Vocal tract modeling in 3D

Olov Engwall

Abstract

A three-dimensional model of the vocal tract is under development. The model consists of vocal and nasal tract walls, lips, teeth and tongue, represented as visually distinct articulators by different colours resembling the ones in a natural human vocal tract. The naturalness of the vocal tract model can be used in speech training for hearing impaired children or in second language learning, where the visual feedback supplements the auditory feedback. The 3D model also provides a platform for studies on articulatory synthesis, as the vocal tract geometry can be set with a small number of articulation parameters, and vocal tract cross-sectional areas can be determined directly from the model.

Introduction

Modeling of the vocal tract has traditionally been limited to two dimensions, both for articulatory synthesis and for visual aids in pronunciation training. Both fields would benefit from incorporating the third dimension in the model, since information is lost when representing the vocal tract only in the midsagittal plane.

Visual aids for articulation training are used for hearing or speech impaired children as well as in some computer guided second language (L2) learning programs to show correct and deviant pronunciation. Extending the model to three dimensions would allow better feedback as lateral variations are shown and the model resembles its human counterpart more closely.

For two-dimensional models used in articulatory synthesis, the area function is predicted using empirically derived formulae relating the cross-sectional area to the midsagittal distance. These formulae postulate the same dependency for all phonemes, but the coefficients differ for different positions along the oral cavity and the pharynx. Using a three-dimensional model allows for direct calculation of the cross-sectional areas, thus avoiding the simulation of a third dimension.

Using the interface for visual speech synthesis developed at TMH by Beskow (1995) an animated 3D model of the vocal tract has been created. This allows for studies of intraoral articulatory movements aiming at improving the visual and articulatory speech synthesis.

The vocal tract model

The model consists of a three-dimensional polygon mesh divided into five different parts, representing vocal and nasal tract walls, lips, teeth and tongue, as shown in Figure 1. The mesh contains 750 vertices joined by approximately 1000 polygons. In accordance with the standard of Beskow (1995), each polygon is defined as an array of three or four vertices, given in Cartesian coordinates, as illustrated in Figure 2.

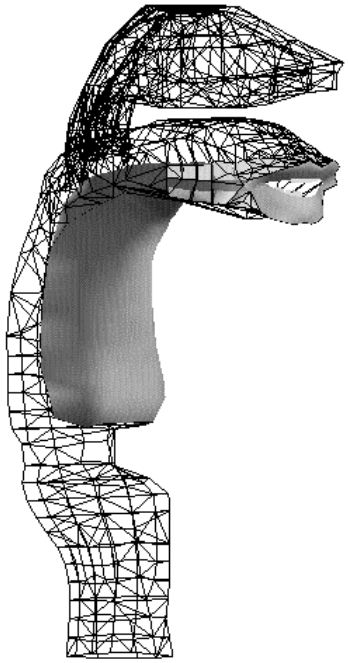
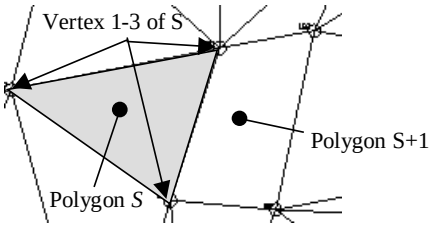


Figure 1. The vocal tract model shown with the oral and nasal walls in wireframe mode and the lips, teeth and tongue in grey scale.

To reduce model complexity the model has been made symmetrical around the midsagittal plane, assuming that relevant articulations can be modeled symmetrically (Engwall, 1998). This assumption is obviously not true for many human



$$\begin{cases} Array\ of\ vertices: \vec{v}_i = (x_i, y_i, z_i) \\ Array\ of\ polygons: \vec{S}_j = \{\vec{v}_{j1}, \vec{v}_{j2}, \vec{v}_{j3}, (\vec{v}_{j4})\} \\ Arrays\ of\ parameter\ weights\ (deformation\ D): \\ \vec{w}_D = \{w_{Di}, ...\}, w_{Di} = 1\ \ for\ prototype. \end{cases}$$

Figure 2. Definitions of polygons, vertices and deformation coefficients.

articulations, but is generally an acceptable approximation.

The vocal tract model is fully compatible with the talking heads at TMH, thus allowing the vocal tract to be combined with a facial model (Figure 3). This facility will be useful in a visual aid for articulation training, as the vocal tract anatomy becomes clearer when seen in spatial relation to its corresponding face (refer to the section on Intraoral visual speech below for details on visual aids in speech training).

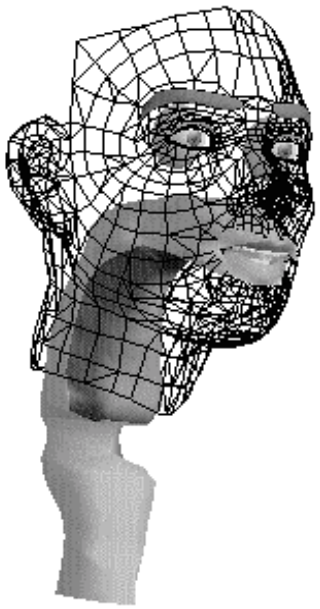


Figure 3. The vocal tract model added to the synthetic face Alf, presented in wireframe.

Articulatory parameters

Ten parameters are used in the articulatory description of phonemes. These are larynx height, jaw opening, lip protrusion, lip rounding, velum movement and five parameters describing the placement of the tongue and separate movements of its parts.

The description of parameters follows, to a large extent, the idea for representing articulator movements outlined by Mermelstein (1973), but has been modified to suit the three-dimensional model, as described below.

Each deformation is defined using a single prototype vertex, a target vertex and weighting coefficient for each vertex that should be affected, as outlined in Table 1 and Figure 2. Rotational deformations have in addition a pivot vertex. The defining vertices can be part of the wireframe or virtual points, used only to define deformations. All parameters are shown in Figure 4, unless stated otherwise.

Table 1. Articulatory parameters and typical values of prototype deformation.

Deformation		**Param.**	**Target range (X_t-X_p)**	**No of affected vertices**
Larynx height		Lp,Lt	6.2 mm	40
Jaw opening		LI,Jt,J	27°	250
Lip protrusion		Pp,Pt	5.4 mm	50
Lip rounding		Rp,Rt	8.7 mm	50
Velum		Vp,Vt	6.8 mm	15
Tongue	promot.	Tp,T	17 mm	130
	raising	T,Tr	12 mm	120
	apex	Ap,At,	9.2°	30
	edges	Ep,Et	6.6 mm	30
	dorsum	Dp,Dt	6.4 mm	25

Larynx height

The parameter describing larynx height has been included to account for changes in vocal tract length caused by changes in the pharynx. The parameter allows both contraction and expansion of the pharynx and is described as a translation, determined by the prototype (Lp) and the target (Lt). These points are chosen so as to give a contraction that is approximately along the midline of the vocal tract.

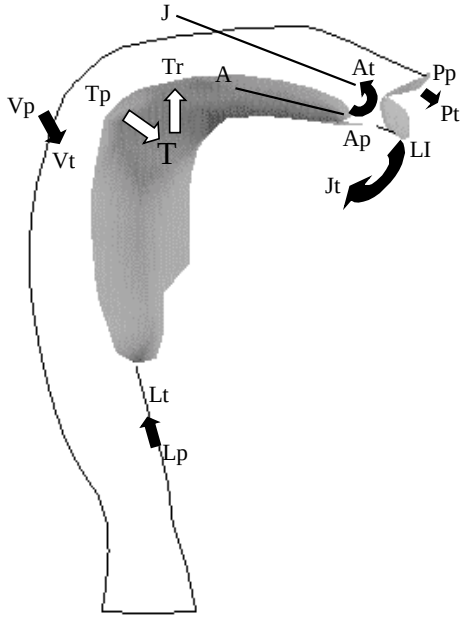


Figure 4. Articulatory parameter definitions in the vocal tract. The tract walls are shown as outlines in the midsagittal plane.

Jaw opening

Human jaw movement in speech shows some minor changes (less than 1 cm) in the length of the axis (J-LI), but in most cases this distance remains constant (Mermelstein, 1973). The jaw opening is thus approximated as a rigid rotation around a virtual point (J). The prototype vertex (LI), the target vertex (Jt) and the pivot vertex (J) determine the degree of opening. The prototype vertex is chosen on the upper part of the lower incisors and both target vertex and pivot vertex are virtual points that give a jaw rotation that is representative for jaw movements in speech. To account for the symmetric movement of the vocal tract, all points mentioned are in the midsagittal plane.

Lip protrusion

Lip protrusion is effectuated through a translation of vertices on or near the lips. The prototype (Pp) and target (Pt) vertices are set in accordance with the definition of protrusion in the facial model (cf. Beskow, 1995).

Lip rounding

As for protrusion, the prototype (Rp) and target (Rt) agree with the corresponding definition applied to the synthetic face. The rounding is accomplished as a pull of affected points towards the target point, placed on the vocal tract midline between the lips, as indicated in Figure 5.

Velum

The velum is lowered using a translation, where the prototype (Vp) on the back edge of the velum represents a closed velum when this parameter is 0 and the target point (Vt) defines a maximally lowered velum. This parameter will account for nasal sounds when the nasal cavity is introduced in the synthesis module. As for now, a lowered velum only indicates nasalization visually by changing the colour of the nasal cavity.

Tongue movements

As the tongue is the most movable and complex articulator it is of no great surprise that its description requires the greatest number of parameters.

Several different three-dimensional tongue models have been proposed, notably by Perkell (1974), Stone (1990) and Wilhelms-Tricario (1995). These models represent the tongue physiologically or biomechanically as subparts or finite elements.

The representation of the tongue in this model is more simplified, considering the tongue body as a whole, with the motion of the apex, tongue edges and dorsum superimposed.

a) The tongue body

Two sets of parameters describe the placement of the tongue body; tongue promotion and tongue raising. The promotion parameter moves the tongue forward using a translation defined by moving the prototype (Tp) towards a target (T) within the tongue body.

It has proven convenient to use a translation towards a non-fixed target (Tr) on the tongue dorsum to specify raising of the tongue. The target moves with the tongue body and hence the displacement of the prototype (T) (the point that also serves as target for the tongue promotion) will always cause the uppermost part of the tongue dorsum to move towards the palate.

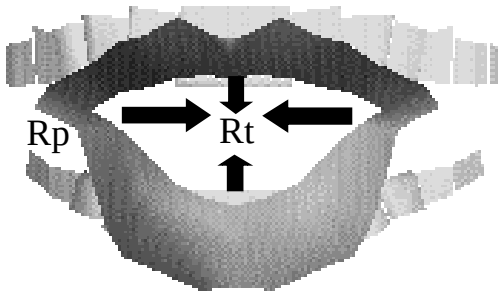


Figure 5. The pull defining lip rounding.

b) The apex

The tongue tip and tongue blade have an additional degree of freedom that is relevant for speech sounds, allowing the anterior part of the tongue to be folded up and back with respect to the tongue body, as can be seen for alveolar or retroflex sounds. Lateral tongue tip movements on the other hand, as when touching the inside of the cheek with the tongue tip are not modeled, since they are not part of normal speech.

The flexion of the tongue tip is effectuated as a rotation around an axis passing by the pivot point (A) on the surface of the tongue dorsum. The prototype (Ap) is placed on the outermost tip of the apex and the target (At) 9.25 degrees counterclockwise from A, as indicated in Figure 4.

c) The tongue edges

In many speech sounds, but most importantly for laterals, the positions of the edges of the tongue deflect from the ones they would have if they followed the tongue body completely. Hence a parameter is needed to account for lowering and raising of the edges with respect to the tongue body. A translation defined to be downward and inward from the prototype (Ep) towards the target (Et) as in Figure 6 allows both lowering and raising of the edges.

d) Dorsum

Tongue grooving is present in many speech sounds and is thus relevant to incorporate in a 3D-tongue model. Actual grooving is much more complex than can be modeled with one parameter as the dorsal height may vary along the tongue. For simplicity, however, the dorsal parameter applies equally to the entire dorsal ridge, raising or lowering it by a translation in the midsagittal plane as defined by prototype (Dp) and target (Dt), shown in Figure 6.

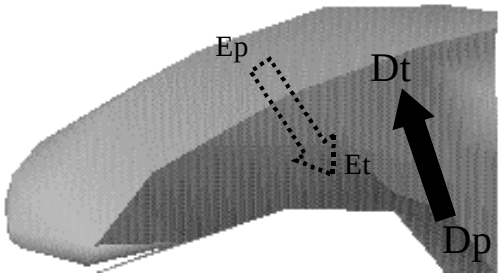


Figure 6. The prototypes and targets for the tongue edge and dorsal movements, indicated in the frontal part of the left tongue half, seen from the cut in the midsagittal plane.

The only constraint imposed on the tongue motion is by limiting the parameter values to the range 0 to 1, or -1 to 1 and hence no physical restraint handles boundary collisions (i.e. tongue surpassing the palate or teeth surface). The method proposed by Cohen et al. (1998) to detect a boundary violation resembles closely how the cross-sectional intersections are calculated in this model (refer to the section on Articulatory synthesis in 3D below) and implementation of detection should therefore be straightforward. Correction, i.e. deciding where vertices violating the boundary condition should be placed, on the other hand, necessitates a more complex method (Cohen et al, 1998). As a consequence correction needs to be tested further in this model before being included.

Generally, the tongue, as well as other parts of the vocal tract would need a larger number of parameters to describe human articulations correctly. Improving the model would include dividing the tongue body into smaller parts with greater internal degrees of freedom. As for the vocal tract walls, a parameter determining the configuration in the lower part of the pharynx might be needed, and labiodentals, such as [f] and [v], would require a parameter that folds the lower lip inwards.

Intraoral visual speech

Visual feedback has always been used in speech training and in communication when noise or some other cause degrades the auditory signal reaching the addressee. The group of addressees includes such different subjects as hearing-impaired persons, second language learners and children with speech problems. A variety of visual aids have been constructed, focusing on prosody as well as pronunciation. Visual feedback in computer learning programs may be in the form of spectrograms (e.g. Nouza & Mádlíková, 1998), colour-coded speech patterns (Öster, 1997) or even computer games (Álvarez et al, 1998), but these give feedback on the acoustical rather than the articulatory features. Mashie (1995) points out that speech training can benefit substantially from visual feedback of articulation. Furthermore, the evaluation studies in the Teleface project (Agelfors et al, 1998) have shown the clear benefit of a synthetic three-dimensional face for lip-reading by hearing-impaired persons or normal hearing persons in noise. This suggests that pronunciation training with a three-dimensional model of the intraoral parts may prove valuable.

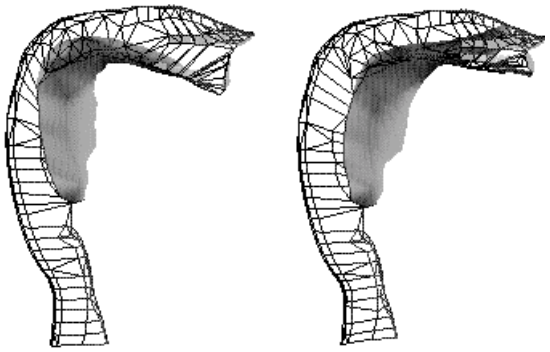


Figure 7. The vocal tract model used to show the different articulations of the vowel [ɑ] (left) and the retroflex consonant [ɖ](right). The walls are shown in wireframe for reasons of resolution of printing.

Visual feedback in the 3D vocal tract model would be of the type illustrated in Figure 7; focusing on place and manner of articulation.

Hearing-impaired children, lacking adequate auditory feedback of a pronounced sound, often acquire a speech with low comprehensibility as many phonemes are reduced or replaced (Erber, 1983). Adults learning a second language often experience a related, but milder, problem. Lacking the ability to distinguish between phonemes in the new language and in the mother tongue, the correct phonemes are often replaced by their counterparts in the speaker's first language, resulting in an accent (Flege, 1998).

Both groups may be aided by a three-dimensional view of the correct pronunciation, especially if articulatory speech recognition can be applied to the user's own pronunciation. The user's phonemes can be analysed using sound-to-gesture inversion (Maeda, 1993-95), giving the corresponding vocal tract geometry. The correct and deviating pronunciations could then be contrasted automatically in the model, providing important feedback of how the pronunciation should be corrected.

At this stage the model only represents prototypes of Swedish phonemes extrapolated from two-dimensional drawings based on earlier X-ray measurements (e.g. Fant, 1964) and some volumetric data obtained from magnetic resonance imaging (MRI) (Baer et al, 1991; Narayanan et al, 1994). The latter have provided information on the overall size and 3D shape of the vocal tract. More measurements, especially volumetric, need to be made to fill the knowledge gap of vocal tract geometry outside the midsagittal plane.

Articulatory synthesis in 3D

Articulatory synthesis has two major advantages compared to other synthesis methods: generally it requires much fewer parameters to describe a speech sound and the description is intuitively attractive due to its close relation to human speech production.

The main objection, that articulatory synthesis requires a large number of computations, and it is therefore not possible to run dynamic synthesis in real time, is no longer as valid, thanks to the evolution of computer performance. Articulatory synthesizers allowing dynamic synthesis are becoming quite frequent (e.g. Browman et al, 1984, and Boersma, 1998).

One other disadvantage remains, however. Traditionally, articulatory synthesis has been based on two-dimensional models representing the vocal tract in the midsagittal plane, as exemplified by Coker et al. (1973), Rubin et al. (1981), Maeda (1988), and Boersma (1998). The area function in such a model is calculated using the formulae

$$\textit{cross-sectional area} = a \cdot (\textit{midsagittal width})^b$$

where the coefficients a and b have been determined empirically through X-ray measurements and casts of the vocal tract (e.g. Sundberg, 1969), or volumetric methods such as MRI (e.g. Baer et al. (1991) for vowel studies, Narayanan et al. (1995, 1997) and Alwan et al. (1997) for studies of fricatives and liquids, Yang & Kasuya (1994) for volumetric dependencies of sex and age, and Matsumura et al. (1994) for measurements of both vocal and nasal cavities.

The relationship is, however, not straightforward, as the coefficients vary greatly with the height over the larynx, as shown in the measurements by Ladefoged et al. (1971). The variations require that different coefficient values be used for the pharyngeal and the oral parts of the vocal tract, but even within each part the relation differs, depending on the distance from the pharynx. The pharyngeal data fit the power law hypothesis rather closely, yielding an average value of approximately 0.94 for a and 1.75 for b in the study by Baer et al. (1991).

Moreover, the determined coefficients vary from study to study. For the upper vocal tract, Sundberg et al. (1987) found $2.07<a<2.63$ and $1.33<b<1.47$, depending on the subject. Baer et al. (1991), on the other hand, argued that the midsagittal exponent b is approximately constant at $b=1.97$ and the coefficient a follows a regression line from 2.0 to 0.6, as the lips are approached. Consequently, the step from midsagittal distance to cross-sectional area depends

on a relationship, where the height over the larynx intervenes in a somewhat disputed way.

Determining the cross-sectional areas directly from a three-dimensional vocal tract model is one solution to this problem, since the area is then given directly from the model.

The co-ordinate system in which the vocal tract is represented defines anatomical distances in mm, and the areas can thus be calculated in four steps, without further scaling, as outlined below.

First, the centreline of the vocal tract is determined using 30 pairs of reference points on the tract wall surface in the midsagittal plane. For each reference pair the midpoint of the segment between the two reference vertices is assumed to lie on the centreline.

Next, 21 points equally spaced on the centreline are taken to form the centre of each plane cutting the vocal tract. Assuming a vocal tract length of about 17.5 cm, the planes will be 0.875 cm apart, except for the distance between the last but one and the last plane, which will depend on the vocal tract length. This length is determined by the larynx height and the lip protrusion. The centreline and the distribution of the cutting planes are shown in Figure 8.

The normal of each cutting plane is employed when searching for intersecting points on the vocal tract wall or on the tongue surface. The normal is assumed to coincide with the midline from the current cutting plane to the next. This assumption is correct when the planes are close enough. Intersections are detected by taking the dot product between the surface normal and a vector from the surface to every vertex point P in the neighbourhood[1]. The sign of the dot product indicates which side P is on, and if a polygon has two corners on opposite sides of the cutting plane it is intersected. The point of intersection is then calculated putting the parametric equation of the line between the two corners $P_1=(x_1,y_1,z_1)$ and $P_2=(x_2,y_2,z_2)$

$$\begin{cases} x = x_1 + (x_2 - x_1)t \\ y = y_1 + (y_2 - y_1)t \\ z = z_1 + (z_2 - z_1)t \end{cases} \quad \text{with } 0 < t < 1 \qquad (1)$$

[1] The neighbourhood is taken to be a cylinder with a radius large enough to contain the vocal tract locally and a height that is 0.875 cm in each direction along the midline as counted from the current cutting plane. The neighbourhood definition is necessary as faraway polygons otherwise may be classified as intersecting the current cutting plane.

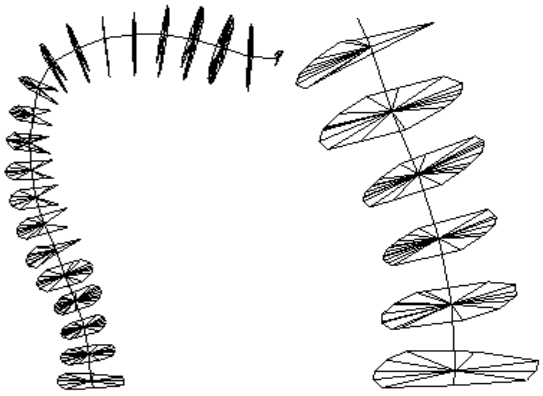


Figure 8. The 21 cross-sectional areas in the vocal tract are shown to the left, and a few cross-sections, showing the triangular structure of the parts are shown to the right.

into the equation of the cutting plane as given by its normal vector $\vec{n}$=(a,b,c) and a point $P_0=(x_0,y_0,z_0)$ belonging to the plane

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0 \qquad (2)$$

Solving for t yields

$$t = \frac{a(x_0 - x_1) + b(y_0 - y_1) + c(z_0 - z_1)}{a(x_2 - x_1) + b(y_2 - y_1) + c(z_2 - z_1)} \qquad (3)$$

The coordinates of the intersection are determined by putting equation (3) into equation (1).

The area of a polygon can be calculated as (Goldman, 1991)

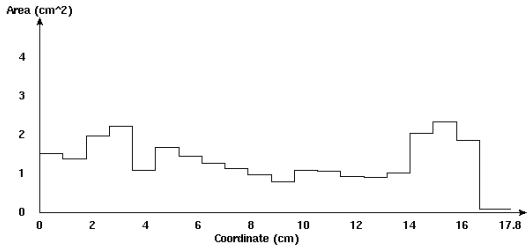
$$A = \tfrac{1}{2}\left|\frac{\vec{n}}{|\vec{n}|}\right| \bullet \left|\sum_{i=0}^{N-1} \vec{v}_i \times \vec{v}_{i+1}\right| \qquad (4)$$

where $\vec{n}$ is the normal of the cutting plane and v_i are the ordered intersection points. As each polygon has an even number of intersections (0, 2 or 4), counter-clockwise ordering of the intersection points can be done using chained lists to which new pairs of intersection points v_1 and v_2 are added under the constraint that

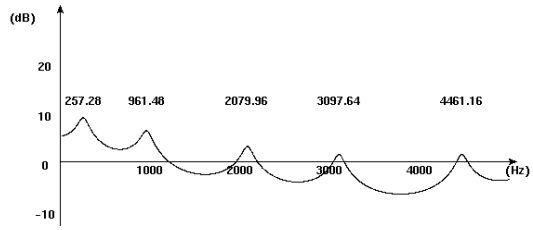
$$(\vec{v}_1 \times \vec{v}_2) \bullet \vec{n} > 0 \qquad (5)$$

Each pair of intersection points forms a triangle together with the point on the centerline. These triangles are added to the polygon model to visualise the cutting planes, as shown in Figure 8. The cross-sectional area is then calculated using equation (4).

The area function is then presented using a Tcl/Tk interface (Figure 9a), displaying the vocal tract as a number of sections with fixed length



a) The area function



b) The transfer function

Figure 9. Presentation of the cross-sectional area (a) and the transfer function (b).

0.875 cm (except for the last section that varies in length) and uniform area for each section. The total length of the vocal tract is shown by the last indicated coordinate on the x-axis. A similar display shows the transfer function, indicating the five first formants numerically (Figure 9b).

As the Tcl/Tk interface for presenting the transfer function is purely for handling the display, whereas all calculations are done in separate C++-functions (adopted from the formant calculation program formf by Liljencrants, 1975), the graphical interface is totally independent of the calculations, allowing this module to be exchanged at will. Incorporating the corresponding parts of the articulatory synthesizer FLEA/Tracttalk, described in Lin (1991), would allow for a more complete simulation, as parameters of wall impedance and other cavities like nasal branch and sinus piriforms are included. These effects are important in the development of a 3D articulatory synthesis method, close to the human original, and should hence be accounted for.

Discussion

For both applications outlined in the previous sections, it is clear that anatomical and articulatory correctness is crucial for the success of the model. This correctness can only be achieved by incorporating data from measurements of static as well as dynamic human vocal tracts. Such data is to a large extent lacking in the current state of the model, reducing it to a platform for future development rather than a complete three-dimensional model. Effort will thus have to be concentrated on using existing or new measurements done mainly with X-ray, MRI and ultrasound to obtain the correct geometrical shapes of the vocal tract.

Such evaluations have lately been done using e.g. Principal Component Analysis for images from both X-ray (Beautemps et al, 1996) and MRI (Badin et al, 1998) or even with self-organising artificial neural networks trained on an X-ray microbeam database (Blackburn & Young, 1996). The two methods allow an articulatory and a pseudo-articulatory model respectively to be driven directly from measurement data. These examples, together with e.g. the ASY at Haskins Laboratories (Rubin et al, 1981), show that articulatory measurements can be incorporated successfully in an articulatory model.

It will also be of great value to test the model consistently regarding both articulatory synthesis and recognition, to judge how well the relation between the 3D geometry and the speech signal can be modeled. Evaluation of generated transfer functions, formants and produced sound on the one hand and vocal tract geometry from articulatory recognition on the other should thus be carried out continuously as the model evolves. The latter is in itself an intriguing problem, but progress has been made in the quest for automatically determining vocal tract shape from acoustic data (Maeda, 1993-1995).

Acknowledgement

This work is carried out within CTT (the Centre for Speech Technology), jointly sponsored by KTH, NUTEK and Swedish industry.

References

Agelfors E, Beskow J, Dahlquist M, Granström B, Lundeberg M, Spens K-E & Öhman T (1998). Synthetic faces as a lipreading support, *Proc ICSLP'98,* 7: 3047-3050.

Álvarez A, Martínez R, Gómez P, Domínguez JL (1998). A signal processing technique for speech visualization, *Proc StiLL*, 33-36.

Alwan A, Narayanan S, Haker K (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics, *J Acoust Soc Amer*, 101: 1078-1089.

Badin P, Bailly G, Raybaudi M, Segebarth C (1998). A three-dimensional linear articulatory model based on MRI data, *Proc 3rd ESCA/COCOSDA Intl Workshop on Speech Synthesis*, 249-254.

Baer T, Gore JC, Gracco LW, Nye PW (1991). Analysis of vocal tract shape and dimensions using

magnetic resonance imaging: Vowels, *J Acoust Soc Amer*, 90: 799-828.

Beautemps D, Badin P, Bailly G, Galván A, Laboissière R (1996). Evaluation of an articulatory-acoustic model based on a reference subject, *Proc 1st ESCA Tutorial and Research Workshop on Speech Production Modeling – 4th Speech Production Seminar*, 45-48.

Beskow J (1995). Rule-based visual speech synthesis. *Proc Eurospeech '95*, 299-302.

Blackburn CS, Young SJ (1996). Pseudo-articulatory speech synthesis for recognition using automatic feature extraction from X-ray data, *Proc ICSLP'98*, 2: 969-972.

Boersma P (1998). Functional Phonology, LOT International Series 11, The Hague: Holland Academic Graphics. Pages i-ix, 1-493. *Doctoral thesis*, University of Amsterdam.

Browman P, Goldstein L, Kelso JAS, Rubin P, Saltzman E (1984). Articulatory synthesis from underlying dynamics, *J Acoust Soc Amer*, 75: S22-S23 (A).

Cohen M, Beskow J, Massaro D (1998). Recent development in facial animation: An inside view, *Proc AVSP '98*, 201-206.

Coker C, Fujimura O (1966). Model for the specification of the vocal tract area function, *J Acoust Soc Amer*, 40: 1271.

Engwall O (1998). A 3D vocal tract model for articulatory and visual speech synthesis, *Proc Fonetik 98, The Swedish Phonetics Conference*, 196-199.

Erber NP (1983). Speech perception and speech development in hearing impaired children. In: Hochberg Levitt, ed., *Speech of the Hearing Impaired*; 131-145.

Fant G (1964). Formants and cavities, *Proc Fifth Intl Congress of Phonetic Sciences*, Münster, 120-141.

Flege J (1998). Second-language learning: The role of subject and phonetic variables, *Proc StiLL*, 1-8.

Goldman R (1991). Area of planar polygons and volume of polyhedra. In: *Graphic Gems II*, 170-171, Academic Press.

Ladefoged P, Anthony JFK, Riley C (1971). Direct measurement of the vocal tract. *UCLA Working Papers in Phonetics* 19: 4-13.

Liljencrants J, Fant G (1975). Computer program for VT-resonance frequency calculations, *KTH STL-QPSR* 4: 15-20.

Lin Q (1991). Speech Production Theory and Articulatory Speech Synthesis", *Doctoral thesis*, KTH.

Maeda S (1988). Improved articulatory models, *J Acoust Soc Amer*, 84: S146.

Maeda S (ed.) (1993-95). WP2 - From speech signal to vocal tract geometry. In: *Speech Maps*, Year 1-3 Report, Vol III.

Mashie J (1995). The use of sensory aids for teaching speech to children who are deaf. *Profound Deafness and Speech Communication*, 461-491.

Matsumura M, Niikawa T, Shimizu K, Hashimoto Y, Morita T (1994). Measurement of 3D shapes of vocal tract, dental crown and nasal cavity using MRI: Vowels and fricatives, *Proc ICSLP'94*, S12-13.1-S12-13.4.

Mermelstein P (1973). Articulatory model of speech production. *J Acoust Soc Amer*, 53: 1070-1082.

Narayanan S, Alwan A, Haker K (1995). An articulatory study of fricative consonants using magnetic resonance imaging, *J Acoust Soc Amer*, 98: 1325-1347.

Narayanan S, Alwan A, Haker K (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals, *J Acoust Soc Amer*, 101: 1064-1077.

Nouza J, Mádlíková J (1998). Evaluation tests of visual feedback in speech and language learning", *Proc StiLL*, 151-154.

Perkell J (1974). A physiologically-oriented model of tongue activity in speech production", *PhD thesis*, MIT.

Rubin P, Baer T, Mermelstein P (1981). An articulatory synthesizer for perceptual research, *J Acoust Soc Amer*, 70: 321-329.

Stone M (1990). A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data, *J Acoust Soc Amer*, 87: 2207-2217.

Stone M (1991). Toward a model of three-dimensional tongue movement. *J Phonetics*, 19: 309-320.

Stone M, Lundberg A (1996). Three-dimensional tongue surface shapes of English consonants and vowels, *J Acoust Soc Amer*, 99: 3728-3737.

Sundberg J (1969). On the problem of obtaining area functions from lateral x-ray pictures of the vocal tract, *KTH STL-QPSR* 1: 43-45.

Sundberg J, Johansson C, Wilbrand H, Ytterbergh C (1987). From sagittal distance to area: A study of transverse, vocal tract cross-sectional area, *Phonetica*, 44: 76-90.

Wilhelms-Tricario R (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *J Acoust Soc Amer*, 97: 3085-3098.

Yang C-S, Kasuya H (1994). Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects, *Proc ICSLP'94*, S12-14.1-S12-14.4.

Öster A-M (1997). Auditory and visual feedback in spoken L2 teaching, *Phonum* 4: 145-161.