

SIMULATION OF VOCAL TRACT GROWTH FOR ARTICULATORY SPEECH SYNTHESIS

Peter Birkholz¹ and Bernd J. Kröger²

¹Institute for Computer Science, University of Rostock, 18051 Rostock, Germany

²Department of Phoniatrics, Pedaudiology, and Communication Disorders
University Hospital Aachen (UKA) and Aachen University (RWTH), 52074 Aachen, Germany
piet@informatik.uni-rostock.de, bkroeger@ukaachen.de

ABSTRACT

We present a three-dimensional articulatory model of the vocal tract with the capability to simulate growth from infancy to adulthood. This model is intended to be applied for the articulatory synthesis of children's speech and the study of speech acquisition. To generate the vocal tract shape for a given age (1–20 years) and sex, we rescale the anatomic structures of an adult reference vocal tract according to natural growth patterns. Furthermore, we discuss the transformation of the articulatory state from the reference vocal tract to a vocal tract with a different anatomy by example of an 11-year-old boy. To reproduce the formant frequencies of children of that age, it is not enough to scale the articulation analogous to the changes of the palatal and pharyngeal length. Instead, our results suggest systematic differences in the articulation between adults and children.

Keywords: Vocal System, Vocal Tract Growth, Articulatory Speech Synthesis

1. INTRODUCTION

A fundamental part of any articulatory speech synthesizer is a model of the human vocal tract. Typically, such models are derived from radiographic or magnetic resonance images (MRI) of the the vocal tract of an adult speaker. However, there are hardly any computer models that replicate the vocal tract and articulation of children. To our knowledge, the only articulatory models capable to simulate growth of the vocal tract from infancy to adulthood are the midsagittal models due to Goldstein [6] and Boë and Maeda [3]. A major benefit from such models is that they permit to study speech acquisition and speech development during childhood [4]. On the other hand, they are a good starting point for the development of *generic* vocal tract models that can easily be fitted to the anatomy of arbitrary speakers of different age, sex and morphology. These models could help to imitate different voices in articulatory speech synthesizers.

The above mentioned studies, as well as the recent MRI study by Vorperian *et al.* [9], show that the vocal tract does not simply scale up uniformly from birth to adulthood. The proportions of diverse parts of the vocal tract are rather different between an infant and an adult, and they grow at different rates. Goldstein [6] summarizes the major differences between infants and adults as follows: The infant has a much shorter vocal tract than the adult, the pharynx of the infant is shorter relative to the overall vocal tract length, and the lateral dimension of

its vocal tract is wider relative to its overall length. Interestingly, acoustic data indicate that the vowel space can be *linearly* scaled between adults and children [4, 7, 8] and thus do not reflect the non-uniform growth. This implies changes of articulatory strategies during growth.

In this paper we present the first *three-dimensional* model of the growing vocal tract. The intended areas of application for this model are the study of speech acquisition and articulatory speech synthesis. Compared to the existing midsagittal models, a 3D anatomic model is more complex, but represents the vocal tract anatomy and the vocal tract area function more precisely. The basic idea was to use our existing vocal tract model representing an adult man as a reference (cf. Fig. 1a and [1, 2]) and scale different parts of the model according to the growth curves given in [6]. These transformations will be presented in the next section. Section 3 describes the mapping of articulatory configurations of vowels from the reference vocal tract to that with a changed anatomy. A short discussion and conclusions follow in Section 4.

2. MODELING GROWTH OF THE VOCAL TRACT STRUCTURES

The simulation of vocal tract growth was based on the ample collection of data by Goldstein [6]. In her dissertation, she collected numerous craniofacial measurements for various ages between birth and adulthood around age 20 (which is the end of the scale in her and our model). All of these measurements define either the angle or the distance between cephalometric landmarks in the region of the head and the neck. The data points for the various measurements were fitted with parameterized growth curves that express the values of the measurements as a function of age and sex. The subset of distance measurements that we used for the simulation of growth are summarized in Fig. 2. Many of these measurements do not directly correspond to dimensions of the vocal tract model. Therefore, we had to define suitable vocal tract dimensions and appropriate transformations between Goldstein's measurements and these dimensions.

Fig. 3 shows the selected anatomic dimensions of the vocal tract model. Each of them controls the growth of a particular structure in one direction. When one of these dimensions has a value different from that of the reference vocal tract, the affected structure of the reference vocal tract is shortened or stretched accordingly. The length of the mandible and the lower lip is not specified individually, but their length is kept proportional to W_1 and W_0 . The teeth are not scaled as W_1 changes, but

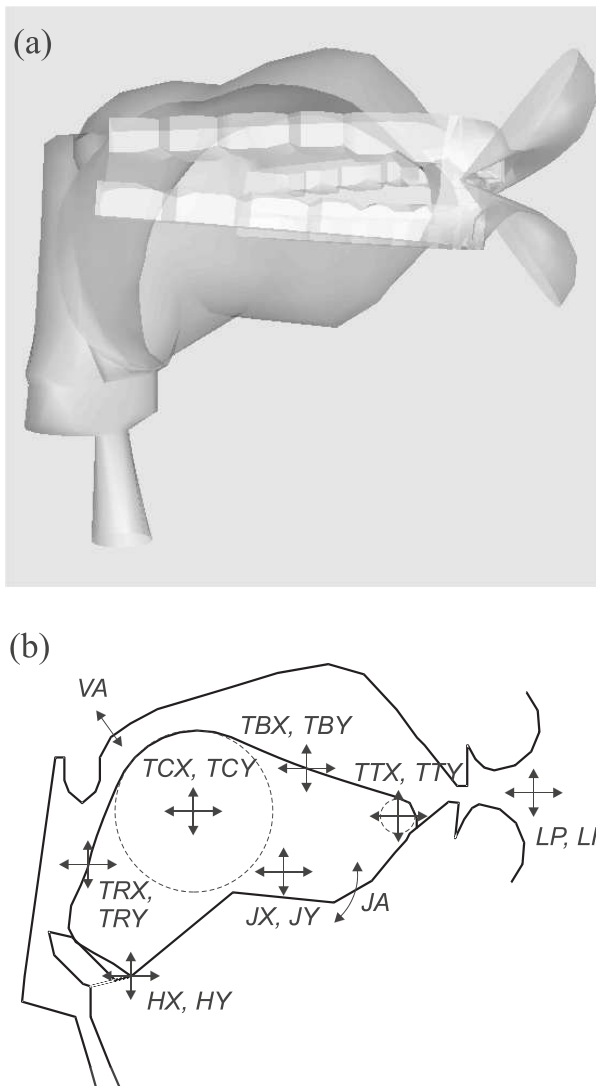


Figure 1: (a) 3D model of the vocal tract of an adult man. (b) The vocal tract parameters and their respective areas of influence on the articulators.

teeth are added or removed depending on the length of the hard palate and the mandible and hence the available space for teeth. Little data is available about the lateral growth of the vocal tract. Therefore, the lateral size of the whole model is scaled according to the scaling factor for D_0 . Some of the dimensions (e.g., H_0 and W_4) apply to the size of soft rather than rigid structures, and therefore affect the resting position or resting shape of that structure instead of its definite position or length.

For the mapping from the measurements in Fig. 2 to the anatomic parameters of the model we made the following assumptions. The highest point of the lingual surface of the palate is 1 mm below the palatal line, the width of the larynx grows proportional to the length of the soft palate ($W_4 = W_2$), the upper edge of the hyoid is covered with 1 mm of soft tissue, the measurement of *HYGLOTT* is oriented at an angle of 120° with respect to the palatal line, and *SYMPH* at an angle of 70° . This leads to the

following equations:

$$\begin{aligned} W_0 &= LIPS - 5 \text{ mm} \\ W_1 &= ANSPNS - 3 \text{ mm} \\ W_2 &= ATLPTM + 3 \text{ mm} - 5 \text{ mm} \\ W_3 &= LENVC \\ W_4 &= W_2 \\ H_0 &= SNHY - SNPNS - THIH/2 - 2 \text{ mm} \\ H_1 &= HYGLOTT \cdot \sin(120^\circ) + THIH + 1 \text{ mm} \\ H_2 &= PALHIT \\ H_3 &= SYMPH \cdot \sin(70.0^\circ) - 10 \text{ mm} \\ D_0 &= PALWI \end{aligned}$$

The height of the molars (H_3 and H_4) and the angle of the posterior pharynx wall with respect to the palatal line (A_0) were calculated according to Goldstein [6].

It should be noted that the frames of reference are not exactly the same in Fig. 2 and 3. In the former, the horizontal is defined by the palatal line, but in the latter, it is the tangential approximation between the upper gums and the teeth. However, the angle between these two lines rarely exceeds 8° and the resulting length changes are thus negligible [5].

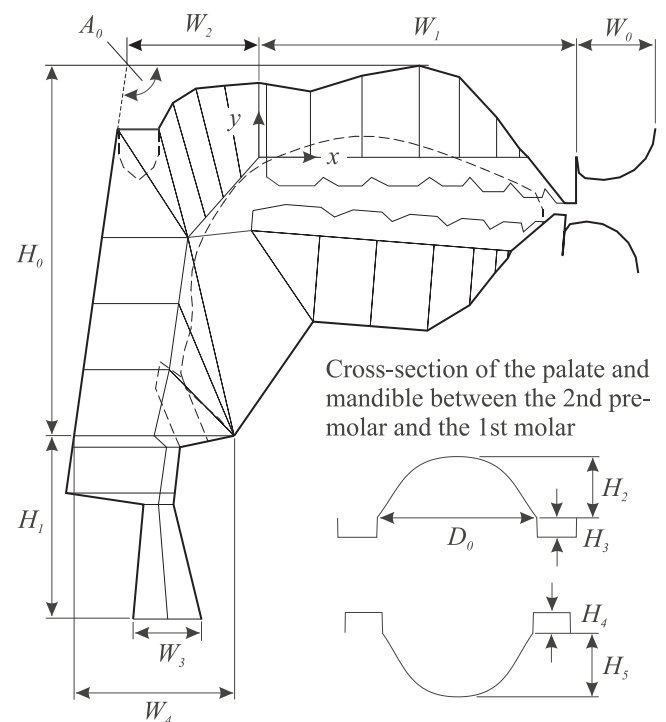


Figure 3: Contour of the vocal tract model with anatomic parameters.

3. ARTICULATORY TRANSFORMATION

While the parameters discussed in the previous section define the basic anatomic dimensions of a speaker's vocal tract, the articulatory state is specified in terms of 23 *vocal tract parameters* [1]. These parameters define the position and shape of the tongue, the lips, the jaw, the velum, and the larynx. Fig. 1b illustrates the influence of

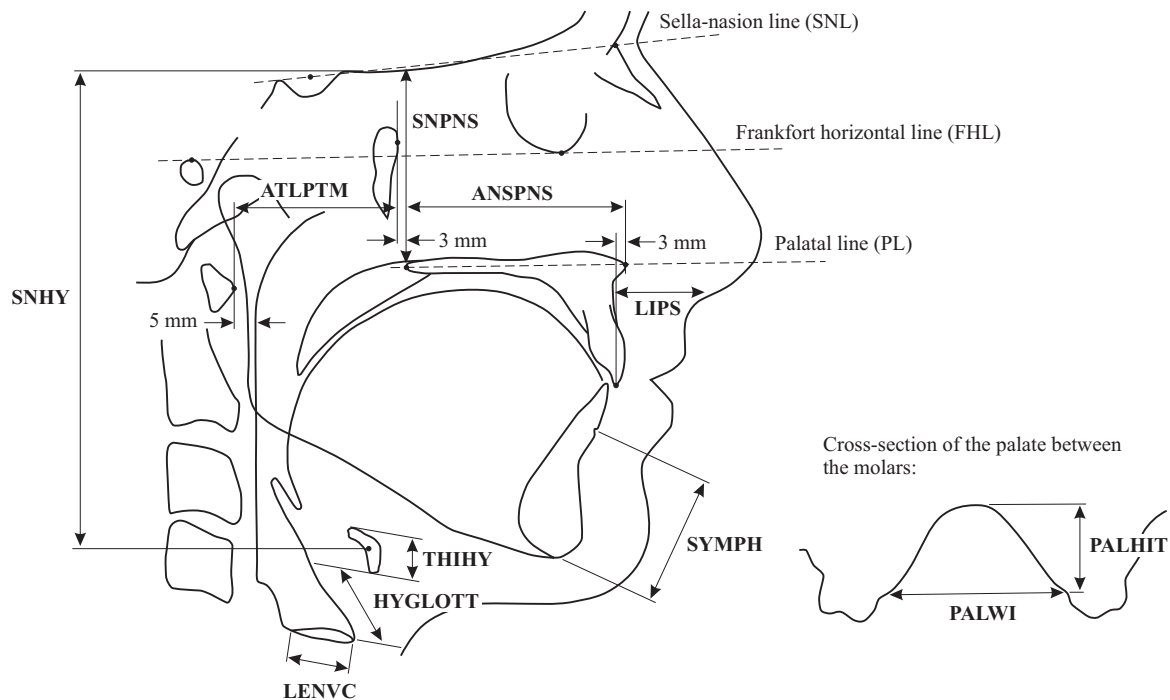


Figure 2: The craniofacial measurements used in this study (summarized from Goldstein [6]). The abbreviations for the measurements and the cephalometric landmarks they refer to are explained in detail in [6].

these parameters on the state of the articulators (cf. [1] for details). By means of magnetic resonance images of our reference speaker, parameter values were determined for all German speech sounds [2].

In the following we will discuss to what extent it is possible to transfer the articulatory state for a given speech sound from the reference vocal tract to a vocal tract with different anatomic dimensions. The most natural thing to try is to tune the vocal tract parameters in such a way that the articulators, especially the tongue, are scaled in accordance with the anatomical length changes of the oral and pharyngeal cavity. To test this approach, we transformed the articulatory state of the vowels [i], [e], [ɛ], [ə], [a], [o], and [u] from the reference vocal tract to that of an 11-year-old boy. When the approach is valid, the synthetic formant frequencies should change from those of an adult man to those of an 11-year-old boy. As shown by Lee *et al.* [8] (and very well confirmed by the formant data in [7]), a *clear* linear-scaling trend of male formant frequencies exists as a function of age. This means that children's formant frequencies can be estimated with a high degree of confidence from those of adult male speakers by multiplication with an age-dependent factor. The scaling factors of *male* speakers are approximately the same for all formants [8]. For an 11-year-old boy, the data in [8] yield a factor of 1.23, and the data in [7] a factor of 1.25. In this study, we assume a factor of 1.24 to calculate the boy's formant frequencies from that of our reference speaker.

The location of the vowel formants in F_1 - F_2 space is shown in Fig. 4. Obviously, there are some major dif-

ferences between the synthetic formants that result from a transformed articulatory state and the "ideal" formants for the boy. The frequency of the synthetic F_1 is for all vowels either equal ([i], [u], [o]) or lower ([e], [ə], [ɛ], [a]) than the target value, the difference being greater for low and mid vowels than for high vowels. According to the acoustic theory of speech production, this suggests that the tongue body position is too high in the synthetic articulations of the low and mid vowels. For the second formant, the deviations do not seem to follow such an obvious pattern. The mean difference between the synthetic and natural vowels averaged over all three formants and all vowels is about 9.1%.

In an attempt to improve the match we performed an automatic search in the vocal tract parameter space for articulatory configurations that result in formant frequencies closer to the natural ones. The search was confined to parameter values in the near vicinity of the values obtained by the above procedure. This was done to consider only those articulatory states that are still similar to the ones obtained by the transformations. Figure 4 shows that the optimized vowels were indeed much closer to the natural vowels. The mean formant error reduced to 2.1%. Figure 5 depicts the synthetic articulatory configurations for the vowels [a], [ə] and [u] before and after the optimization. The optimization for [a] and [ə] confirm the "articulatory correction" predicted above very well, i.e., the need for a lower tongue position. For [u], the improved articulatory state has a slightly retracted tongue in order to compensate for the original difference in F_2 . Note that the outline for the vowel [a] is shown together

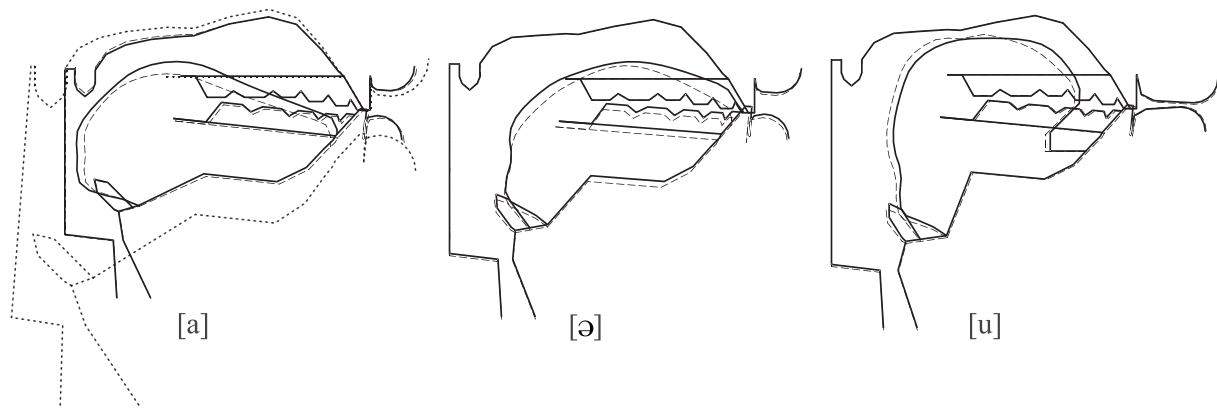


Figure 5: Synthetic articulatory configurations for the vowels [a], [ə], and [u] by an 11-year-old boy. Configurations generated by proportional scaling the corresponding articulations of the adult reference vocal tract are drawn with solid lines, and those after acoustic optimization with dashed lines.

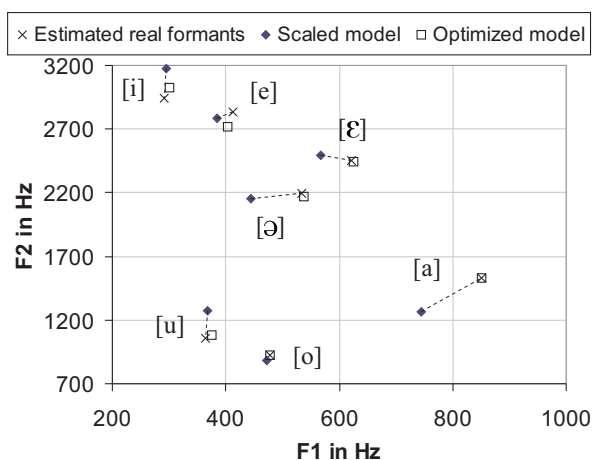


Figure 4: Formant chart for seven vowels of an 11-year-old boy. The (estimated) real formant frequencies (crosses) are those obtained by scaling the synthetic formants of the adult reference vocal tract by the factor 1.24. Synthetic formant frequencies are shown after proportional scaling of the articulatory states (black diamonds) and after optimization (squares).

with the vocal tract outline of an adult male speaker to demonstrate the different proportions between the pharyngeal and oral cavities of children and adults.

4. DISCUSSION AND CONCLUSIONS

We have developed a 3D model of the vocal tract that can change its anatomy and articulation to that of any male or female speaker between about 1 and 20 years. It was shown that a simple transformation of the size and position of the model articulators proportional to the length changes of the oral and pharyngeal cavity cannot account for the formant differences between adults and children. Instead, our results indicate in accordance with Boë *et al.* [4] that children use slightly different articulatory configurations than adults to produce the same phonemes. The major difference in the articulation of vowels seems to be that children use *relatively* lower tongue positions than adults for the articulation of low and mid vowels.

However, the results must be regarded as preliminary and need further examination.

5. ACKNOWLEDGMENTS

This research was funded by grant no. JA 1476/1-1 from the German Research Foundation. We would like to thank Louis-Jean Boë from the ICP in Grenoble for helpful discussions.

6. REFERENCES

- [1] Birkholz, P., Jackèl, D., Kröger, B. J. 2006. Construction and control of a three-dimensional vocal tract model. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)* Toulouse, France. 873–876.
- [2] Birkholz, P., Kröger, B. J. 2006. Vocal tract model adaptation using magnetic resonance imaging. *7th International Seminar on Speech Production (ISSP'06)* Ubatuba, Brazil. 493–500.
- [3] Boë, L.-J., Maeda, S. 1998. Modélisation de la croissance du conduit vocal. *Journées d'Études Linguistiques, La voyelle dans tous ses états* Nantes, France. 98–105.
- [4] Boë, L.-J., Ménard, L., Maeda, S. 2000. Adaptation of control strategies during the vocal tract growth inferred from simulation studies with an articulatory model. *Proceedings of the 5th Seminar on Speech Production* Kloster Seeon, Germany. 277–280.
- [5] Enlow, D. H. 1982. *Facial Growth*. W. B. Saunders Company.
- [6] Goldstein, U. G. 1980. *An Articulatory Model for the Vocal Tracts of Growing Children*. PhD thesis Massachusetts Institute of Technology.
- [7] Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97(5), 3099–3110.
- [8] 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.
- [9] Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., Yandell, B. S. 2005. Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America* 117(1), 338–350.