A three-dimensional model of the vocal tract for speech synthesis

Peter Birkholz and **Dietmar Jackel**

Institute for Computer Graphics, Department for Computer Sciences, University of Rostock
18055 Rostock, Germany; e-mail: {piet,dj}@informatik.uni-rostock.de

Abstract

A three-dimensional model of the vocal tract is presented. The vocal tract walls and the tongue are represented by three individual grids. The shape of the grids is determined by a set of parameters specifying the form and position of the tongue, the lips, the velum, the larynx and the jaw. Both articulatory speech synthesis and the visualization of speech production can benefit from this three-dimensional representation of the vocal tract.

1 Introduction

The increasing use of magnetic resonance imaging (MRI) in speech research has let to the development of the first three-dimensional models of the vocal tract (e.g., [2], [1]). These 3D-models are based on large amounts of MRI-data and are therefore difficult to replicate. We have designed a 3D-model of the vocal tract which does not reflect the vocal tract anatomy of one particular person, but has a clearly arranged structure and preserves the advantage of the third dimension compared to traditional 2D-models (e.g., [3]). The main advantage in the context of speech synthesis is, that the vocal tract area function can directly be obtained from the model. However, even a comprehensive three-dimensional simulation of sound wave propagation is conceivable.

The shape of the proposed model is based on measurements on midsagittal X-ray films and diverse anatomical data that can be found in the literature. The parameters defining the shape and position of the articulators are mainly based on the proposals of Mermelstein [3], but have been extended to take account of the three-dimensional geometry.

2 Model Description

2.1 General structure

The proposed model is composed of three grids, forming different parts of the vocal tract (see Fig. 1 (a)).

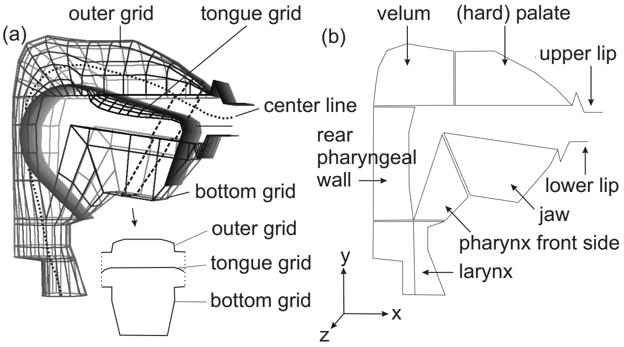

Figure 1: Wire frame representation of the model in (a) and segmentation of the geometry in (b).

The outer grid represents the upper-back wall of the vocal tract. It consists of the upper lip, the (hard) palate, the velum and the rear pharyngeal/laryngeal wall. The bottom grid, consisting of the lower lip, the jaw and the front side of the larynx, forms the lower-front part of the vocal-tract. The tongue surface is represented by the tongue grid. All three grids are assumed to be symmetric with respect to the midsagittal plane. Therefore, we only describe the geometry of the left half of the vocal tract.

The palate (cf. Fig. 1 (a)) is a little reminiscent of an upside-down boat hull, which is shaped by means of a number of ribs. We use this concept to define the shape of all parts of the model by means of ribs. The outer grid has 27 ribs and the bottom and tongue grid have 18 and 36 ribs, respectively. The ribs, which are forming the left half of the vocal tract, are defined by 7 points for the outer grid, 5 points for the bottom grid and 12 points for the tongue grid. Following the terminology in boat building, we call the piece-wise linear connections of corresponding rib points “stringers”. Thus we have 7, 5 and 12 stringers on the outer, bottom and tongue grid, respectively.

2.2 Vocal tract wall geometry

The outer and bottom grid is divided in sub-grids as shown in Fig. 1 (b). We will define the shape of these grids in separate local coordinate systems first and then combine them in a global coordinate system. Figure 6 illustrates the geometry of the sub-grids. The

ISBN 1-876346-48-5 © 2003 UAB

ribs and stringers of each individual grid are denoted by $Q0$, $Q1$,... and $H0$, $H1$, ..., respectively.

The palate and the jaw have a fixed shape in contrast to the larynx, jaw and lips, whose shape depends on several parameters. The shape of the velum is specified by the parameter VO ($0 \leq VO \leq 1$). A maximal raised and a maximal lowered velum correspond to $VO = 0$ and $VO = 1$, respectively. The orientation of the ribs for these two states is shown in Fig. 6. Any states between them (for $0 < VO < 1$) are obtained by linear interpolation. The ribs are assumed to have an elliptic shape with a minimal z-value of -1.8 for $Q6$ and -1.5 for $Q0$.

The shape of the larynx is specified by the parameter HX ($2.0 \leq HX \leq 2.75$ cm) corresponding to the horizontal position of the hyoid. According to Mermelstein [3], the hyoid position slightly influences the midsagittal distance in the laryngeal region. Therefore, we define the shapes for the "maximal wide" and the "maximal narrow" larynx, and obtain the actual shape for a given HX-value by linear interpolation again. The ribs have an elliptic shape again, and are illustrated for $Q0$ and $Q5$.

The shape of the lips is specified by two parameters. The protrusion is expressed by LP ($0.8 \leq LP \leq 1.6$ cm) and equals $p + q$ (with $p : q = 1 : 2$). The active opening of the lips is given by LH ($-0.8 \leq LH \leq 0.5$ cm) and corresponds to the height h in Fig. 6. In the local coordinate system, the upper and lower lip are symmetric with respect to the x, z-plane. The width w of the lips for $Q3$ is assumed to be a function of the lip protrusion and is given by $w = 2.8 - 1.625 \cdot LP$.

The combination of the locally defined sub-grids in the global coordinate system is illustrated in Fig. 2.

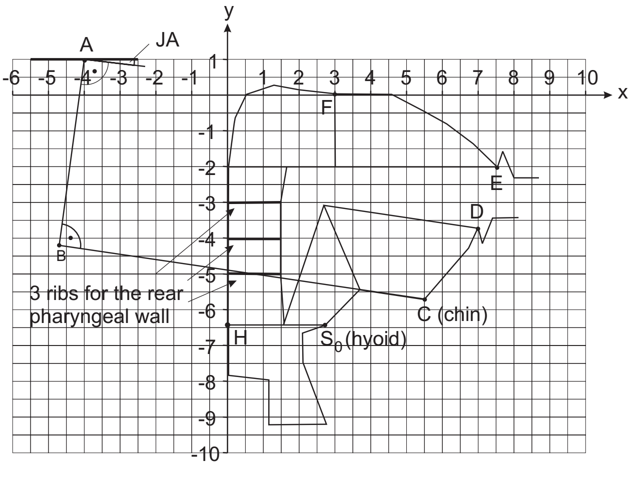

Figure 2: Combination of the locally defined sub-grids in the global coordinate system.

The local origins of the hard and soft palate are simply moved to the fixed point F, and the local origin of the larynx is moved to the point H. The y-value of H is specified by the parameter HY ($-7.0 \leq HY \leq -6.1$ cm) and allows the vertical position of the larynx to change. The local abscissa of the jaw is aligned along the straight line BC. The distance $\overline{BC}$ equals $10.3 + JP$ cm ($-0.5 \leq JP < 0.3$ cm). Forward/backward movements are thus expressed by JP. The opening of the jaw is specified by the angle JA ($-0.15 \leq JA \leq 0.0$ rad). The upper and lower lip are attached to the palate in the point E and to the jaw in the point D, respectively. The rear pharyngeal wall is defined by three fixed ribs, which are elliptic in shape. The expansion of the ribs in the z-direction are 1.2, 1.4 and 1.5 cm for the lowest, middle and highest rib, respectively. The pharynx front side is implicitly defined by the last rib of the larynx front side and the first rib of the jaw.

2.3 Tongue geometry

The tongue surface is modeled in two steps. Firstly, we define the contour in the midsagittal plane and then we attach the ribs to the contour. The contour definition is illustrated in Fig. 3.

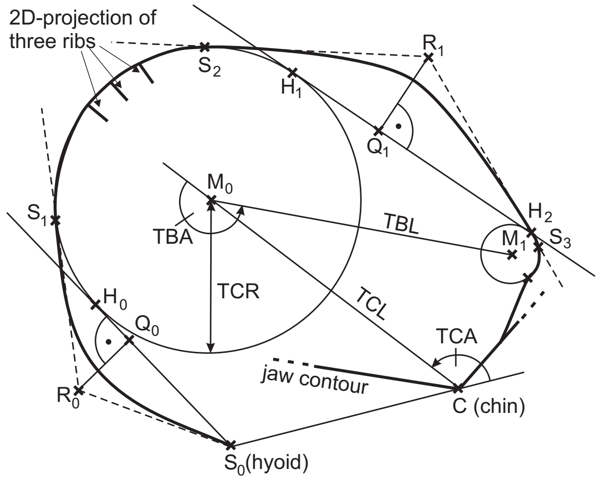

Figure 3: Tongue contour.

The outline consists of a big and a small circle representing the tongue body and the tongue tip. Two quadratic splines are forming the tongue root and the tongue blade. The position M_0 of the tongue body is defined by the angle TCA ($1.47 \leq TCA \leq 3.14$ rad) and the distance $\overline{CM_0} = TCL$ ($2.5 \leq TCL \leq 6.0$ cm) with respect to the straight line between the hyoid (S_0) and the chin (C). The hyoid and the chin are lying on the bottom grid contour (cf. Fig. 2). The position M_1 of the tongue tip is defined in the same way by the angle TBA ($3.14 \leq TBA \leq 4.71$ rad) and the distance $TBL = \overline{M_0M_1}$ ($1.9 \leq TBL \leq 5.0$ cm) with respect to CM_0. The small circle has a fixed radius of 4 mm and the radius of the big circle is determined by the parameter TCR ($1.5 \leq TCR \leq 2.25$ cm).

The splines are defined by the points $S_0R_0S_1$ and

$S_2R_1S_3$. The position of R_0 is specified by the parameters (coordinates) TAX and TAY with respect to the tangent S_0H_0 and can be on either side of the tangent. The position of R_1 on the other hand is specified by the parameters TBX and TBY with respect to the straight line H_1H_2. This line is a tangent to both the big circle and the small circle. S_1, S_2 and S_3 are tangents to the circles, too. The shape of the tongue ribs is defined in a local coordinate system as illustrated in Fig. 4.

The shape of each rib is determined by two (local) parameters s and t. The value t ($0 \leq t \leq 1$) defines the (relative) width of the tongue and s ($0 \leq s \leq 1$) varies the (relative) depth of a tongue groove. The values of s and t for the individual ribs depend on the vocal tract parameters RTW and RTG, which allow to modify the width of the tongue blade and tip (for example for the laterals) and the depth of a tongue blade groove (for the production of fricatives). The locally defined ribs are transferred into the global coordinate system by attaching the origin of the local coordinate systems to equally spaced points on the tongue contour (cf. Fig. 3). The x'-axis is aligned in the $-z$-direction and the y'-axis is oriented perpendicular to the tongue contour.

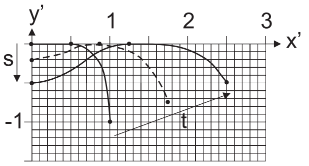

Figure 4: Tongue rib.

2.4 Vocal tract center line and area function

The vocal tract center line is assumed to lie in the midsagittal plane and is calculated in three steps. The first crude approximation to the center line is the μ-line shown in Fig. 5 (a).

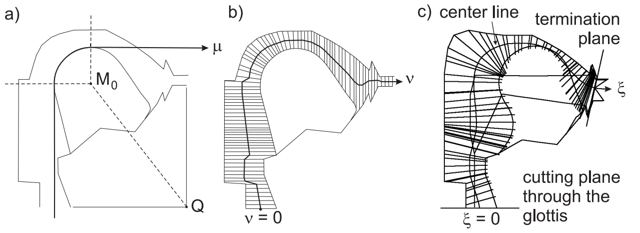

Figure 5: Calculation of the vocal tract center line.

It consists of a vertical line section, a horizontal line section and a 90°-circular arc around the tongue body. Any normal of the μ-line between the glottis and the lips intersects the upper and lower vocal tract contour. Several of these normals are shown in Fig. 5 (b). The positions of the normals along the μ-line are chosen in such a way, that at least one normal runs through every grid point of the piece-wise linear contours. When the mid-points between the intersections with the upper and lower contour are connected, we obtain the ν-line, which gives a good approximation to the center line. However, the ν-line has some unwanted sharp "cracks". Therefore, we apply a low-pass filter to the ν-line in order to obtain the smoothed final center line shown in Fig. 5 (c).

The vocal tract area function can now be sampled by intersecting the stringers of the model with planes that are aligned perpendicular to the center line. The resulting cuts will look similar to the one in Fig. 1 (a), and the area above the tongue in each cut must be calculated.

3 Conclusion

Our goal was to create a vocal tract model that can contribute to more naturalness and intelligibility of synthetic speech. The complexity of the model should only increase moderately compared to traditional 2D-models. The qualification of our model for speech synthesis was confirmed in preliminary experiments with a synthesizer similar to [4]. Vowel sequences generated from the vocal tract area functions sounded very natural and could be calculated in real-time on a 2 GHz Pentium 4 computer. Apart from the area functions, we also computed the "circumference functions" which are important for the energy losses at the vocal tract walls. However, the current number of 18 vocal tract parameters is very high. We aim to reduce this number to 12-14 by utilizing dependencies between individual parameters, especially for the tongue.

4 Acknowledgment

We are grateful to K. G. Munhall and the ATR Human Information Processing Laboratories in Kyoto, Japan, for supplying us with the "X-ray film database for speech research".

References

[1] Dang, J., and Honda, K. (**2000**). "Estimation of vocal tract shape from speech sounds via a physiological articulatory model," Proceedings of the 5th Seminar on Speech Production, Bavaria, pp. 233-236

[2] Engwall, O. (**2002**). "Tongue Talking - Studies in Intraoral Speech Synthesis," Doctoral Dissertation, Royal Institute of Technology, Department of Speech, Music and Hearing, Stockholm

[3] Mermelstein, P. (**1973**). "Articulatory model for the study of speech production," Journal of the Acoustical Society of America 53(4), pp. 1070-1082

[4] Sondhi, M. M., and Schroeter, J. (**1987**). "A hybrid time-frequency domain articulatory speech synthsizer," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 35, no. 7, pp. 955-967

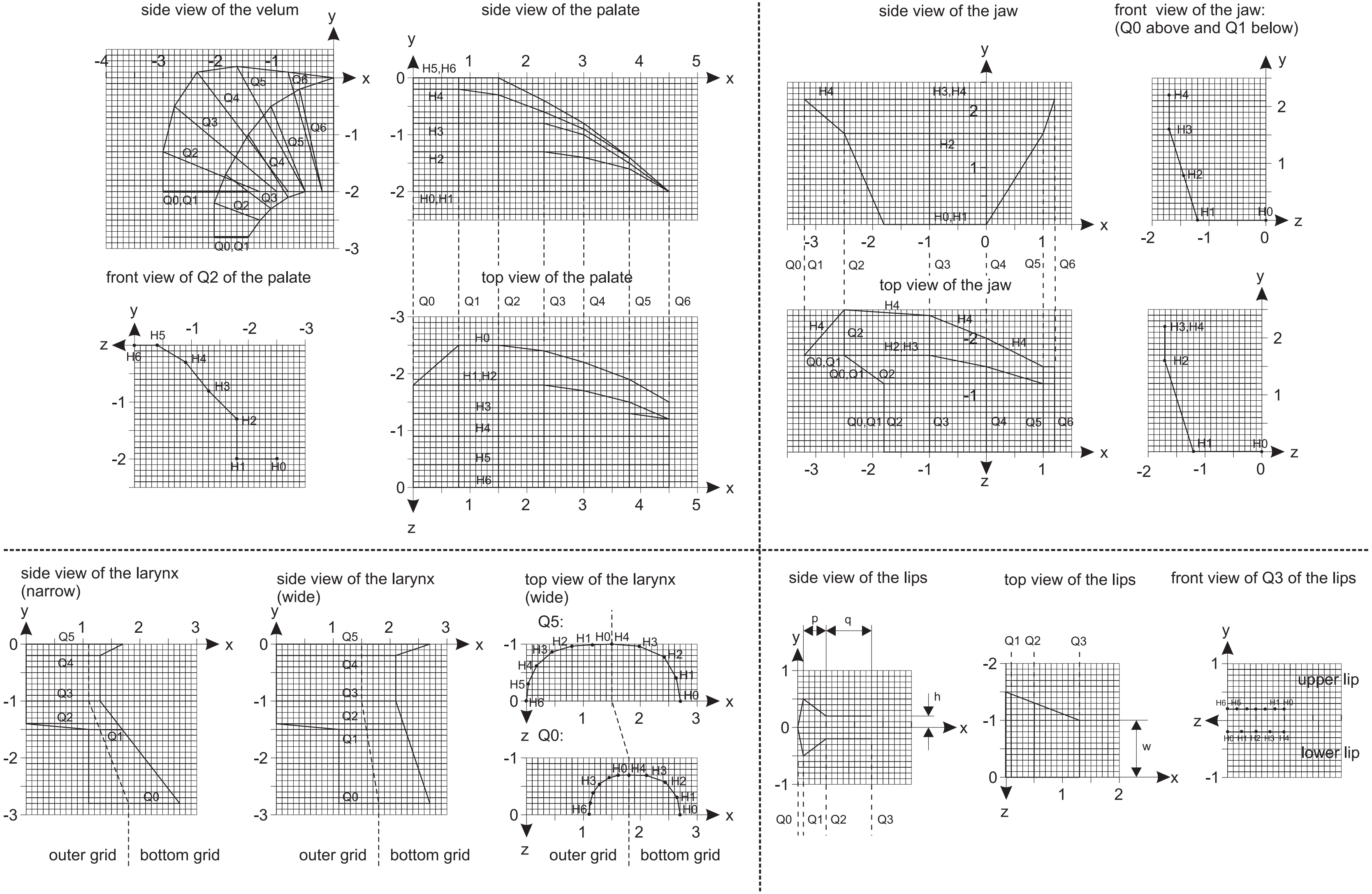

Figure 6: Shape of the velum, palate, jaw, larynx and lips, defined in local coordinate systems. Ribs are denoted by $Q0, Q1, \ldots$ and stringers by $H0, H1, \ldots$.