A 3D TONGUE MODEL BASED ON MRI DATA

Olov Engwall

olver@speech.kth.se
Centre for Speech Technology (CTT), KTH, Drottning Kristinas väg 31, SE-100 44 Stockholm, Sweden

ABSTRACT
A new three-dimensional tongue model has been developed within the KTH 3D vocal tract project using manually extracted tongue contours from MR Images of a reference subject producing 43 artificially sustained Swedish articulations. The six linear parameters jaw height, tongue body, tongue dorsum, tongue tip, tongue advance and tongue width were determined using an ordered linear factor analysis controlled by articulatory measures. 88% of the variation in the midsagittal plane and 78% of the overall sagittal variation was explained by the first five factors of the analysis. The six parameter model is able to reconstruct the modeled articulations in 3D with an overall RMS reconstruction error of 0.13 cm sagittally and 0.12 cm laterally, and it specifically handles lateral differences and the observed asymmetries in tongue shape.

1. INTRODUCTION
A number of articulatory-oriented two- ([1]-[3]) or three-dimensional ([4]-[7]) tongue models have been proposed over the years. The three-dimensional models have used Finite Element Modeling ([4]) to implement biomechanical constraints or statistical analysis of a measurement corpus collected with ultrasound ([5]-[6]) or Magnetic Resonance Imaging ([7]) to cover the articulatory space. This article presents the new statistically based tongue in the KTH 3D Vocal Tract model ([8]). It has been redeveloped using tongue contours extracted from MR Images both to generate an asymmetrical reference model and to determine articulatory parameter values and parameter influence on the vertices.

2. MR IMAGE PROCESSING
2.1 Image Acquisition
The three-dimensional MR Images used for developing the 3D tongue model are identical to the ones described in [9]. A total of 44 configurations were acquired: one neutral reference and 43 Swedish articulations. The neutral tongue position was defined as that of the subject at rest with upper and lower incisors touching and aligned. The corpus consisted of 13 vowels and 10 consonants in the three symmetric vowel contexts [a r o], artificially sustained during the 43 second acquisition time. For the consonants, the subject made the initial VC transition before the acquisition, then held the articulation while breathing out very slowly or holding his breath and finally made the CV transition after the scan.

The image set consists of three stacks of 18 parallel slices each (cf. Fig. 1): one axial stack of the pharynx, one oblique stack at 45° in the velar region and one coronal stack of the oral cavity. The final image resolution was 1 mm/pixel and the interslice distance 4 mm.

2.1 Contour extraction
The tongue contour (cf. Fig. 1) was manually extracted from each image using a Bézier curve editor (developed by P. Badin, ICP, CNRS, INPG). The whole tongue was extracted as one unit, without subdividing it into its constituent muscles. Moreover, the epiglottis and the external muscles (palatoglossus, styloglossus and stylohyoid) were excluded when extracting the tongue and only one contour was extracted from each image, i.e. when the tongue tip was separate from the tongue root in the coronal images, the contour included only the tip. The contour extraction process resulted in up to 42 planar contours (maximally 13, 18 and 11 contours from each set).

1.2 Three-dimensional Reconstruction
The ICP semipolar grid ([10], cf. Fig. 2a) was employed for the initial 3D reconstruction, but limiting the tongue by gridplane 8 (tongue root) and 27 (tongue tip). The tongue was hence reconstructed using 20 gridplanes, 10 of which were polar. The same reconstruction process as for the vocal tract shape model ([9]) was used, resulting in a three-dimensional tongue surface with substantial overlap between slices from the different stacks (cf. Fig. 2a).

Figure 1. Examples of tongue contours in /apa/: the axial set (top left), the oblique set (top right), the coronal set (bottom left) and the acquisition grid (bottom right).
In the subsequent transformation to the KTH 3D model, the overlap was removed by limiting the tongue contours in the axial and semi-polar parts of the grid by the first grid-plane in the alveolar part of the grid (plane 23 in Fig. 2a). The resulting contours were then resampled to have equally spaced points, such that the half-Posteriors in the axial and the semi-polar parts of the grid have 18 evenly spread points and those of the frontal part 30. A polygon mesh of 420 vertices and about 800 polygons was constructed by connecting each vertex (vi) to its neighbour in the same grid-plane (v(i+1)) and to the corresponding vertex (vi) and its neighbour on the previous grid-plane (v(i−1)) (cf. Fig. 2b).

The neutral tongue shape (subject at rest with closed jaw) was used as the reference shape for the polygon model as well as in the parameter extraction process. All other articulations were hence modelled as deformations from the neutral shape using the articulatory control parameters defined in the component analysis described below.

3. THE ARTICULATORY MODEL

3.1 The Linear Component Analysis

The extraction of the model’s parameters was done by decomposing the geometrical points of the tongue in linear components, through a Linear Component Analysis (LCA), where the factors to be extracted were imposed on the model using articulatory measures from the MR Images. LCA was chosen instead of PCA, at the cost of sub-optimal data variance explanation, as it suited the control parameter definitions in the KTH Vocal Tract model better.

The order in which factors were extracted, the region of influence and the direction of activation were chosen based on an earlier extraction of the midsagittal tongue control parameters for the same subject ([9]), employing guided PCA ([10]), i.e. pure PCA alternated with LCA. Five parameters JH, TB, TD, TT and TA were determined in the same order as for the midsagittal model and a sixth parameter, tongue width (TW), was added to account for width variations of the blade and tip.

3.2 Parameter Definitions

The parameters in the KTH visual speech synthesis system ([11]) are defined using the activation A (-1<Δ<1) of the movement of a prototype vertex (P) towards a target vertex (T) and a weight vector Wc, determining the influence of the parameter on every vertex of the mesh. The only deformation type used in the present tongue model is translations, and the displacement (Δx, Δy, Δz) of vertex i due to an activation A is hence (Δx, Δy, Δz) = A(Tc×P - Tc×P).

All tongue control parameters have been redefined and reduced in number compared to those in [8], as the asymmetric model and statistically defined weight vectors allow for lateral variations, such that tongue grooving and lateral asymmetries are handled automatically by the midsagittal control parameters.

The five parameters JH, TB, TD, TT and TA were defined by midsagittal translations, with midsagittal prototypes and targets, whereas TW was defined orthogonal to the midsagittal plane.

The prototype, target and activation were set according to articulatory measures and the weight vector of each parameter was extracted through the factor analysis, minimising the difference between the Cartesian vertex coordinates of the reference shape and those of the corpus in the least square sense.

When one parameter had been extracted, its contribution was withdrawn from all the articulations of the corpus and the next parameter was determined using the residual.

Jaw height – JH was defined as a linear translation such that its vertical component is equal to the articulatory measure JawHei, the jaw height, and its horizontal component is the value of JawAdv, the jaw advance, predicted from JawHei ([9]). The configuration with the maximal jaw height has a JH value of 1.0 in the vertical direction and JH of the other articulations are proportional to the JawHei quota.

JawHei was measured in the MR Images as the increase in the mean distance Δ, between the centres of gravity of the frontal air sinuses of the nose and the pulp of the lower incisors compared to the reference with closed jaw.

The weight vector wJH is laterally asymmetric, reflecting the fact that the subject consistently lowers the left tongue edge more than the right, as shown in the nomogram in Fig. 3a).

Tongue Body – TB controls the front-back movement of the tongue (corresponding to genioglossus activation), raising the tongue relative the palate and at the same time contracting in the pharynx. In the present model TB consists of two correlated translational deformations, using separate prototypes and targets for the oral and pharyngeal parts, as the parameters in the model are limited to uni-directional translations.

TB was determined as the normalised deviation from the reference shape of the measure TngBody. TngBody is the Euclidean distance from the grid centre to the midsagittal tongue contour along the median line of maximal deviations in the alveopalatal part. The range of TB is shown in Fig. 3b).

Tongue Dorsum – TD controls the velar arching of the tongue body (corresponding to activation of the styloglossus). The activation of TD for each articulation is given by the articulatory measure TngDors: the mean value of the Euclidean distance from the grid centre to seven midsagittal points in the velar region. TngDors, centred on the reference shape and normalised, determines TD, that also controls the grooving of the tongue, increasing the groove with decreasing TD, as shown in Fig. 3c).
Tongue Tip – TT and Tongue Advance – TA model the raising-lowering of the tongue tip and the advancing-retraction of the tongue tip and blade, respectively. The parameters are defined as the normalised values of the deviation from the neutral tongue shape, parallel to (TngTip<sub>rel</sub>) and orthogonal to (TngAdv<sub>rel</sub>) the gridlines. Fig. 3d)-e) show the nomograms.

Tongue Width – TW was added to the parameter set used for the midsagittal model, to control the width variation at the tongue blade and tip. The lateral widening of the tongue, TngWidth, is measured as the mean difference in width relative the reference shape at the tongue blade and tip edges. TW, the normalised value of TngWidth, is largest for palatal plosives and fricatives, as the tongue is widened when pressed against the palate, and smallest for velar plosives and fricatives, as most of the tongue volume is shifted backwards (cf. Fig. 3f).

3. MODELING RESULTS

4.1 Midsagittal Model

The variance explained by the control parameters in the midsagittal plane amounts to about 88%, with the contribution of each parameter shown in Fig. 4. Compared to the midsagittal vocal tract contour model in [9], where the parameters were extracted through guided PCA and calculated over the entire inner vocal tract contour, the contribution of JH (16.8% vs. 20.1%) and especially TD (23.1% vs. 45.6%) is lower and the parts explained by TB (30.9% vs. 20.5%), TT (12.1% vs. 4.1%) and TA (5.1% vs. 0%) are higher.

![Figure 4](image)

Figure 4. Standard deviation (in cm) of the midsagittal contour as a function of gridline number, when removing the contribution of JH, TB, TD, TT and TA successively.

4.2 Three-dimensional model

The data variance in the 3D model was calculated over the lingual region delimited by the two main axes of the semi-polar grid (12 and 23 in Fig. 2a), excluding the frontal lower part of the genio-hyglossus. The percentage of the total sagittal data variance explained by the five factors JH (13.5%), TB (27.6%), TD (14.8%), TT (16.0%) and TA (6.4%) amounts to over 78%.

The residual variation at the tongue root is largest in the midsagittal plane and decreases with the fibre distance from the midsagittal plane, whereas the residual at the tongue tip is smallest in the midsagittal plane and increases outwards.
The sagittal residual variance in the pharyngeal, velar and anterior oral part as a function of lateral fibre is shown in Fig. 5, where fibre numbers run from left to right and fibre number 10 corresponds to the midsagittal plane. The pharyngeal part is almost exclusively controlled by TB, the velar mainly by TB and TD and the tongue blade and tip mostly by JH, TT and TA.

The root mean square (RMS) reconstruction error relative the initial data was calculated to assess the model’s reconstruction abilities. The six parameter model is able to replicate the three-dimensional articulations with an overall RMS reconstruction error of 0.13 cm sagittally and 0.12 cm laterally. The maximal RMS error is 0.17 cm sagittally and 0.16 cm laterally and both lateral variations and the observed left-right asymmetries are handled.

5. DISCUSSION & PERSPECTIVES

A decrease in the pharyngeal part was found when reconstructing the vocal tract shape from the MR images in this study ([9]) and some changes in tongue position were found when comparing static MRI measurements to dynamical measures of simultaneous EPG and EMA ([12]). The tongue articulations of the MRI corpus may hence be somewhat shifted backwards, due to the supine position or the sustained articulation.

The presented three-dimensional tongue shape data nevertheless provide important information on 3D articulations, especially as the evaluated corpus is larger than in earlier studies of 3D tongue shapes using MRI (e.g. 25 articulations in [7]). More importantly, three-dimensional tongue shapes are shown to be replicable with fairly low reconstruction error using linear articulatory parameters defined in the midsagittal plane.

Future work on the tongue model will focus on modeling dynamical movements of the 3D tongue using real-time articulatory measurement methods (e.g. EPG and EMA).

6. ACKNOWLEDGEMENT

This work is supported by the Centre for Speech Technology (CTI) at KTH. The software for contour extraction and part of the three-dimensional reconstruction was provided by Pierre Badin, Institut de la Communication Parlée, UMR CNRS 5069, INPG, Université Stendahl, Grenoble, France. The image acquisition was done by Christoph Segerberth, INSERM U438, Centre Hospitalier Regional Universitaire, Grenoble, France with the assistance of Pierre Badin.

7. REFERENCES


Figure 5. Variance explained by each parameter in each part: pharyngeal (top left), velar (top right) and alveolar (bottom).