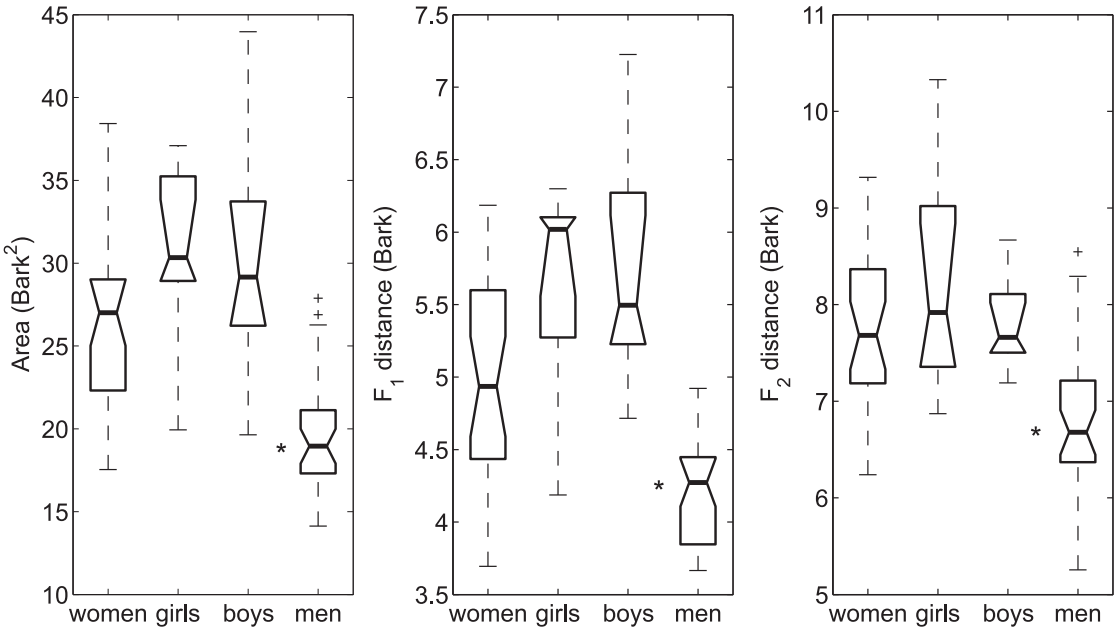

Figure 6: Statistics of the comparison of human data taken from Peterson and Barney's dataset. Data from adult women, girls, boys and adult men are shown. Horizontal fat lines indicate the medians, box edges indicate the lower and upper quartiles, and whiskers indicate the total extent of the data, except outliers, which are indicated with pluses. Note that exact values of the areas should not be compared with those of the modeling experiments, as the test conditions and the way of measuring acoustic area are different.

The different measures of the human data were compared using the Wilcoxon rank sum test. It was found that the female, boy and girl vowel spaces were larger than the male vowel spaces with $p < 0.001$ for all measurements. It was also found that there was no significant difference between women, boys and girls. The measures are given in figure 6, and statistics are given in table 3. Similar results were obtained from datasets on Dutch vowel systems by (Pols *et al.*, 1973; Van Nierop *et al.*, 1973), and from (Weenink, 1985). These datasets are also available in PRAAT. It should perhaps be stressed that these results do not indicate that there are no differences at all between boys, girls and women. Indeed differences between boys and girls have been found (Perry *et al.*, 2001; Vorperian *et al.*, 2009). It only appears that with respect to the area of acoustic/perceptual space, children and adult women form one group, and adult men another.

Table 3: statistics of the comparisons performed on Peterson and Barney's data set.

comparison		**p (**	**rank sum**	**p**	**rank sum**	**p**	**rank sum**
female (n=28)	male (n=33)	1.17×10^{-6}	1186	2.76×10^{-6}	1176	1.93×10^{-5}	1152
female	girls (n=8)	0.0191	209	0.0401	202	0.4429	169
female	boys (n=7)	0.146	162	0.0538	173	0.8561	131
male	girls	7.26×10^{-6}	287	3.46×10^{-5}	281	7.27×10^{-4}	265
male	boys	1.26×10^{-4}	241	2.04×10^{-6}	254	4.45×10^{-4}	235
girls	boys	0.6126	51	0.8665	54	0.6943	52

note that for total significance at the $p < 0.05$ level, *p*-values must smaller than 0.05/6 (Bonferroni correction) because of multiple comparisons

As can be observed in the figure, there is considerable individual variation and overlap. This might be caused among other things by individual variation in anatomy, by behavioral compensation and because the larynx height can to some degree be controlled. However, the data do show that on average the males, whose larynx has made an extra descent during puberty, have a significantly smaller acoustic space than women, boys and girls, with vocal tracts where the length of the vertical (pharyngeal) part is approximately equal to the horizontal (oral) part. The differences between the human averages and between the averages of the models are comparable, but the spread in the human case is much larger. The average difference could be explained as the result of average anatomical differences between men and women, while the larger spread could reflect anatomical variation and behavioral compensation due to either different articulatory habits, or due to control over larynx height. Although there is considerable overlap between the groups, the difference in median is significant and considerable. It seems that the anatomical changes after puberty lead to reduced size of the signaling space.

The fact that female and child data fall in one group, and male data in a different group weakens the argument that the difference between men and women is behavioral. One would then expect boys to be different from women and girls. It also suggests that individual variation (e. g. Laver, 1980) and the ability to move the larynx vertically do not negate the difference between men and women.

One way to further test the relation between behavior, development and vowel space size is to investigate the way vowel spaces change over puberty. The anatomy-is-important hypothesis predicts that boys and girls are indistinguishable before puberty, but become more different during puberty. Unfortunately, research into the effect of development on formant frequencies only presented averages over age groups (Lee *et al.*, 1999), or averages over all vowels of each individual (Perry et al., 2001). For testing the hypothesis presented here, data of vowel categories of individual speakers are needed.

7 Discussion

This study has investigated the influence of larynx position on the size of the acoustic space that can be generated by the vocal tract. In a highly simplified computer model of the human vocal tract, which nevertheless implemented essential constraints of anatomy and muscular control, 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 this model the optimal position occurred when the vertical part of the vocal tract was slightly shorter than the horizontal part. Comparable results were found when more realistic models of the male and female vocal tracts were used. Here too, the model in which the vertical part of the vocal tract was slightly shorter than the horizontal part – the female model – could generate a larger repertoire of signals than the models in which the vertical part was longer – the male and mixed models.

Analysis of data from speakers of American English showed that whereas women, boys and girls have comparable vowel spaces, men had a significantly smaller vowel signaling space, thus corroborating the modeling results. The similarity of boys' and girls' vowel spaces appears to argue against a behavioral explanation. These observations suggest that the human (female) vocal tract, with almost equally long horizontal and vertical parts is optimal for producing as large a range of speech sounds as possible.

The reason for this is that this configuration of the vocal tract allow for the range of deformations that results in the largest range of acoustic signals (assuming identical control over the different articulators, most importantly tongue, lips, pharynx and larynx). 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.

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. Together with the findings that in noisy communication larger acoustic spaces always have an advantage over smaller ones (Nowak et al., 1999; Zuidema & de Boer, 2009), and that signaling systems tend to fill the available acoustic space through cultural evolution (de Boer, 2000b; Oudeyer, 2005; de Boer & Zuidema, to appear), it could be argued that this would create a strong (biological) evolutionary pressure for the larynx to descend

The results then are compelling from an evolutionary perspective. The female larynx position appears to be very close to the one found to be optimal in the simplified model. This is consistent with the hypothesis (first articulated by Lieberman et al., 1969) that the human vocal tract has evolved to produce as distinctive articulations as possible. The fact that the male larynx is slightly lower than the optimal position might be explained in evolutionary terms by the fact that this helps to exaggerate size (Ohala, 1984; Fitch, 2000; Fitch & Hauser, 2002). It has been found that this is important for animals, and it

has also 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 perhaps then be explained by the fact that the male vocal tract still needs to be able to produce a sufficient repertoire of distinctive speech sounds.

Acknowledgements

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.

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., & Weenink, D. (2008). Praat: Doing phonetics by computer (Version 5.0.16). Amsterdam.

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

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* (pp. 73–89). Amsterdam: Elsevier.

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

Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1993). *Introduction to algorithms*. Cambridge (MA): The MIT Press.

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

de Boer, B. (2000b). Self organization in vowel systems. *Journal of Phonetics, 28*(4),

441–465.
de Boer, B. (2007). Investigating the acoustic effect of the descended larynx with articulatory models. *ACLC Working papers, 2007*(2), 61–86.
de Boer, B., & Zuidema, W. (to appear). An agent model of combinatorial phonology. *Adaptive Behavior*.
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.
DuBrul, E. L. (1958). *Evolution of the speech apparatus*. Springfield (IL): Charles C. Thomas.
Engwall, O., & Badin, P. (1999). Collecting and analysing two- and three-dimensional mri data for swedish. *Speech, Music and Hearing Quarterly Progress and Status Report, 40*(3–4), 11–38.
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. (2006). Production of vocalizations in mammals. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (pp. 115–121). Oxford: Elsevier.
Fitch, W. T., & de Boer, B. (2010). Computer models vocal tract evolution: An overview and a critique. *Adaptive Behavior, 18*(1), 36-47.
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* (pp. 65–137). New York: Springer.
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).
Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevikova, E. V., Rysinka, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*, 684-686.
Kuhl, P. K., & Meltzoff, A. 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, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.
Langton, C. G. (1989). *Artificial life*. Reading, MA: Addison Wesley.
Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America, 105*(3), 1455–1468.
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., Klatt, D. H., & Wilson, W. H. (1969). Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science, 164*, 1185-1187.
Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems. *Language, 48*, 839–862.
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* (pp. 181-203). Berlin: Walter de Gruyter & Co.
Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.
Maddieson, I., & Precoda, K. (1990). Updating upsid. *UCLA Working Papers in Phonetics, 74*, 104–111.
Martínez, I., Rosa, M., Arsuaga, J.-L., Jarabo, P., Quam, R., Lorenzo, C., et al. (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.
Metropolis, N., & Ulam, S. (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.
Nordström, P.-E. (1977). Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics, 5*, 81–92.
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.
Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology, 233*(3), 435–449.
Parker, G. A., & Maynard Smith, J. (1990). Optimality theory in evolutionary biology. *Nature, 348*, 27–33.
Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *Journal of the Acoustical Society of America, 109*(6), 2988–2998.
Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175–184.

Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America, 53,* 1093-1101.

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.

Rosner, B. S., & Pickering, J. B. (1994). *Vowel perception and production*. Oxford: Oxford University Press.

Sachs, J., Lieberman, P., & Erickson, D. (1973). Anatomical and cultural determinants of male and female speech. In R. W. Shuy & R. W. Fasold (Eds.), *Language attitudes: Current trends and prospects* (pp. 74–84). Washington DC: Georgetown University Press.

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

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics, 25*, 255–286.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27*, 379–423,623–656.

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

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

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. J. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51–66). New York: McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*(1), 3-45.

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.

Traunmüller, H. (1984). Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. *Speech Communication, 3*, 49–61.

Van Nierop, D. J. P. J., Pols, L. C. W., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acustica, 29*, 110–118.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., et al. (2009). Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *Journal of the Acoustical Society of America, 125*(3), 1666-1678.

Watrous, R. L. (1991). Currrent status of Peterson-Barney vowel formant data. *Journal of*

the Acoustical Society of America, *89*(5), 2459–2460.
Weenink, D. (1985). Accurate algorithms for performing principal component analysis and discriminant analysis. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, *19*, 45–52.
Zuidema, W., & de Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, *37*(2), 125–144.