CONSTRUCTION AND CONTROL OF A THREE-DIMENSIONAL VOCAL TRACT MODEL

*Peter Birkholz, Dietmar Jackèl**

Institute for Computer Science
University of Rostock
Germany

Bernd J. Kröger

Department of Phoniatrics, Pedaudiology,
and Communication Disorders,
University Hospital Aachen, Germany

ABSTRACT

We present a novel 3D vocal tract model and a method to control the articulatory movements of the model. The vocal tract model consists of 7 wireframe meshes that represent the three dimensional surfaces of the articulators and the vocal tract walls. 23 parameters determine the shape of the meshes. The articulatory movements in terms of the parameter curves are generated from a gestural description of an utterance. The work presented here is an integral part of a complete articulatory speech synthesizer for high quality synthesis.

1. INTRODUCTION

Vocal tract models can be divided into two and three dimensional models on one hand, and into geometric, statistical and biomechanical models on the other hand. 2D models define the vocal tract shape in the midsagittal plane by the contour lines of the articulators whereas 3D models provide the genuine 3D shape of the vocal tract. With 2D models, the area function that is needed for calculating the speech sounds must be inferred solely from midsagittal distances. This obviously requires some sort of empiric transformation that complements the missing information. 3D models do not depend on such approximations. Furthermore, 2D models are not able to represent lateral consonants, whereas this does not pose a problem for a 3D model. Many of the newer vocal tract models for the study of speech production are therefore 3D models.

Statistical 3D models (e. g. [1]) have the advantage of relatively few uncorrelated parameters. However, these models require huge amounts of MRI or CT data for their construction and they are usually specific for a particular speaker. *Biomechanical* vocal tract models simulate the behaviour of the articulators by means of finite element methods (e. g. [2]). They are especially suited to facilitate new insight in the relation between muscle activation and articulatory movements. On the other hand, they have many degrees of freedom, are difficult to control and require much computational power. *Geometric* vocal tract models are similar to statistical models in that their parameters define the vocal tract shape directly in geometrical terms, but the kind and number of parameters is choosen *a priori* and fitted to particular data *a posteriori*.

In this paper, we present a novel *3D geometric* vocal tract model in combination with a *gestural dominance* control model. Compared to our previous 3D vocal tract model [3], the new model has a more intuitive visual appearance and allows a better control of the jaw, tongue and lips (Sec. 2). The control model was devised as a simple and effective method to generate the parameter curves for whole utterances and is presented in Sec. 3.

2. VOCAL TRACT MODEL

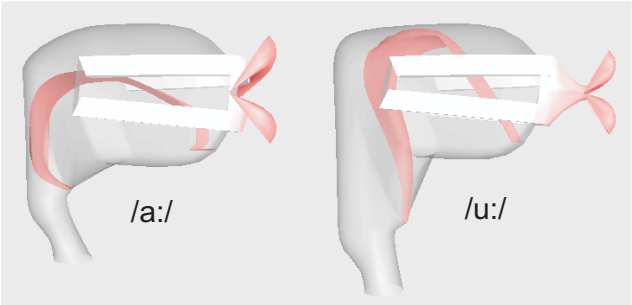

Fig. 1. Vocal tract geometry for the vowels /a:/ and /u:/.

The vocal tract model is composed of 7 wireframe meshes representing different model parts. One mesh, the "upper cover", forms the palate and the posterior wall of the larynx and the pharynx. The "lower cover" represents the anterior side of the larynx and pharynx and the lower jaw. The 5 remaining grids form the upper and lower lip, the upper and lower teeth and the tongue. Figure 1 shows renderings of the vocal tract geometry for the vowels /a:/ and /u:/ with a transparent upper and lower cover.

Figure 2 shows the wireframe representation of the individual parts of the vocal tract model. X-ray pictures of a male English speaker served as templates for the sagittal outline of the model, and the lateral shape and dimensions were set to typical values given in the literature. The rigid parts of the vocal tract have a static shape (like the palate, the mandible

*This research was funded by DFG (grant JA 1476/1-1).

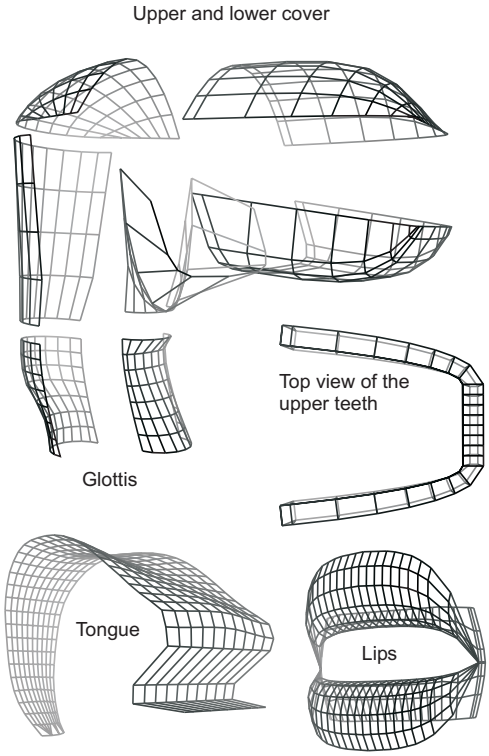

Fig. 2. Parts of the vocal tract model.

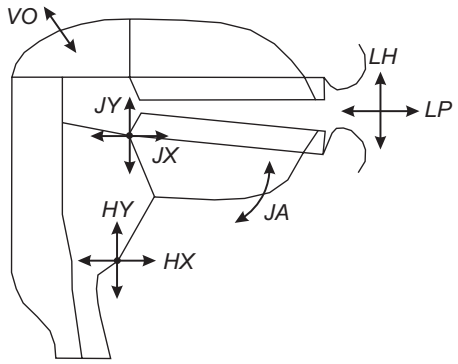

Fig. 3. Vocal tract parameters (except tongue parameters).

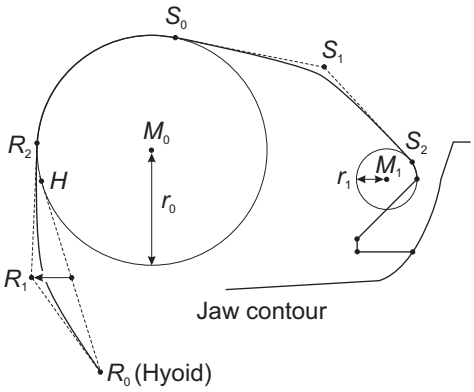

Fig. 4. Midsagittal geometry of the tongue.

and the teeth), while the shape of the other parts depends on the vocal tract parameters described next.

The shape of the velum and hence the size of the velopharyngeal opening is determined by one paramter, VO, ranging from -1 to +1. We defined different shapes for $VO = -1$ (velum tightly closed), for $VO = 0$ (velum just closed) and for $VO = 1$ (velum wide open), and interpolate between these shapes for intermediate values. The lips depend on two parameters, LP and LH, defining the protrusion of the lip corners and the vertical distance between the upper and lower lip. From these lip parameters we derive all other important lip dimensions according to Abry et al. [4] and therewith construct the lip meshes. The position and orientation of the rigid lower jaw ist determined by three parameters. JX and JY define its position in Cartesian coordinates and JA determines the opening angle. The position of the hyoid is defined by two parameters in accordance with the vocal tract model by Mermelstein [5]. HX defines the horizontal and HY the vertical position. Since we keep the distance between the hyoid and the glottis constant, the larynx moves up and down conjointly with the hyoid. The parameter HX changes the distance between the hyoid and the posterior wall of the pharynx and such controls the air volume in the larynx. Figure 3 gives an overview of the parameters described above.

Figure 4 shows the midsagittal contour of the tongue, which is essentially composed of two circular arcs and two rational Bézier curves. The big circle with the center M_0 and the radius r_0 constitutes the tongue body, and the small circle with the center M_1 and the radius r_1 constitutes the tongue tip. The positions of the two circles are given by the four parameters TCX, TCY, TTX, TTY, where $M_0 = (TCX, TCY)$ and $M_1 = (TTX, TTY)$. The radius of the tongue body circle is given by the parameter TCR and $r_1 = 0.4$ cm is fixed. The back of the tongue is formed by a Bézier curve defined by the points $R_0R_1R_2$, where R_0 represents the hyoid, R_1 depends on the parameter TRE, and R_2 is the point where the tangent from R_1 at the tongue body contacts the circle. The vertical position of R_1 is fixed at half the distance between R_0 and H, while its horizontal position depends on TRE. The tongue blade is formed by the second Bézier curve defined by the points $S_0S_1S_2$. S_1 is given in Cartesian coordinates by the parameters TBX and TBY, and S_0 and S_2 are the osculation points of the tangents through S_1 at the two circles. Three additional short line segments form the anterior termination of the tongue contour.

The elevation of the tongue sides relative to the midsagittal outline of the tongue is defined at four equidistant locations along the midsagittal outline by four parameters: $TS1 \ldots TS4$. They determine the elevation in the region of the tongue root, in the center of the tongue dorsum and the tongue blade, and at the tongue tip. In intermediate positions along the contour,

the tongue side elevation is interpolated.

The vocal tract parameters described above were chosen to provide a very flexible control over the geometric properties of the vocal tract model. However, this flexibility was achieved at the expense of a large *number* of parameters compared to typical statistical vocal tract models. Therefore, some of our parameters are probably correlated and therefore redundant. The possibility to individually set correlated parameters holds the risk that physiologically unrealistic vocal tract shapes are generated. However, instead of trying to eliminate possible dependencies among the parameters *a priori*, we decided to leave the responsibility for the avoidance of unrealistic vocal tract shapes to the control model described in the next section.

3. CONTROL OF THE VOCAL TRACT MODEL

In this section we briefly describe the control model for the generation of the articulatory movements. As in the well-known task-dynamic model by Saltzman *et al.* [6], the articulatory gesture is the basic unit of articulatory action in our model. However, both the specification of gestures and their transformation into parameter trajectories differ from the task-dynamic model. In our approach, gestures generally describe the coordinated movement of one or more articulators within a specific time interval. A typical example is the formation and release of the oral constriction for a stop consonant. A gesture in our approach defines the temporal course of the *degree of realization* of such an action by means of a realization function $y(t)$. A realization function is similar to a trapezoid pulse with smoothly raising and falling edges. Therefore, the parameters of a gesture are the durations of the onset phase, the steady state portion and the offset phase, as well as the amplitude of the pulse [7].

An utterance is represented as a coordinated structure of the constituting gestures, called a gestural score. What a gestural score looks like in our system is illustrated in Fig. 5 for the utterance /ipa/. Here, each gesture is represented by its realization function.

We distinguish between six types of gestures: Vocalic, consonantal, glottal, and pulmonary gestures as well as F_0-phrase and accent commands. In this paper, we focus on the vocalic and consonantal gestures, because only they affect the state of the supraglottal system and therefore determine the vocal tract parameters. The remaining four gesture types affect the excitation of the vocal tract (pulmonary pressure, abduction/adduction of the glottis, F_0 for voiced excitation) and are discussed in detail in [8]. The distinction between vocalic and consonantal gestures reflects the intrinsic differences between both classes of sounds. While vowels are characterized by approximately invariant vocal tract configurations, only a reduced set of articulators is needed for the realization of consonants, giving rise to vowel-consonant coarticulation. Figure 5 shows two vocalic gestures for the realization of the

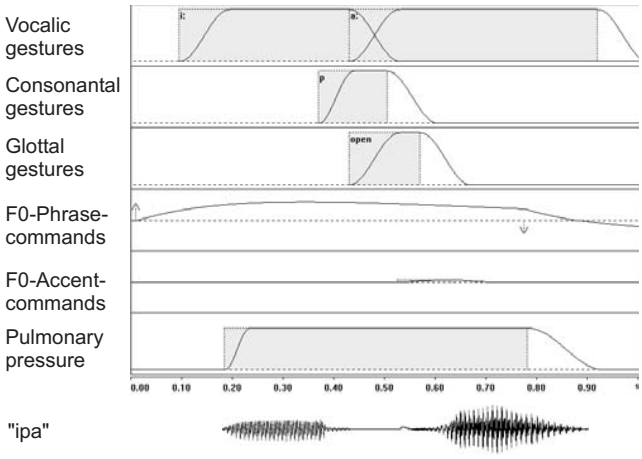

Fig. 5. Gestural score for the utterance /ipa/.

vowels /i:/ and /a:/ superimposed with a consonantal gesture for /p/. The /p/-gesture is accompanied by a glottal gesture that is responsible for the abduction and adduction of the vocal folds.

The transformation from the gestural score to the vocal tract parameter trajectories is achieved with a dominance model that is described next. Generally, each vocalic and consonantal gesture is associated with a predefined set of parameter values and such with a specific vocal tract target shape. We set up the target parameters manually by matching the midsagittal contours of our vocal tract model with sagittal x-ray contours. While the target shapes are approximately invariant for vowels, the articulatory realization of consonants is strongly influenced by the underlying vocalic context. Our consonantal target parameters roughly represent the vocal tract shapes for the consonants produced in schwa-context at the moment, when the constriction/closure is maximally narrow. In order to consider the coarticulatory influence of the vocalic context, each parameter of a consonantal target is complemented with a dominance value between 0.0 to 1.0. When the dominance of a particular parameter is high, then it is important for the articulatory realization of the consonant. For example, the parameters TTX and TTY for the tongue tip position have a high dominance for the consonants /n/ and /t/ but a low dominance for /p/ or /k/.

Given a gestural score, the calculation of the vocal tract parameters at a particular moment t_0 works as follows. First, all parameters are initialized with the values for the neutral vowel schwa: $p^{(0)}(i) := n(i)$. Here, $n(i)$ denotes the value of the parameter i of the neutral vowel. When a vocalic gesture with the realization degree y_v $(0 \leq y_\mathrm{v} \leq 1)$ exists at the moment t_0, the initial parameter values $p^{(0)}(i)$ are replaced by $p^{(1)}(i) := y_\mathrm{v} v(i) + (1 - y_\mathrm{v}) p^{(0)}(i)$, where $v(i)$ are the parameter values for the target shape of the vocalic gesture. When two vocalic gestures with the target parameter values $v_1(i)$ and $v_2(i)$ and the realization degrees $y_{\mathrm{v}1}$ and

y_{v2} overlap at time t_0, the parameters are merged according to $v(i) = [v_1(i)y_{v1} + v_2(i)y_{v2}]/(y_{v1} + y_{v2})$. The resulting realization degree is $y_v = \min\{y_{v1} + y_{v2}, 1\}$. The vocalic gestures themselves are superimposed by the consonantal gestures that in turn modify the parameter values $p^{(1)}(i)$.

Assume, a consonantal gesture with the realization degree y_c exists at time t_0. This gesture is not only associated with a target shape defined by the parameters $c(i)$ but also with corresponding dominance values $d(i)$. The vocal tract parameters resulting from the superimposed consonantal gesture are calculated as $p^{(2)}(i) := \mu(i)c(i) + [1 - \mu(i)]p^{(1)}(i)$, where $\mu(i) = y_c d(i)$. Hence, a vocal tract parameter is affected by a consonantal gesture the more the greater the product of the realization degree *and* the dominance of the parameter. When two consonantal gestures overlap, like for example in the consonantal cluster /sp/, each parameter is modified by the gesture with the greater dominance value for that parameter first, and then by the gesture with the smaller dominance value.

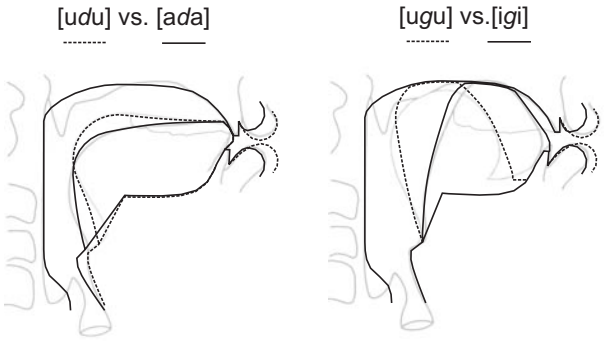

Fig. 6. Simulation of coarticulation.

An important point for a realistic coarticulatory behaviour is the correct setting of the dominance values for the individual consonantal targets. For now we manually adjusted these values for a good fit between sagittal model contours and x-ray contours of consonants in different vocalic environments. Fig. 6 shows the midsagittal contours of our model for the consonant /d/ in /u/- and /a/-context, and for the consonant /g/ in /u/- and /i/-context calculated with the adjusted dominance values. The figure shows that the context dependent change of the articulatory positions is well simulated by our vocal tract/control model (e. g., for /g/, the tongue body is more anterior in /igi/ than in /ugu/). For comparision, x-ray contours measured by Oehman [9] for the same articulations are plotted in gray.

4. CONCLUSIONS

We have presented a novel 3D geometric vocal tract model and a gestural dominance model to generate the parameter curves. The vocal tract model is both flexible and accurate in representing the complete 3D vocal tract shape. The control model combines elements of gesture and dominance based techniques, is conceptually simple and permits the realistic simulation of (co-)articulatory movements. The models presented here have not only proven to yield realistic visual representations of speech, but also natural sounding speech output in combination with an acoustic simulation method [10]. Our medium-term goal is to create a complete *articulatory* TTS system including the presented models. Current work is focusing on the adaption of the vocal tract geometry to a particular speaker of German by means of MR-images and on the automatic calculation of the dominance values for consonantal vocal tract targets.

5. REFERENCES

[1] Olov Engwall, *Tongue Talking: Studies in Intraoral Speech Synthesis*, Ph.D. thesis, Royal Institute of Technology, Stockholm, 2002.

[2] Jianwu Dang and Kiyoshi Honda, "Construction and control of a physiological articulatory model," *Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 853–870, 2004.

[3] Peter Birkholz and Dietmar Jackèl, "A three-dimensional model of the vocal tract for speech synthesis," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2003, pp. 2597–2600.

[4] Christian Abry and Louis-Jean Boë, "Laws for lips," *Speech Communication*, vol. 5, pp. 97–104, 1986.

[5] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[6] E. L. Saltzman and K. G. Munhall, "A dynamic approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, pp. 333–382, 1989.

[7] Bernd J. Kröger, Georg Schröder, and Claudia Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1878–1889, 1995.

[8] Peter Birkholz, *3D-Artikulatorische Sprachsynthese*, Ph.D. thesis, University of Rostock, 2005.

[9] Sven E. G. Öhman, "Numerical model of coarticulation," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320, 1967.

[10] Peter Birkholz and Dietmar Jackèl, "Artikulatorische Sprachsynthese mit dem Programm TractSyn - Ein überblick," in *Fortschritte der Akustik, DAGA '05*, 2005.