



ELSEVIER

Available at  
[www.ComputerScienceWeb.com](http://www.ComputerScienceWeb.com)  
POWERED BY SCIENCE @ DIRECT®

Speech Communication 41 (2003) 303–329

**SPEECH**  
COMMUNICATION

[www.elsevier.com/locate/specom](http://www.elsevier.com/locate/specom)

## Combining MRI, EMA and EPG measurements in a three-dimensional tongue model <sup>☆</sup>

Olov Engwall <sup>\*</sup>

*Centre for Speech Technology, KTH (Royal Institute of Technology), Drottning Kristinas v. 31, SE-100 44 Stockholm, Sweden*

Received 8 October 2001; received in revised form 14 March 2002; accepted 4 June 2002

### Abstract

A three-dimensional (3D) tongue model has been developed using MR images of a reference subject producing 44 artificially sustained Swedish articulations. Based on the difference in tongue shape between the articulations and a reference, 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. The first five factors explained 88% of the tongue data variance in the midsagittal plane and 78% in the 3D analysis. The six-parameter model is able to reconstruct the modelled articulations with an overall mean reconstruction error of 0.13 cm, and it specifically handles lateral differences and asymmetries in tongue shape. In order to correct articulations that were hyperarticulated due to the artificial sustaining in the magnetic resonance imaging (MRI) acquisition, the parameter values in the tongue model were readjusted based on a comparison of virtual and natural linguopalatal contact patterns, collected with electropalatography (EPG). Electromagnetic articulography (EMA) data was collected to control the kinematics of the tongue model for vowel-fricative sequences and an algorithm to handle surface contacts has been implemented, preventing the tongue from protruding through the palate and teeth.

© 2002 Elsevier B.V. All rights reserved.

### Résumé

Un modèle à trois dimensions de la langue a été élaboré à partir des images obtenues par Résonance Magnétique (IRM) sur un sujet prononçant 44 articulations suédoises. En s'appuyant sur la différence entre les contours de la langue mesurés pour les différentes articulations et une position de référence, six paramètres de contrôle de la mâchoire et la langue ont été déterminés par application d'une analyse factorielle sur des mesures articulatoires. Les cinq premiers facteurs ont expliqué 88% de la variance des contours de la langue dans le plan sagittal et 78% de la variance tridimensionnelle. Ce modèle à six paramètres est capable de reconstruire les articulations mesurées avec une erreur moyenne de 0,13 cm et peut également prendre en compte les différences latérales et les asymétries des contours de la langue. En vue de corriger l'hyper-articulation résultant des expositions prolongées durant l'acquisition d'IRM, les valeurs des paramètres ont été ajustées en comparant les contacts linguopalataux virtuels et ceux mesurés par électropalatographie. Des données de mouvement ont été mesurées pour des séquences voyelle-fricative à l'aide d'un

<sup>☆</sup> Supplementary data associated with this article can be found at [doi:10.1016/S0167-6393\(02\)00132-2](https://doi.org/10.1016/S0167-6393(02)00132-2).

<sup>\*</sup> Tel.: +46-8-790-8965; fax: +46-8-790-7854.

E-mail address: [olov@speech.kth.se](mailto:olov@speech.kth.se) (O. Engwall).

URL: <http://www.speech.kth.se/~olov>.

articulographe électromagnétique, afin de déterminer le contrôle cinématique du modèle. De plus, un algorithme de gestion des surfaces de contact a été développé, afin d'éviter que la langue ne traverse le palais et les dents.

© 2002 Elsevier B.V. All rights reserved.

*Keywords:* Three-dimensional modelling; Magnetic resonance imaging; Electromagnetic articulography; Electropalatography; Linear component analysis; Kinematic control; Parameter tuning; Boundary contact handling

---

## 1. Introduction

### 1.1. Tongue modeling

A number of articulatory models of the tongue have been proposed over the years, approaching the complexity of the tongue muscles from different viewpoints and making different simplifications to arrive at a working model, depending on the pertinent application. The main distinctions are between physiological and geometrical or statistical modelling, between two- (2D) and three-dimensional (3D) models, and if the model is real-time or not.

The physiological approach aims at understanding and modelling the muscular structure and functions of the tongue and the biomechanical constraints involved, such as volume conservation and tissue deformation.

An early attempt was made by Perkell (1974), who modelled the tongue in the midsagittal plane as a simplified structure of muscles, each represented as a line element of lumped springs and dampers. The articulation was changed by modifying the stiffness of an active spring, and volume conservation and boundary collisions were handled using mechanical forces.

The representation with mass-points and viscoelastic springs has been further used by Honda et al. (1994) and Dang and Honda (1998), with a physiological articulatory model that was quasi-3D, having three sagittal planes: the midsagittal and one plane on each side, displaced 2 cm laterally from the midsagittal. The tongue geometry was reconstructed from MR images of a male Japanese speaker and the dynamics of the tongue was improved using X-ray microbeam data for vowel and VCV sequences for 11 Japanese speakers (Dang and Honda, 2000).

An alternative to using springs to model the muscle tension is to divide the tongue into small units, often tetrahedrons or prisms, and to define the strain and the elastic properties for these units using finite element modelling (FEM).

This was first proposed by Kiritani et al. (1976) for a 3D static model of the tongue, where the 14 elements of the tongue were given linear isotropic elastic properties and the reactive strain within each element was related to the displacement of its vertices. The model was further developed to interpret control strategies in vowel articulations (Fujimura and Kakita, 1979; Fujimura, 1990). The FEM approach has also been investigated by Payan et al. (1995, 1997), who modelled the tongue in the midsagittal plane and simulated vowel transitions, controlled by an equation of motion that simulated the dynamics of the tongue.

The most ambitious use of FEM for tongue modelling this far has been carried out by Wilhelms-Tricarico (1997), who managed to define an exact mathematical method for simulations of the dynamic tongue movements and deformations, determined as the solutions to non-linear second-order differential equations that approximate the energy functions of the structures. The model was able to compute large tissue deformations under the volume conservation constraint exactly, by simulating incompressibility with a system for computing Lagrange multipliers. The muscle morphology of the 3D model was further refined by mapping data from 3D magnetic resonance imaging (MRI), the Visible Human project and anatomical literature onto the model (Wilhelms-Tricarico and Wu, 1997; Wilhelms-Tricarico, 2000).

Physiological modelling is however very computationally intensive (50 times real-time for the model proposed by Dang and Honda (1998) and

substantially more for FEM models), and if a real-time constraint is imposed, simplifications are called for. The argument behind these simplifications is that what really matters for the produced speech is the articulator shape, not how this shape was generated, and the focus is shifted from muscular modelling to representations of the outline of the tongue.

In geometrical modelling, assumptions are made on the geometry of the articulators and how they move, based on observations of speech production, but without a direct coupling between measurements and the model. This approach is exemplified by the midsagittal model in (Mermelstein, 1973), further developed in (Rubin et al., 1981, 1996), and the functional 2D model by Coker and Fujimura (1966). The tongue is approximated as a geometrical shape, e.g. a circle segment in the midsagittal plane, that is deformed using rotations around fixed axes and translations in predefined directions. The idea in this type of modelling is to produce a model with fewer details that is still able to produce all observed articulations with simply defined parameters.

Instead of observing and adjusting the model to speech production data, it is possible to let the measurements define the model statistically, as done by Lindblom and Sundberg (1971), who used a purely statistical decomposition of tongue contour variations in X-ray data into articulatory parameters. The two factors place and degree of maximum distortion were defined, where [i, u, a] represented the maximal distortions in the palatal, velar and pharyngeal regions, respectively.

An effective way to do the statistical analysis is to sample the tongue contour with a grid, represent the sampled tongue shape as a vector, with the distance from the inner part of the grid as a function of the gridline number, and then find components that sum up to the observed vectors. Harshman et al. (1977) introduced the PARAFAC analysis, where the tongue positions of 10 English vowels produced by 5 speakers were quantified in terms of 13 superimposed lines, and then analysed into two factors: firstly, the forward movement of the root accompanied by an upward movement of the front of the tongue and secondly, the upward

and backward movement of the tongue. The two factors gave large correlations (over 0.96) between the observed data and the model predictions.

A similar method to extract optimal factors explaining the tongue contour is through the principal component analysis (PCA). However, PARAFAC and PCA have a weakness for articulatory modelling, in that they do not guarantee that the extracted components represent elementary articulatory gestures.

Maeda (1988, 1990) hence instead proposed the arbitrary factor analysis, in which the tongue contour was decomposed using PCA, but only after the effect of the jaw position had been removed using linear regression. The tongue shape could then be described using four articulatory parameters interpreted as jaw position, front-back tongue body position, arching-flattening of the dorsal shape and raising-lowering of the tongue blade.

The statistical analysis in the tradition of Maeda has been used by Badin et al. (2000) to construct a 3D tongue model based on planar contours from MRI data. The position of each contour point was controlled by six articulatory parameters, defined through a factor analysis of the initial data set.

This article presents another 3D tongue model defined from MR images using a similar approach. In the longer run, the KTH 3D Vocal Tract project (Engwall, 1999) aims at generating a 3D vocal tract model that can be used for multimodal synthesis, producing both articulator animation and acoustics from the same parameter set. For the time being it is however the short-term goal of using the model in a text-to-audiovisual synthesis system that is in focus, to introduce a more realistic tongue model in the synthetic faces than the simple one defined in (Beskow, 1995). The modelling approach in this study should be seen in this perspective, regarding implications on real-time constraints and definitions of parameters.

These constraints are the main reason for using a linear model rather than FEM, as the model has to be fast and simple enough to be incorporated in a real-time system. In this respect the KTH tongue is a follower to the b-spline model by Cohen et al. (1998), where 9 sagittal and  $3 \times 7$  coronal

parameters were used to replicate natural tongue shapes observed with ultrasound and MRI.

### 1.2. Tongue measurements

MRI data was chosen as the basis for the modelling as it is the 3D measurement method that produces the most detailed tongue images without any known harmful effects on the subject. The disadvantages of MRI when measuring speech, such as supine position, artificially sustaining and high amplitude noise, were considered acceptable. The main alternative, ultrasound, used for 3D modelling by Stone (1990) and Stone and Lundberg (1996), has some benefits over MRI (mainly shorter acquisition time and upright position), but is unable to image the tongue tip and gives less detailed tongue surface data. A more thorough discussion on the choice of measurement methods is presented in (Engwall, 2002b).

In a comparative study (Engwall, 2000a) of coarticulation measured by static MRI and combined real-time electromagnetic articulography (EMA) and electropalatography (EPG), it was found that the artificially sustained articulations in the MRI acquisition were hyperarticulated. The conclusion of the study was that the static MRI data needed to be complemented with real-time data, in order to generate a model representative of running speech.

Following this conclusion, the parameters of the 3D tongue model have been empirically adjusted, using data on the natural linguopalatal contact, collected with EPG. The concept of using 3D models of the tongue and palate to determine virtual linguopalatal contact patterns that can be compared to natural EPG data has been proposed earlier by Schwartz and Boë (2000) and Cohen et al. (1998), but no results from EPG adjustment of 3D tongue models have yet been presented.

The static MRI data also needs to be complemented with real-time parameter control, in order to generate a kinematic model showing articulatory movements. This has been studied for sequences of fricatives and vowels using EMA.

The generation of the model from MR images, the tuning and the kinematic parameter control are described in the following sections.

## 2. Measurements

All measurements in this study used one 27–28 year-old reference subject: a male native speaker of standard Swedish with no dental fillings that could distort the MR images, and no record of speech disorders.

### 2.1. Magnetic resonance imaging acquisition

In the development of the 3D tongue model a set of 3D MR images was used, with a total of 44 configurations: one reference and 43 Swedish articulations. The reference tongue position was defined as that when the tongue rested on the floor of the mouth with the tongue tip touching the lower incisor and with upper and lower incisors touching and horizontally adjusted to be in line. This reference was used as it was a well-defined position that could be reproduced in the upright EMA–EPG measurements. The acquisition was made according to the same protocol and set-up as in (Badin et al., 1998), with details on the acquisition for this study given in (Engwall and Badin, 1999).

The corpus consisted of the 13 vowels [y:] (as in *byt*—change), [i:] (vit—white), [ɥ:] (hus—house), [ø] (hund—dog), [e:] (vet—know), [ɑ:] (mat—food), [a] (matt—feable), [u:] (bo—live), [o:] (gå—walk), [ɔ] (gått—walked), [æ:] (här—here), [ø:] (hö—hay), [œ:] (hör—hear) and the 10 consonants [p, t, k, l, r, s, f, ʃ, ʂ, ʃ] in three symmetric contexts with the short vowels [a, ɪ, ʊ]. In the subject's mid-Swedish dialect of the Stockholm area, [ʂ] represents the voiceless retroflex fricative, such as in *fors* (rapids), [ç] the voiceless alveolo-palatal fricative, in e.g. *tjur* (bull) and [ʃ] the voiceless velar fricative (simultaneous [ʃ] and [x]) in e.g. *sju* (seven). The subject's articulation of these fricatives is illustrated with midsagittal tracings in (Engwall and Badin, 2000).

The VCV pseudo-words were produced with a long consonant, surrounded by lax vowels, e.g. 'appa' [ap:a], with the stress on the consonant. The VCV pseudo-words do occur, with a few exceptions, as parts of real Swedish words, and the subject moreover practiced on all the VCV words beforehand, to ensure that the vowel context specification was followed.

All articulations were artificially sustained during the 43 s acquisition time. For the consonants, the subject made the initial VC transition before the acquisition, then held the articulation while breathing out very slowly (for fricatives) or holding his breath (for stops) and finally made the CV transition after the scan.

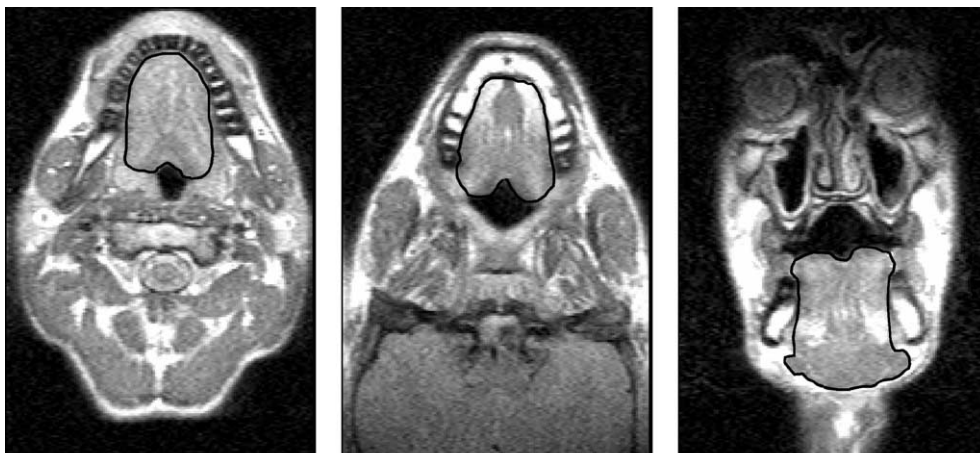
The 43 articulations included all the long Swedish vowels, the short vowels that were judged to differ substantially from the corresponding long vowel (e.g. [a] compared to [ɑ:]) and all the voiceless consonants, assuming that the voiced consonants could be modelled on the voiceless counterparts. The number of articulations resulted from the wish to collect a corpus that was as large as possible within the time allowed at the MRI scanner and taking subject fatigue into account. The retroflexes [ʈ, ʂ, ʐ] were not collected in the 3D set, as they could not be analysed in the grid used (as it assumes that the tip is the front-most part of the tongue, which is not the case for these retroflexes).

The 3D set used consists of 54 images in three subsets of 18 parallel MRI slices each: one axial

subset of the pharynx, one oblique subset at 45° in the velar region and one coronal subset of the oral cavity. The final image resolution was 1 mm/pixel with a slice thickness of 3.6 mm and an interslice centre distance of 4 mm.

The tongue contour was manually extracted from each image using interpolation with Bézier curves (Badin et al., 2000). The whole tongue was extracted as one unit, without subdividing it into its constituent muscles. Moreover, the following choices were made in the contour extraction:

- (1) The epiglottis was considered as a separate part of the vocal tract and it was hence not included in the tongue contour when it was present in the axial images.
- (2) The external muscles connected to the tongue body, the palatoglossus and styloglossus, were excluded when extracting the tongue contour from the images, as the purpose was to provide a model of 3D tongue movements, not to model the biomechanics of the tongue. The jaw muscles and the lateral tissue at the base in the coronal images (cf. Fig. 1c) were excluded in



(a) The axial pharyngeal set (slice 16). View angle 0° with respect to the vertical axis.

(b) The oblique velar set (slice 10). View angle 45° with respect to the vertical axis.

(c) The coronal set (slice 2). View angle 90° with respect to the vertical axis.

Fig. 1. Examples of tongue contours in images of [ʌpʌ]. Air passages and teeth appear in black, fat and marrow, containing much free hydrogen, appear in white.

the tongue model for the same reasons, but as it was possible to remove this part at a later stage, it was included in the tongue contours, as it might be wanted in future modelling.

- (3) Only one contour was extracted from each image, meaning that when the tongue tip was separate from the tongue body 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 the respective sets), exemplified by Fig. 1, which shows tongue contours from each of the different stacks.

## 2.2. Electropalatography data acquisition

The corpus consisted of the subset of the MRI corpus of the 25 articulations with clear linguopalatal contact. The vowels [e:, i:, y:, ʌ:] were acquired in isolation, whereas the consonants /s, ʃ, ʒ, t, k, l/ were collected in VCV context with  $V=[a, ɪ, ʊ]$ . The fricatives were acquired simultaneously with EMA data, as a part of a fricative study, described in (Engwall, 2000b).

The linguopalatal contact data was collected using a Reading 62 electrode EPG system (Jones, 1977).

## 2.3. Electromagnetic articulography data acquisition

The corpus consisted of the fricatives [s, ʃ, ʒ, ʎ] in symmetric VCV context with  $V=[a, ɪ, ʊ]$ . The EMA measurements were collected with the Move-track (Branderud, 1985) electromagnetic measurement system developed at the Department of Linguistics, University of Stockholm. Five receiver coils were used in the study, all placed in the midsagittal plane. The first, placed on the upper incisor, served as reference to adjust for head movements. The remaining four were placed on the lower incisor and on three points on the tongue, 11, 36 and 55 mm from the tip of the protruded tongue (cf. Wrench and Hardcastle, 2000, for a general illustration of EMA coil placement). The front-most coil was placed 11 mm from the tip, as positioning it closer to the tip interfered

with the production of the fricatives. It hence measured the movement of the blade rather than the tip, but none of the fricatives had such retroflexion that the correlation between the tongue tip and the anterior coil was greatly diminished. [ʃ] is a retroflex, but the subject still produced it without any important backward arching of the tongue tip (cf. Engwall and Badin, 2000).

## 3. The articulatory model

### 3.1. Three-dimensional reconstruction

The semi-polar grid defined in (Beautemps et al., 2001) was employed for the initial 3D reconstruction (cf. Fig. 2b), with the modification that the tongue was reconstructed using the 20 gridlines from the tongue root to the tongue tip, instead of the 28 that cover the entire vocal tract (gridline 1 in this study hence corresponds to gridline 8 in the original semi-polar grid).

The reconstruction was based on the fact that all contours had the same number of points, grouping points with the same index on different contours into 3D fibres (Badin et al., 1998) running from the tongue root to the tip. The intersection of these fibres with the planes that are orthogonal to the midsagittal plane and associated with the semi-polar grid defined the contour in each gridplane (cf. Fig. 2a).

When these contours were connected, a 3D tongue surface with substantial overlap between slices from different stacks was created, as shown in Fig. 2.

In the subsequent transformation of the 3D tongue surface to the KTH 3D model, the overlap was removed by limiting the tongue contours in the axial (lines 1–5) and semi-polar parts (lines 6–15) of the grid to the parts that did not surpass the first grid-plane in the second linear part of the grid (gridplane 16 in Fig. 2).

The trimmed contours were then resampled to have equally spaced points along the contour, such that the half-contours in the axial and the semi-polar parts of the grid each have 18 evenly spread points and those of the frontal part 30. The contours in plane 16–20 were divided by the last

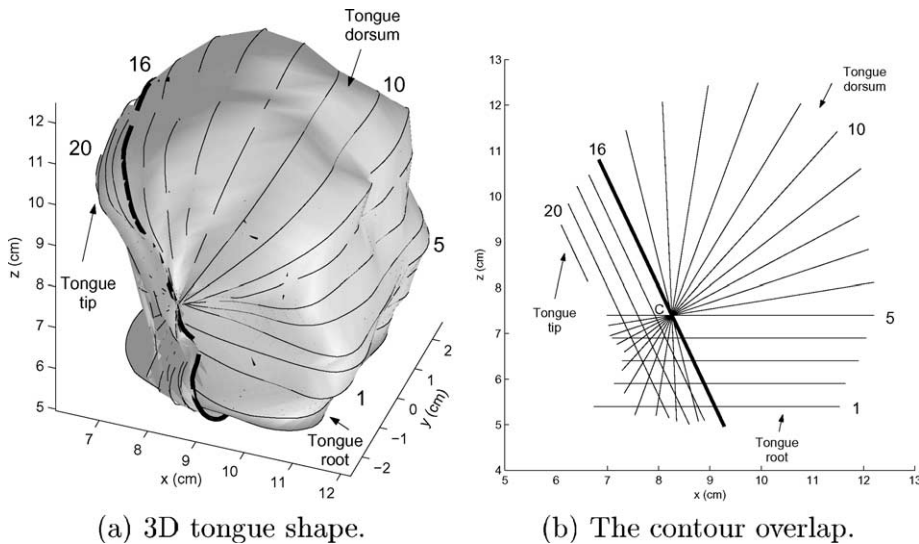


Fig. 2. Initial 3D tongue shape reconstructions of [ʰf], with gridline numbers indicated.

(upper-most) axial gridplane (gridplane 5 in Fig. 2b) before the resampling, so that they were resampled with 18 points above that plane and 10 below (5 on each side).

This allowed for a polygon mesh construction of the tongue by connecting each vertex ( $v_i$ ) to its neighbour in the same gridplane ( $v_{i+1}$ ) and to the corresponding vertex ( $v_j$ ) and its neighbour ( $v_{j+1}$ ) on the adjacent gridplane. In the junction between gridplane 16 and the axial and semi-polar parts of the grid, the 18 vertices of contour 16 that were above gridline 5 were connected to the 18 vertices of contour 15, as outlined above for the other gridplanes, whereas the 10 that were below were connected to the ends of the 5 axial contours no. 1–5.

This resulted in an ordered mesh consisting of 420 vertices and approximately 800 polygons. In this mesh the sagittal coordinates refer to the coordinate from the inner part of the grid to the outside of the tongue. The lateral coordinates run from left to right.

The tongue shape when the subject was at rest with closed jaw was used as the reference shape for the polygon model as well as in the parameter extraction process. This means that tongue shapes for all other articulations were created in the model as deformations from the reference shape

using the articulatory control parameters defined in the component analysis described below.

In the last part of the reconstruction process, the sagittal fibres were binomially smoothed to suppress some local variations. This smoothing was mainly for visual purposes, reducing tongue shape variations due to reconstruction artefacts, and had only minor influence (4%) on the model's ability to explain the data variability (cf. Section 3.4).

### 3.2. The linear component analysis

The extraction of the model's parameters was done by decomposing the geometrical points describing the tongue in linear components. In the present study this was done through linear component analysis (LCA), where the factors to be extracted were imposed on the model.

The advantage of using LCA is that every extracted control parameter has a well-defined articulatory influence on the model and that articulatory measures, such as the jaw height can be used in the extraction process. The disadvantage is that the data variation is not as efficiently explained as with PCA or PARAFAC. LCA was chosen nevertheless, due to its compatibility with the definition of

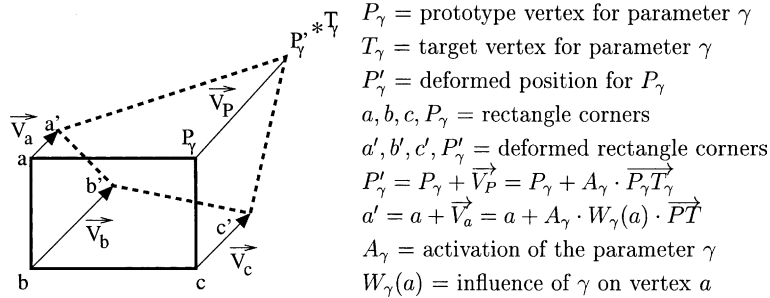


Fig. 3. The definition of translational deformations in the model exemplified for a rectangle. The solid rectangle contour  $abcP_\gamma$  is deformed into the dashed quadrangle  $a'b'c'P'_\gamma$ , due to differences in the influence of the parameter on the four corners. In this example  $A_\gamma = 0.8$ ,  $W_\gamma(a) = 0.25$ ,  $W_\gamma(b) = 0.75$  and  $W_\gamma(c) = 0.5$ .

control parameters in the existing KTH visual synthesis system, explained below.

The system by Beskow (1995) uses deformable wireframe meshes, that can be controlled by a set of parameters, each parameter  $\gamma$  using the activation  $A_\gamma$  ( $-1 \leq A_\gamma \leq +1$ ) of the movement of a prototype vertex ( $P_\gamma$ ) towards a target vertex ( $T_\gamma$ ) along an activation line and a weight vector  $W_\gamma(i)$ , determining the influence of the parameter on every vertex  $i$  of the mesh, as illustrated in Fig. 3. The displacement  $(\Delta x, \Delta y, \Delta z)$  of vertex  $j$  caused by setting the activation of a control parameter to  $A_\gamma$  is hence

$$(\Delta x_j, \Delta y_j, \Delta z_j) = A_\gamma \cdot W_\gamma(j) \cdot (T_{\gamma x} - P_{\gamma x}, T_{\gamma y} - P_{\gamma y}, T_{\gamma z} - P_{\gamma z}) \quad (1)$$

for translations, which is the only deformation type used in the present model.

For the LCA, this means that the parameters have to be defined by choosing a prototype  $P_\gamma$  and a target  $T_\gamma$  based on articulatory measures and then letting the weight function  $W_\gamma(i)$  be set by the statistical analysis of the tongue shapes in the corpus. The order in which factors were extracted, the region of influence and the direction of activation were however not chosen arbitrarily, but based on an earlier study of the parameters controlling the midsagittal tongue contour for the same reference subject (Engwall and Badin, 1999). That study employed guided PCA, described in Beautemps et al. (2001), consisting in alternating pure PCA and LCA, to extract, in order, the five parameters jaw height JH (explaining 20.1% of the

variance), tongue body TB (20.5%), tongue dorsum TD (45.7%), tongue tip TT (4.1%) and tongue advance TA (negligible).

In this study, the five parameters JH, TB, TD, TT and TA were determined in the same order and one sixth parameter, tongue width (TW), was added to account for variations of the width of the tongue in the oral cavity.

### 3.3. Parameter definitions

The tongue control parameters described here are similar to those in (Engwall, 1999), but all have been redefined, changing prototypes, targets and weights. Two improvements of the model were made possible in redefining the parameters based on 3D MRI data:

Firstly, using an asymmetric model (as opposed to the symmetric model proposed in (Engwall, 1999), where the right part was a reflection of the left) and a statistically defined weight vector allows for lateral variations of the control parameter influence  $W_\gamma(i)$ , such that tongue grooving and lateral asymmetries are handled automatically by midsagittal control parameters. The parameters defined in (Engwall, 1999) to control grooving and dorsum arching are hence not needed, as these aspects are simply consequences of TB, TD or TT activation, a finding replicating that by Badin et al. (2000).

Secondly, the parameter activations are defined from the corpus, meaning that the articulatory range of each parameter is constrained by the training data in the corpus (no limitation has

however been imposed on the combination of different control parameters, meaning that physically impossible articulations could be created when combining extreme activation of different parameters. For the articulations of the corpus this is no problem, as the combination of parameter activation is fixed in the component analysis. The corpus also contains a large part of the Swedish phonemes, and specifically the border phonemes of the articulatory space, so that parameter combination for articulations that were left out can be determined through interpolation of parameter combinations in the corpus).

The five parameters JH, TB, TD, TT and TA were defined by midsagittal translations; i.e. both prototype  $P_\gamma$  and target  $T_\gamma$  are in the midsagittal plane and the deformation is parallel to this plane. The parameter TW was defined according to the same principle, but its activation is 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. The weights were determined by minimizing the difference between the Cartesian vertex coordinates of the reference shape and those of the corpus in the least square sense, i.e. using Eq. (1) to determine  $W_\gamma(i)$  as

$$W_\gamma(i) = (A_{\gamma \text{ art}} \cdot \Delta_{P_\gamma T_\gamma}) \setminus \Delta_{\text{art-ref}} \quad (2)$$

where  $W_\gamma(i)$  is the weight array,  $A_{\gamma \text{ art}}$  is the activation array for all articulations,  $\Delta_{P_\gamma T_\gamma}$  is the Euclidean difference between the prototype and target,  $\Delta_{\text{art-ref}}$  is a vector with the Euclidean differences between the reference vertices and the vertices of the articulations in the corpus and  $\setminus$  stands for the matrix division giving the least square solution to an overdetermined system of equations. 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.

The influence of each of the six parameters defined below is shown in the animated image files.<sup>1</sup>

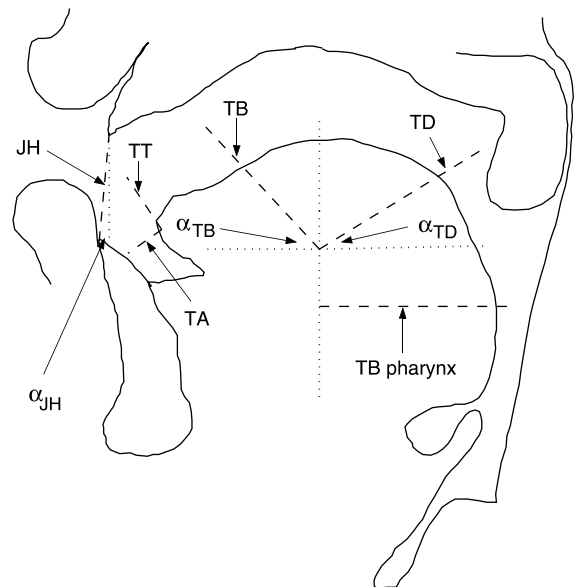


Fig. 4. The articulatory measures used for the parameter definitions. The dashed lines show the direction of activation for each parameter (coinciding with the median lines of measured maximal deviation). The dotted lines are horizontal–vertical references and  $(\alpha_{\text{JH}}, \alpha_{\text{TB}}, \alpha_{\text{TD}})$  refer to the angles mentioned in the text.

### 3.3.1. Jaw opening—JH

The jaw opening parameter was defined as a linear deformation along the median line of the jaw opening over all articulations for the subject (cf. JH in Fig. 4), such that its vertical component is the measured vertical jaw height, JawHei, and its horizontal component is the horizontal displacement predicted from JawHei, using the median line. The jaw opening parameter, JH, was then defined for each articulation as

$$\text{JH}(\text{art}) = \frac{\text{JawHei}(\text{art})}{\cos(\alpha_{\text{JH}}) \max(\text{JawHei})} \quad (3)$$

which means that the configuration with the maximal jaw height, [æ:], will have an activation of JH that is 1.0 in the vertical direction and that the activation for the other articulations will be proportional to the quota above.

$\alpha_{\text{JH}}$  is the angle from the vertical to the activation line of JH (cf. Fig. 4), and it is positive, reflecting the tendency of the subject to advance the jaw slightly when lowering the jaw.

<sup>1</sup> The animated image files are available in the online version of this paper.

The vertical and horizontal displacements were determined from the MR images. For each articulation, JawHei was measured by first determining the mean distance  $\Delta_{si}$  between the centres of gravity of the frontal air sinuses of the nose and the pulp of the lower incisors. The mean was taken over a set of three images using both left and right sinuses, giving a data set of six jaw heights for every articulation. A set of three reference measurements of the subject with closed jaw provided the reference for JawHei = 0, i.e. JawHei for an articulation is  $\Delta_{si}(\text{art})$  subtracted by  $\Delta_{si}(\text{ref})$  for a closed jaw:

$$\text{JawHei}(\text{art}) = \Delta_{si}(\text{art}) - \Delta_{si}(\text{ref}) \quad (4)$$

The advancing of the jaw, measured as the horizontal displacement of the jaw in the midsagittal images, is very much smaller than the vertical displacement, and  $\alpha_{JH}$  is hence only +7.0°. The target for JH is given by the maximally open articulation, [œ:], as  $T_{JH} = [x_{JH} - \text{JawHei}([\text{œ:}]) \tan(\alpha_{JH}), 0, z_{JH} - \text{JawHei}([\text{œ:}])]$  with the prototype at  $P_{JH} = [x_{JH}, 0, z_{JH}]$  (note that the  $x$ -coordinate grows following the tongue backwards).

### 3.3.2. Tongue body—TB

The tongue body parameter TB controls the front–back movement of the tongue, evidenced e.g. in the study of biomechanical degrees of freedom in tongue movements by Perrier et al. (2000).

However, TB influences the tongue contour in two different directions: raising or lowering the oral part of the tongue towards or from the hard palate, and, at the same time contracting or expanding it in the pharynx. Due to the definition of the parameters in the present model (where the deformation is along one axis, cf. Section 3.2), TB is defined with separate prototypes and targets in the oral and pharyngeal parts, with the division between the two parts given by the pivot point of the control parameter TB in the midsagittal model. In both the oral and the pharyngeal cavity TB was determined using articulatory measures with respect to the centre of the reconstruction grid (C in Fig. 4).

For each articulation, the maximal Euclidean distance in the midsagittal plane from the grid

centre to the tongue contour was determined in the alveo-palatal region, and the direction of the TB activation in the oral part was given by the median line of these maximal distances (TB in Fig. 4). The articulatory measure TngBody was then calculated as the deviation from the reference shape along this median line, and TB is this measure normalized by the maximal deviation, occurring for the palatal plosive [a<sup>k</sup>]:

$$\text{TB}(\text{art}) = \frac{\text{TngBody}(\text{art}) - \text{TngBody}(\text{ref})}{\max(\text{TngBody}) - \text{TngBody}(\text{ref})} \quad (5)$$

The maximal deviation  $\Delta_{\text{TngBody}}$  also determined the target, such that the prototype vertex on the reference shape ( $P_{\text{TB}} = [x_{\text{TB}}, 0, z_{\text{TB}}]$ ) reaches the target position  $T_{\text{TB}} = [x_{\text{TB}} + \Delta_{\text{TngBody}} \cos(\alpha_{\text{TB}}), 0, z_{\text{TB}} + \Delta_{\text{TngBody}} \sin(\alpha_{\text{TB}})]$ , for the activation  $A_{\text{TB}} = 1.0$ .

The pharyngeal part of TB is determined by calculating the maximal horizontal deviation from the centre of the grid (along the line TB pharynx in Fig. 4), then centring on the reference shape (i.e.  $\text{TB}_{\text{pharynx}} = 0$  for the reference shape) and normalising:

$$\begin{aligned} \text{TB}_{\text{pharynx}}(\text{art}) \\ = \frac{\text{TngBody}_{\text{pharynx}}(\text{art}) - \text{TngBody}_{\text{pharynx}}(\text{ref})}{\max(\text{TngBody}_{\text{pharynx}}) - \text{TngBody}_{\text{pharynx}}(\text{ref})} \end{aligned} \quad (6)$$

### 3.3.3. Tongue dorsum—TD

TD controls the velar arching of the tongue body, and as for TB, TD is defined in accordance with the midsagittal control parameter. The activation of TD for each articulation is given by an articulatory measure relative the centre of the grid. The articulatory measure TngDors is the mean value of the Euclidean distance from the grid centre C to seven midsagittal points in the velar region for each articulation. The mean value, rather than the maximum, of the distances is taken as the articulatory measure, as this minimizes the variation over a larger part of the velar region, whereas the maximum minimizes the variation more only in the absolute neighbourhood of the maximum deformation. TngDors, centred on the

reference shape and normalized to 1.0, gives the control parameter TD:

$$TD(\text{art}) = \frac{TngDors(\text{art}) - TngDors(\text{ref})}{\max(TngDors) - TngDors(\text{ref})} \quad (7)$$

The prototype  $P_{TD} = [x_{TD}, 0, z_{TD}]$  is the midsagittal velar point with maximal deviation, that reaches the target  $T_{TD} = [x_{TD} + \Delta_{TngDors} \cos(\alpha_{TD}), 0, z_{TD} + \Delta_{TngDors} \sin(\alpha_{TD})]$  for maximal activation of TD (along the axis TD in Fig. 4). The velar plosive  $[ʋ_k^ʋ]$  has the maximal TD activation  $TD = 1.0$ , whereas  $[ʋ_ç^ʋ]$  represents the minimum,  $TD = -0.87$ . The fact that the minimum occurs for  $[ʋ_ç^ʋ]$  is due to the important tongue grooving for this articulation, causing its midsagittal contour to have the least velar expansion of all articulations (cf. Fig. 8d). This grooving is caused by the fact that the tongue body has to be raised and moved forwards substantially to create the alveo-palatal fricative with the jaw lowered in  $[ʋ_ç^ʋ]$  and the tongue dorsum is hence lowered even in the back vowel context. The grooving is negatively correlated to TD, so that the groove decreases with positive TD and increases with negative TD.

### 3.3.4. Tongue tip—TT, tongue advance—TA

TT and TA model the raising–lowering and advancing–retraction of the tongue tip and blade, respectively. The deviation of the tongue tip for an articulation compared to the reference is measured as  $TngTip_{rel}$ , the part parallel to the gridline (TT in Fig. 4), and  $TngAdv_{rel}$ , the part orthogonal to the gridline (TA in Fig. 4). Note that  $TngTip_{rel}$  and  $TngAdv_{rel}$  are defined relative the reference tongue shape and not relative the centre of the grid; both measures can hence be negative as well as positive. TT is the normalized value of  $TngTip_{rel}$  and TA is the normalized value of  $TngAdv_{rel}$ . The maximal TT activation is for the lateral  $[ʃ^a]$ , where the tongue tip has to be raised more relative the tongue body, as the surrounding vowel is open, and the minimal is for  $[ʋ_k^ʋ]$ , where the tongue tip is actively lowered to increase the front cavity prescribed by the vowel  $[ʋ]$ . TT is also the main contributing parameter to grooving in the front part of the tongue. For TA, the activation range is from  $-1.0$  to  $0.18$  and the majority of the articu-

lations have negative values of TA, which is normal with the definition given above, as the tongue tip is quite advanced (touching the lower incisor) in the reference position. Apart from controlling the active advancing and retraction of the tongue tip, TA also contributes to the volume conservation at the tongue blade.

### 3.3.5. Tongue width—TW

One parameter, TW, was added to the control parameter set used for the midsagittal contour, to model the width variation at the tongue blade and tip, that are caused by volume conservation and deformation of the tongue when it is in contact with the palate. The addition of TW is necessary as the midsagittal parameters, that could be linked statistically to the lateral width of the tongue, can only inflict deformations parallel to the midsagittal plane (cf. Section 3.2), and because the model has no physiological modelling of the linguopalatal contact that can describe width variations due to pressure against the palate.

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 (i.e. the difference in  $y$ -coordinate for the tongue edge vertices at the gridplanes 14–20). TW is the value of  $TngWidth$  normalized by  $\Delta_{TW} = \max(TngWidth)$ , and influences vertices from gridplane 8 and onwards, using a horizontal translation defined by the target  $T_{TW} = [x_{TW}, y_{TW} - \Delta_{TW}, z_{TW}]$  (the prototype being  $P_{TW} = [x_{TW}, y_{TW}, z_{TW}]$ ). TW is generally largest for palatal plosives and fricatives, the width of the tongue increasing as it is pressed against the hard palate, and minimal for velar plosives and fricatives, having a large part of the tongue body volume retracted towards the velum and hence creating a slender tongue blade.

## 3.4. Modelling results

In order to evaluate the parameters determined for the 3D model, the standard deviations of the residues after subtracting the parameters, the variance explained by each parameter and the root mean squared (RMS) reconstruction error was calculated for the data set.

### 3.4.1. Midsagittal model

As the parameters for the 3D model are based on definitions in the midsagittal plane the variance explained by the control parameters is the highest in that plane, amounting to about 88%. The distribution of the standard deviation of the residues after removing the contribution of each of the five midsagittal parameters is shown in Fig. 5, with the variance explained by each parameter in Table 1. Compared to the midsagittal vocal tract contour model in (Engwall and Badin, 1999), where the parameters were extracted through guided PCA (and calculated over the entire inner vocal tract contour, instead of the tongue), the contributions of JH and especially TD are lower and the parts explained by TB, TT and TA are higher. The lower contribution of JH (16.8% compared to 20.1%) is due to the fact that JH explains variation in the laryngeal part and the part in front of the tongue (lips, mouth floor), that was included in the vocal tract model, but not in the tongue model. The lower contribution of TD (23.1% compared to 45.7%) is partly a consequence of the parameter

Table 1

The midsagittal data variance explained by each of the five articulatory parameters in the order of their extraction and the total variance explained in the midsagittal plane

Parameter	Variance explained (%)
JH	16.8
TB	30.9
TD	23.1
TT	12.1
TA	5.1
Total	87.8

definition, both limiting its possible region of influence and reducing its efficiency in its active region (cf. lines 5–15 in Fig. 5). The first decrease in variance explanation is compensated for by the increased contribution of TB (27.6% vs. 20.5%) and TT (12.1% vs. 4.1%), whereas the second results in a lower total of explained variance, mainly at gridlines 6–8 and 12–14. The role of TA is clearly more important than in the VT model, as it has a much larger importance at the last gridlines,

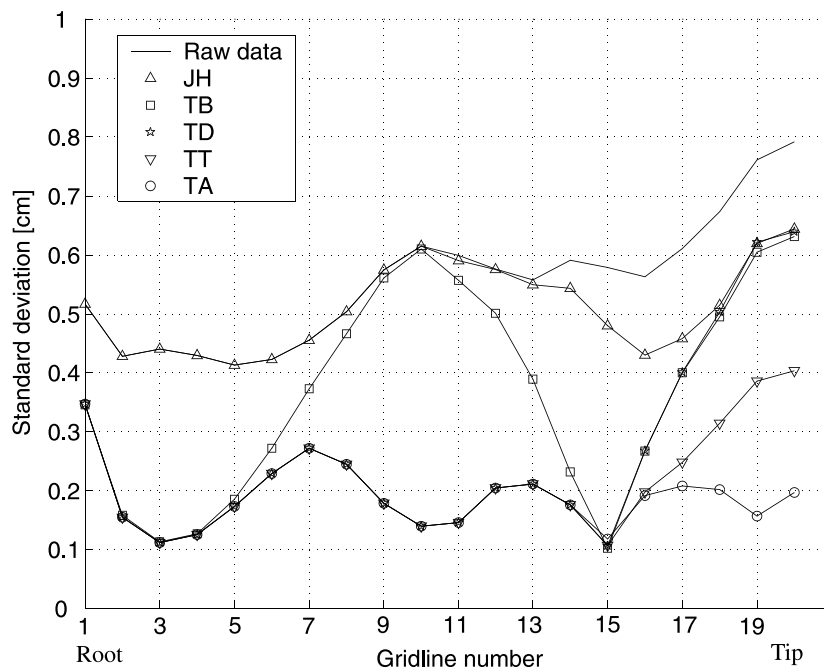


Fig. 5. Standard deviation (in cm) against gridline number of the midsagittal tongue contour for the successive residues of sagittal coordinates when removing the contribution of JH, TB, TD, TT and TA successively.

so that its contribution augments from virtually negligible to about 5%, comparable to the increase found by Badin et al. (2000).

### 3.4.2. Three-dimensional model

Following Badin et al. (2000), the data variance in the 3D model was calculated over the lingual region delimited by the two main axes of the semi-polar grid, excluding the frontal lower part of the tongue base (that is both below line 5 and in front of line 14 in Fig. 2b). That part of the tongue was excluded as it is of little importance to the tongue surface shape and is little influenced by the control parameters. The percentage of the total sagittal data variance explained by the five factors JH, TB, TD, TT and TA amounts to over 78% (if the variance explained is calculated over the excluded vertex points as well it drops to 67%), with the respective contributions shown in Table 2. As the articulatory parameters were established using LCA, they were not orthogonal, and as shown in Table 3, they were actually correlated to varying extents. Note that the correlation between JH and TT is negative because their movements are defined in opposite directions: an open jaw (large JH value) is positively correlated with a low tongue tip position (small TT value), and the correlation between the parameters is hence negative.

The sagittal residual variance in the pharyngeal, velar and anterior oral parts is plotted against lateral fibre number in Fig. 6, where fibre 10 corresponds to the midsagittal plane, lateral fibres below 10 to the left half and fibres above 10 to the right half of the tongue. Fig. 6 illustrates the importance of different parameters in different parts of the tongue, indicating that the pharyngeal part

Table 2  
The variance explained by each factor on the 3D data set, given in the order they were estimated

Parameter	Variance explained (%)
JH	13.5
TB	27.6
TD	14.8
TT	16.0
TA	6.4
Total	78.2

Table 3  
Correlation coefficients between the articulatory parameters

	JH	TB	TD	TT	TA	TW
JH	1.00	-0.11	0.07	-0.57	-0.23	-0.03
TB	-0.11	1.00	0.43	-0.09	0.37	0.09
TD	0.07	0.43	1.00	-0.48	-0.03	-0.56
TT	-0.57	-0.09	-0.48	1.00	0.33	0.44
TA	-0.23	0.37	-0.03	0.33	1.00	0.33
TW	-0.03	0.09	-0.56	0.44	0.33	1.00

is almost exclusively controlled by TB (Fig. 6a), that TB and TD are the main contributors to the variance in the velar region (Fig. 6b) and that JH, TT and TA are the important factors for the tongue blade and tip (Fig. 6c).

To assess the model's reconstruction abilities, the RMS reconstruction error compared to the initial data was calculated for each articulation as

$$\text{RMS} = \sqrt{\frac{\sum_{i=1}^N \epsilon_i^2}{N}} \quad (8)$$

where  $\epsilon_i$  is the reconstruction error at vertex  $i$  and  $N$  the number of vertices. The RMS reconstruction error over the articulations is shown in Fig. 7 and the overall RMS error was 0.13 cm for sagittal coordinates and 0.12 cm for lateral coordinates. The maximal RMS error was 0.17 cm sagittally and 0.16 cm laterally. The reconstruction error was hence slightly larger than the measurement accuracy, as the image resolution is 1 mm/pixel and the extraction was pixel-based.

No clear connection can be found between consonant category and sagittal reconstruction error (Fig. 7a), other than an indication that front fricatives in [ɪ]-context are less well reconstructed. These fricatives are articulated with the front-most part of the tongue blade raised to the alveolar ridge, and the movement of that part of the tongue in that direction is not accounted for by TB, nor is it handled to the full extent with the tongue tip parameters TT and TA (cf. Fig. 5, where a local maximum remains at gridlines 17–18). This larger reconstruction error is probably due to the non-linear compression of the tongue against the alveolar ridge.

The lateral reconstruction error for the consonants is dependent on the vowel context, being

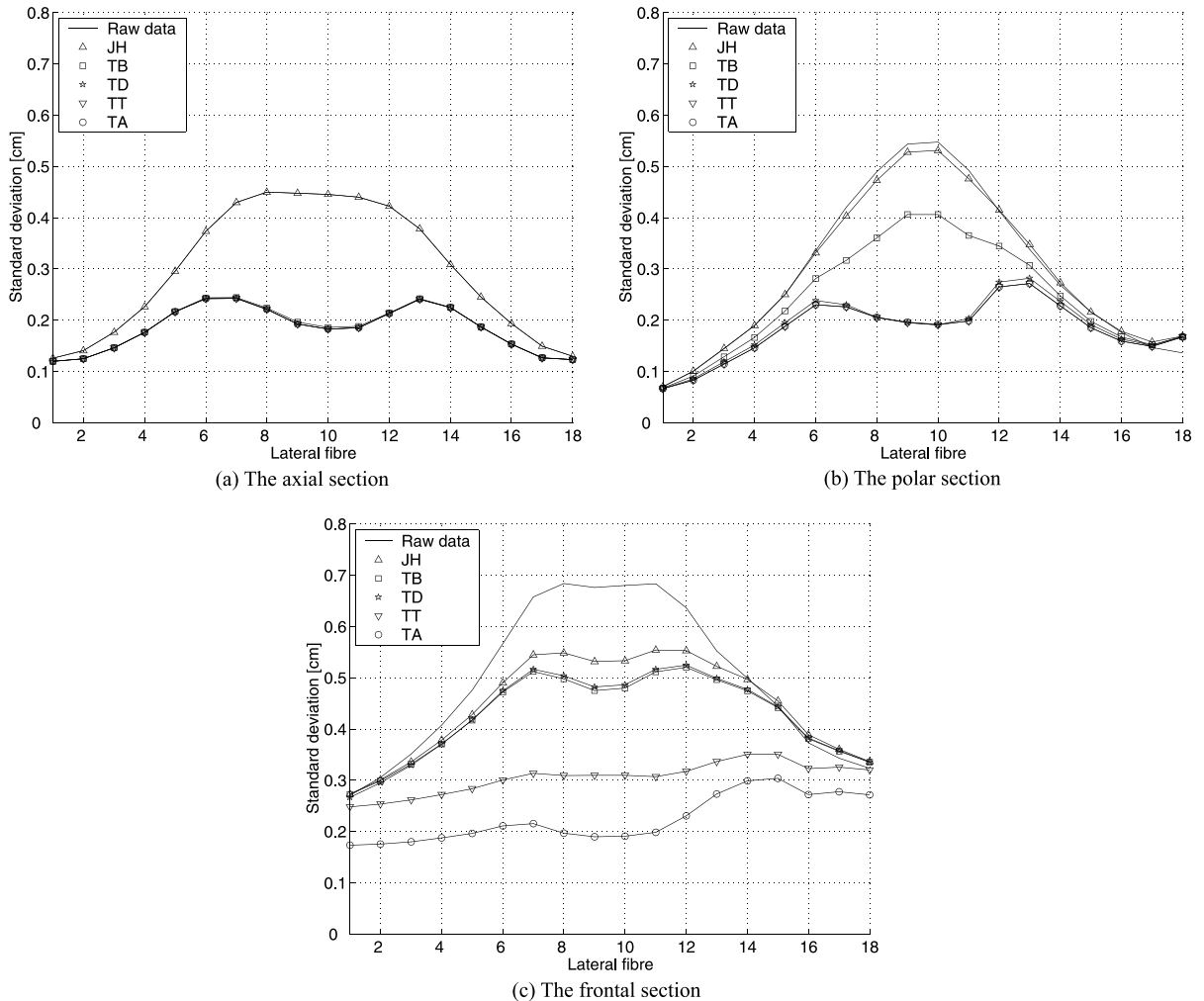


Fig. 6. The residues of sagittal coordinates of the mean standard deviation for the three tongue parts as a function of lateral fibre (summed over all vertices belonging to each part of the fibre), after removing the first five articulatory parameters in the order JH, TB, TD, TT and TA. Fibre 10 corresponds to the midsagittal plane. The graphs are clustered into two in the axial section, as JH, TD, TT and TA have very little influence on the variation in that part.

largest for [ɪ]-context for all consonants with the exception of [t] and [l], and being below or at average in [a]-context for all consonants except [k] (Fig. 7b). The larger lateral reconstruction error for [ɪ]-context can be attributed to the bracing of the tongue against the palate, leading to tongue width variations dependent on the place of articulation.

The sagittal and lateral reconstruction errors were negatively correlated for vowels (Fig. 7c),

so that articulations with lower sagittal reconstruction error had larger lateral error and vice versa. This difference was largest for the most closed, [yɪ, iɪ, ɥɪ], and the most open, [æɪ, øɪ, œɪ], front vowels.

An acoustical evaluation, in which the synthesis was based on the calculation of area functions from the 3D vocal tract shape, has been performed in (Engwall, 2001) to assess the model's ability to reproduce the reference subject's target

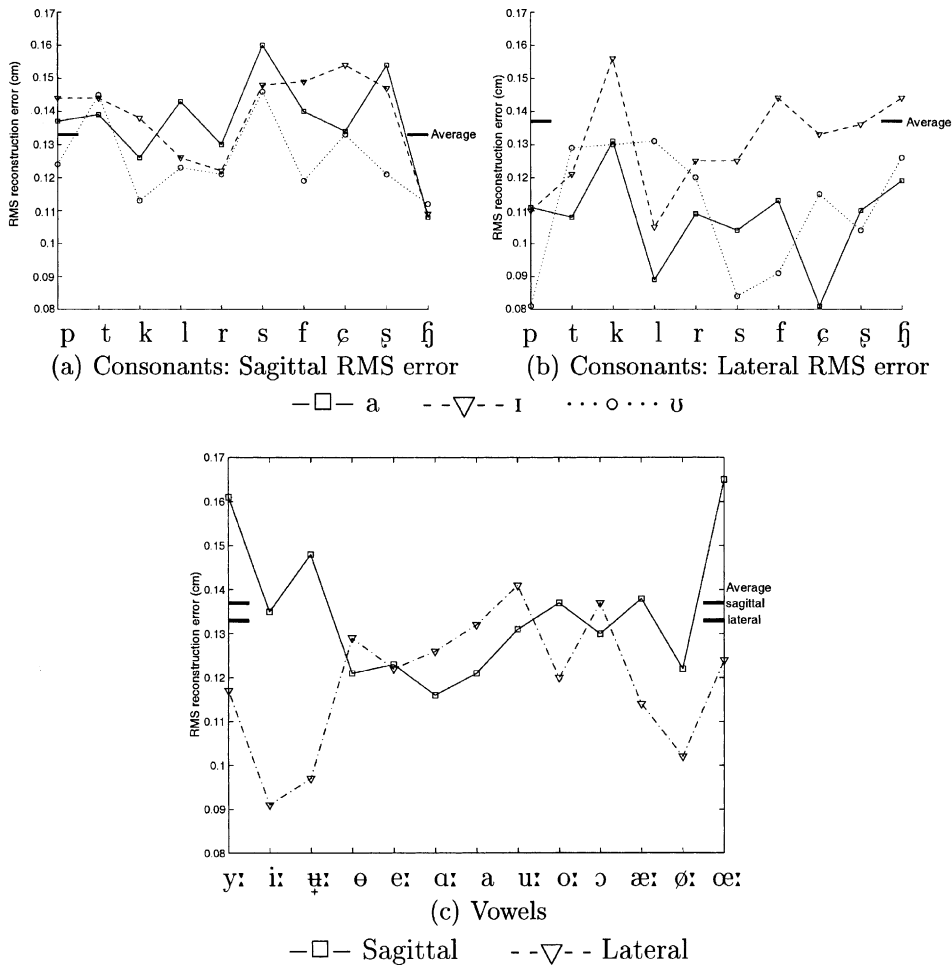


Fig. 7. The sagittal and lateral RMS reconstruction errors.

vowels. The study concluded that [i:, e:] and to some extent [y:] differed more from the reference subject's vowels, both considering formant frequencies and in the perceptual identification test using 20 subjects. This is worth noting, in relation to the statement about the problem with non-linear tongue compression against the palate in the alveolar region, as these vowels are articulated with the tip and blade braced against the front palate, creating a narrow air passage in a region of the model (gridlines 17–18) where the reconstruction error is large (up to about 2 mm). The reconstruction error in the alveolar region was hence showed to have acoustical consequences.

The parameter definitions imposed by the model environment are clearly suboptimal for data variance explanation in some regions and one method of improving the model further would consequently be to redefine the parameter concept in the model, allowing one parameter to create deformations in several directions simultaneously.

The model is nevertheless able to replicate the initial data with fairly low reconstruction error. Moreover, lateral differences and observed asymmetries in tongue shape are handled by the model, as exemplified by the asymmetry in tongue edge height in [æ:] (Fig. 8a) and the important grooving in the velar region for [ʊ<sub>ɕ</sub>ʊ] in Fig. 8d (discussed in Section 3.3.2).

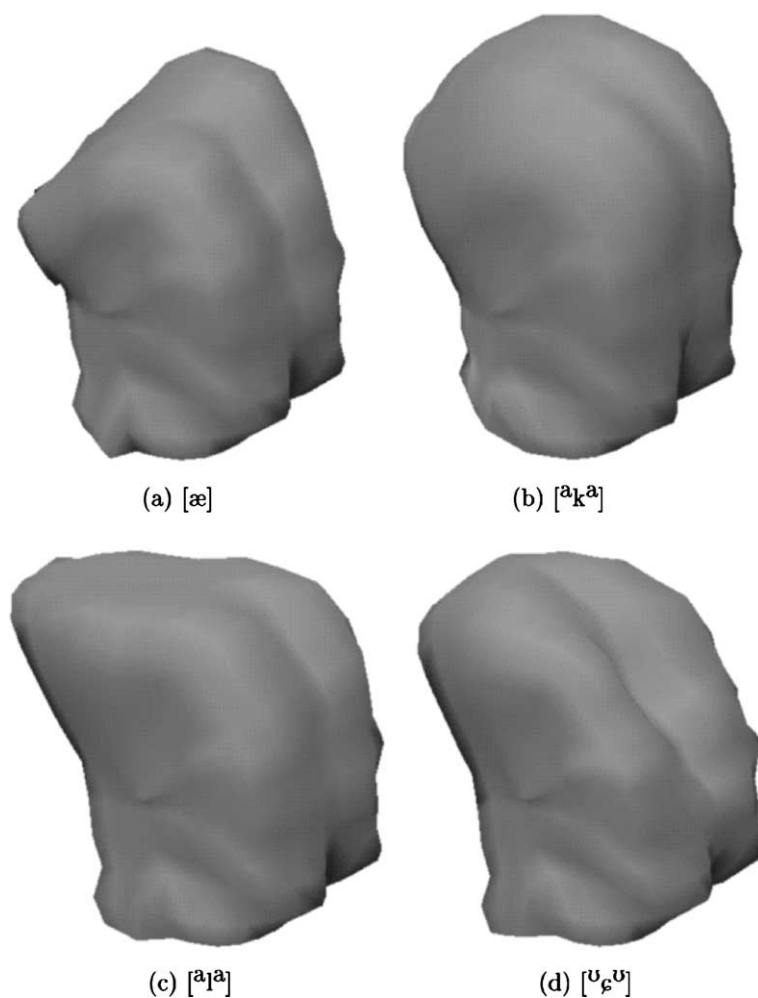


Fig. 8. Examples of replicated tongue shapes in the tongue model.

Badin et al. (2000, p. 262) noted “a fairly clear backward displacement of the tongue in the MR images compared to the X-ray images” when evaluating their midsagittal tongue model. Tiede et al. (1997) also found that the tongue posture was shifted rearwards in supine articulations measured with EMA, and Engwall and Badin (1999) found that the vocal tract shapes reconstructed from the same MR images as in this study were sometimes too narrow in the pharyngeal part, compared to the midsagittal images (due to the longer acquisition time, during which the subject was not able to compensate as well for the gravitational effects)

and to reference tracings of Swedish articulations in X-ray images in (Fant, 1965).

It can hence be argued that the tongue articulations presented here are unusually backward, due to the supine position of the speaker or the sustained articulations during the acquisition. The influence of the supine position has been investigated by Tiede et al. (2000), who found that the two subjects adapted the supine articulation to maintain consistency in the acoustics and that the posture effects were least for running speech and largest for sustained vowels. A combined EPG and EMA study of fricatives (Engwall, 2000a) did in-

deed conclude that the tongue positions in the MRI data differed somewhat from the EMA and EPG data, and judged that the static articulations were hyperarticulated. The correspondence between the MRI data and the set of real-time data was generally good considering the tongue position in the fricative articulations, in particular at the point of constriction (though a backward displacement was found in the MRI data for [s, ʃ]), but the coarticulatory influence on the fricative was substantially reduced in the static condition.

#### 4. Palate and teeth models

New models of the palate and upper and lower teeth were generated for the KTH 3D Vocal Tract project from MR images of the subject's dental cast (Engwall and Badin, 1999). In order to allow real-time display of the model in a text-to-visual speech synthesis system, a simplification was called for. This consisted in four steps:

- (1) The teeth were first identified in every contour, separating them from the palate and the gums.
- (2) The subcontours were subsampled such that all teeth contours had the same number of points, all palate contours the same number etc. The subsampling was made in such a way that teeth contours, which are more complex, were given twice as many points as the palate and the gums.
- (3) A polygon mesh was generated connecting each point to its neighbour and the two corresponding points on the adjacent contour (as for the tongue mesh generation, cf. Section 3.1). The contours to be included in the mesh generation were chosen manually, based on the difference between consecutive contours.
- (4) To reduce the polygon number sufficiently while retaining the wanted level of detail, the jaw and palate models were made symmetrical, discarding the right parts and making them a reflection of the left.

This process simplified the teeth-palatal model to about 800 vertices and 1400 polygons.

#### 4.1. Handling surface contacts

The combination of a moving tongue and fixed structures, such as the teeth and palate, makes a method for handling contacts between surfaces necessary. The tongue movements could otherwise lead to physically impossible articulations where the tongue penetrates the bounding surfaces in the oral cavity. The best solution to this problem would be to introduce physiological constraints, based on dynamical considerations of the effect of inertia, to model the deformation of the tongue when it enters into contact with the palate (cf. Wilhelms-Tricarico, 1997). Perkell (1974) introduced an explicit impenetrability threshold to handle these collisions and Dang and Honda (1998) used reaction forces from the vocal tract wall to bring the tongue mass points to an equilibrium position after the collision.

Such algorithms are however still too complex for the current text-to-visual speech application, with respect to the real-time constraint. A simple detection and correction algorithm has hence instead been introduced in the model to avoid impossible situations, where the tongue penetrates the teeth, the gums or the palate.

The algorithm consists of detecting tongue points that have penetrated the teeth or palate boundary and then correcting these points by placing them on the boundary surface instead, similar to the method proposed by Cohen et al. (1998). The method does hence not model the tongue deformations adequately, but it constrains the tongue to remain in the physically allowed space, which is the chief concern in the current text-to-visual speech application, even if it is too crude an approach for physiological studies.

##### 4.1.1. Boundary meshes

The testing and correction is carried out against one upper boundary mesh,  $B_p$  (cf. the black mesh in Fig. 9), consisting of the palate and the inner parts of the teeth and gums, and one lower,  $B_j$ , consisting of the inner parts of the lower teeth and gums. The boundaries are made up of regular quadrilateral meshes of sagittal and coronal (for the palate) or sagittal and axial (for the jaw) lines, defined by regular sampling of the original

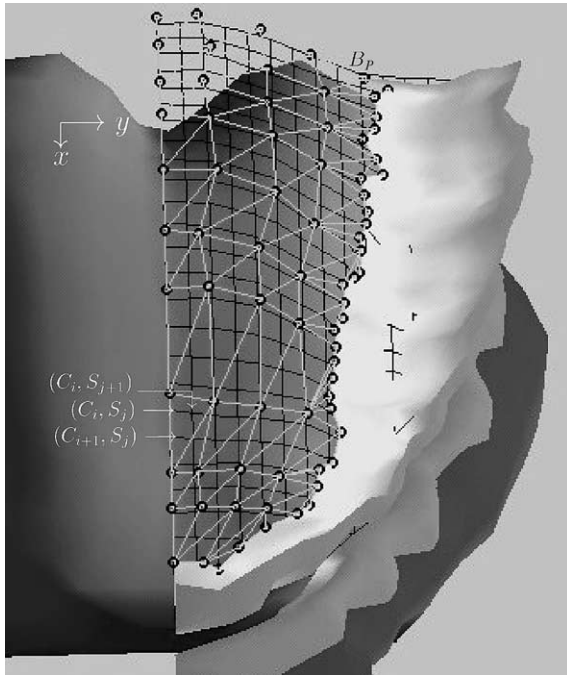


Fig. 9. The upper boundary mesh  $B_p$  (the finer black mesh), created from interpolation and sampling of the palate (the bright mesh with circular markers) and the inner part of the teeth.

polygon surfaces. A mesh size of  $2 \times 2$  mm squares is used, as a compromise between modelling accuracy and a sufficient computational speed.

As the palate and jaw do not change shape over time, the boundary meshes are pregenerated and stored as separate structures in the model. The palate is kept in fixed position during the speech synthesis and the boundary mesh used in the testing is hence defined directly from its orientation in the reference position. For the jaw, an additional step is needed, as the jaw is moved during speech. The jaw boundary mesh is kept in its reference position with closed jaw when the check and correction is to be performed and rather than rotating the boundary mesh, the tongue points to check are transformed to the coordinate system of the jaw. After detection and correction, the corrected points are transformed back to the original coordinate system. This solution was chosen instead of rotating the jaw as slightly fewer points have to be transformed in each frame, and, more

importantly, this allowed the detection algorithm used for the palate to be applied to the jaw as well. The method is described below for the palate, but it also applies to the jaw mesh, with the minor change that the search is based on axial rather than coronal lines.

#### 4.1.2. Correction and detection algorithm

The algorithm is based on a careful definition of the boundary mesh, allowing the detection and correction to be carried out efficiently in one step. Firstly, the palate mesh is ordered with increasing  $x$ -coordinate back-to-front for the coronal lines  $C_i$  and then with increasing  $y$ -coordinate for the midsagittal plane and outwards for the intersections with the sagittal lines  $S_j$  within each coronal line (refer to the  $(C_i, S_j)$  pairs in Fig. 9 for an illustration). Secondly, the sampling of the mesh is made at integer  $x$ - and  $y$ -coordinates (the unit is mm). As a consequence all distance calculations to find the polygon closest to a tongue point can be omitted since that polygon is directly identified by the  $(x, y)$ -pair of the tongue point  $P = [x_P, y_P, z_P]$ . The first corner in the closest boundary polygon has coordinates  $[2 \cdot \text{fix}(x_P/2), 2 \cdot \text{fix}(y_P/2), z_1]$  and the remaining three  $[2 \cdot \text{fix}(x_P/2) + 2, z_2]$ ,  $[2 \cdot \text{fix}(x_P/2) + 2, 2 \cdot \text{fix}(y_P/2) + 2, z_3]$ ,  $[2 \cdot \text{fix}(x_P/2) + 2, 2 \cdot \text{fix}(y_P/2) + 2, z_4]$ , with  $\text{fix}$  denoting rounding towards zero.

The vectors from  $P$  to the four corners are then calculated and the shortest of these approximates the vector from the tongue surface to the boundary. The sign of the dot product between this vector and the surface normal  $\vec{n}$ , given by the vectors spanning the square, shows whether  $P$  passes the surface or not. Once an intersection is detected for  $P$ , it is mapped upon the closest polygon corner found in the detection process. This strategy speeds up the correction step, as it eliminates the search for the polygon closest to  $P$  in the original polygon surface and the projection onto it.

The maximal error in the correction is  $\sqrt{2}$  mm, which is of the same level as the reconstruction error of the synthetic tongue shape compared to the natural, and thus of sufficient accuracy for the model. The risk of clustering of corrected points is negligible, as the mesh is significantly finer than

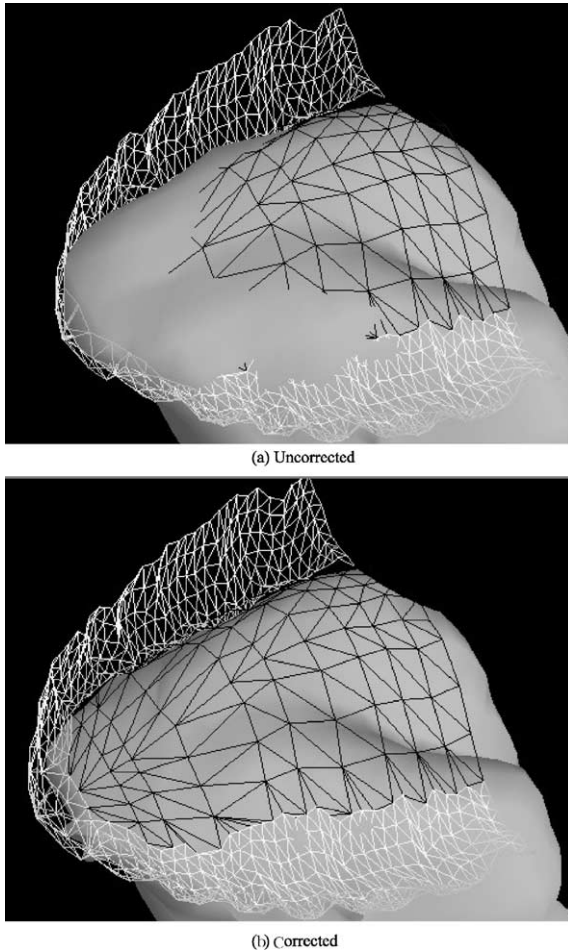


Fig. 10. Example of correction for the tongue.

the tongue polygon mesh. The outcome of the correction method is exemplified in Fig. 10.

The algorithm is further speeded up by testing only relevant parts, instead of the entire tongue. The upper surface and edges of the tongue, from the tip back to the velar region, are tested for boundary violation against the palate, whereas only the tongue tip and edges are tested for boundary violation against the jaw. Instead of checking all 450 vertices of the tongue against two boundary meshes, only about 230 checks need to be performed. The time for detection and correction is less than 10 ms/frame, hence allowing the model to be rendered at 50 frames/s.

## 5. Electropalatography analysis

Following the conclusion in Section 3.4.2 that the tongue model did need some articulatory adjustment to decrease the effect of the supine position and the artificial sustaining, the EPG data was used to find differences between the linguopalatal contact pattern of the model and that of the subject and to adjust the articulatory parameter values accordingly.

The attempted readjustment is an approximation, as there is no guarantee that a tongue vertex in contact with the palate in the model corresponds to the fleshpoint that was measured to be in contact with an EPG electrode in the natural pattern. If, however, the entire EPG pattern of 62 electrodes is considered, and the parameter adjustment starts from the parameter values defined through the analysis of MRI data, the spatial freedom of vertices in the model diminishes, and the correlation between the vertices in contact in the model and real fleshpoints hence increases to a level where the attempted adjustment can be a justifiable approximation.

Three repetitions of each articulation were used in the analysis and the contact pattern was determined at the most constricted phase of the articulation. The mean pattern over the three repetitions was calculated and considered as the natural linguopalatal contact.

### 5.1. The three-dimensional electropalatography palatal model

To be able to compare real EPG data with the model, a computerized 3D virtual EPG palate was generated, allowing the display of the subject's natural contact patterns in 3D and, more importantly, the calculation of virtual contact patterns using the 3D tongue model. The palatal shape was generated from MR images, but without making any of the simplifications described in Section 4, to avoid that these influenced the results of the calculation of virtual EPG patterns.

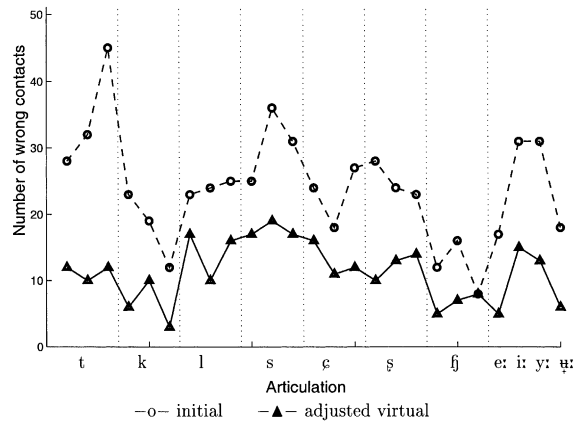
Sixty-two EPG electrodes were placed on the virtual palate according to their placement on the real EPG palate. The electrode coordinates were calculated by superposing a 2D scan of the EPG

palate on an image of the dental cast viewed from the same angle as the EPG palate. The *z*-coordinate of the electrode was then set equal to that of the point on the palate that corresponded to the electrode position in the 2D representation.

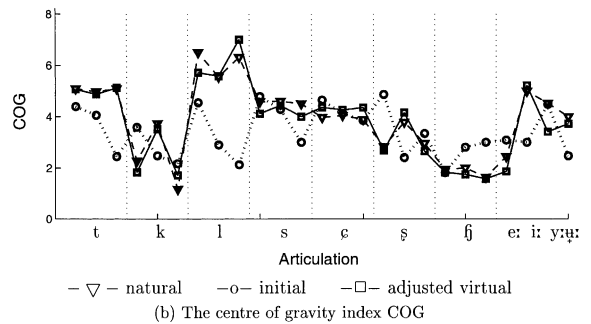
5.2. Determining the virtual electropalatography patterns

The virtual EPG patterns were calculated by checking if the point on the tongue surface closest to the virtual electrode was in contact with the palate surface model, using the dot product, as in the detection algorithm in Section 4.1.2.

The initial contact patterns, based on the parameters determined from the MRI analysis were first calculated and compared to natural EPG patterns. The error was analysed both quantitatively, based on the number of false and missing electrode contacts, and qualitatively, based on the contact pattern displayed with the virtual palate. Fig. 11 gives an example of the difference between the natural and the virtual data. The initial divergence and the type of error vary greatly over the corpus (cf. Fig. 12), not only between consonants, but also between contexts for the same consonant. There was a weak tendency for articulations with much linguopalatal contact to have larger errors (e.g. [t] had larger error than [k, l] and



(a) The total number of contact errors (false and missing contacts) compared to the natural contact pattern.



(b) The centre of gravity index COG

Fig. 12. The improvement after the tuning, measured as the decrease in contact error (a) and centre of gravity resemblance (b). Vowel context from left to right for the consonants: [a, ɪ, ʊ].

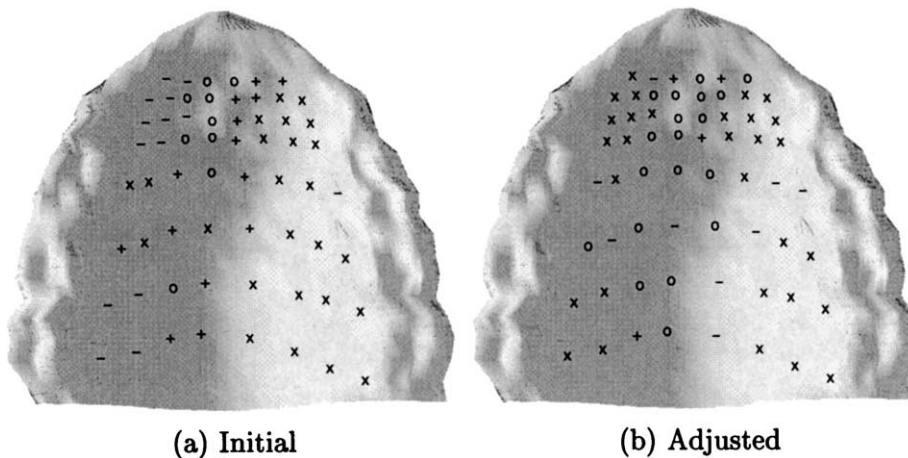


Fig. 11. The virtual EPG palate with information on the resemblance between the natural, initial and adjusted virtual contact patterns for [ʃ]: (x) correct contact, (O) correct non-contact, (+) false contact and (-) missing contact.

['s'] more than ['s<sup>a</sup>']). Some contextual influence on the type of error was also found, with the main error being too many alveolar or palatal contacts for consonants in [a] context, too few alveolar contacts in [ɪ] context and too few contacts in total for front consonants in [ʊ] context (it is worth noting that these observations are consistent with a backward displacement of the tongue in supine position and hyperarticulation, diminishing the influence of the vowel). The relation between closed vowel context and too little alveolar contact did not generalise to vowels, however. [i:, y:] both had high initial errors, but in [i:] it was due to too few alveolar contacts and in [y:] to too many. In total over all articulations, with maximally 1550 contacts, 370 missing and 230 false contacts were registered (a total error of 38%) and 15 articulations had more missing than false contacts, 7 had more false and 3 equal number of missing and false contacts.

Note however that the total error of wrong contacts is not a very good measure of the modelling quality of an articulation in the original model, as it does not indicate the distance by which a contact was missed. The articulations with many missing contacts initially, [t, s, i:], were generally close to the palate even where there was no contact.

The type of error and its distribution, on the other hand, give better indications of weaknesses in the model. The two types of error were considered as equally wrong and the following adjustment of parameter values consisted of minimizing the sum of the two errors, rather than aiming for an equal error rate of missing and false contacts.

## 6. Parameter adjustment

For every articulation, the difference between the natural and virtual EPG patterns was minimized with a combinatory search varying each of the six articulatory parameters  $A_\gamma$  in the interval  $A_{\gamma\text{MRI}} - 0.4 \leq A_\gamma \leq A_{\gamma\text{MRI}} + 0.4$ , where  $A_{\gamma\text{MRI}}$  was the parameter value determined by the component analysis of the MRI data and 0.4 was an empirical choice to provide a suitable interval for the tuning.

The combination of parameter values closest to  $A_{\gamma\text{MRI}}$  that resulted in the least deviation from the natural pattern was chosen as the new parameter values. Only the articulatory parameter values were tuned; the weights  $W_\gamma$  that determine the parameter influence on each vertex in the mesh and the axes of activation were maintained from the MRI based model, as the weights could not be tuned without introducing uncontrolled influence on articulations without palatal contact. In order to adjust the weights as well, measurements of the linguopalatal distance at several points along the tongue surface would be needed for parts that are not in contact with the palate. This will be possible with the optopalatograph system (Wrench et al., 1998) when it becomes more widely available, but with EPG there is not enough information to tune the weights in a manner that would be correct for all articulations. The vertex weights thus remain common for all articulations and each articulation was tuned individually by varying its parameter values in successively smaller steps.

### 6.1. Results

Fig. 12a shows the difference between the natural and the modified virtual patterns and the improvement from the initial virtual patterns. The total number of deviating contacts was lowered substantially for all articulations but [ʊfʊ] and the summed error over the whole corpus was more than halved (183 missing + 101 wrong = 284 in total vs. 370 + 230 = 600), with a mean error of 18%.

The characteristics of the natural contact pattern were further quite well replicated after readjusting the parameters, if measured by the centre of gravity index, COG, displayed in Fig. 12b, where

$$\text{COG} = \frac{8R_1 + 7R_2 + \dots + R_8}{R_1 + \dots + R_8} \quad (9)$$

and  $R_i$  denotes the number of contacts in row  $i$  (Nguyen, 2000). Even if the total number of remaining errors was as high as 18%, the overall contact patterns are thus nevertheless correct and the improvement compared to the initial pattern was clearly evidenced in the COG index. This is

explained by the fact that switching the correct, but different, values of two electrodes in the same row counted as two contact pattern errors, while the COG index was unaffected, as was the case for example for [t] in all vowel contexts.

### 6.1.1. Remaining error

The distribution of the remaining error was investigated for common features. The phonemes with alveolar contact and free electrodes either centrally, [s, i:, y:], or laterally, [l], have larger remaining errors, which is probably caused by the fact that the bracing and non-linear deformation of the tongue against the palate is not taken into account, making it hard to accomplish inhomogeneous contact patterns with the tongue tip and blade. Another problem occurs for articulations with a low tongue body and a high tongue tip, such as [a<sup>s</sup>, u<sup>s</sup>], which have mainly false velar contacts. These articulations have larger errors than articulations with either both high tongue body and tip (e.g. [t<sup>l</sup>]) or only high tongue body (e.g. [e:]).

The most common error was lateral shifts, e.g. false right velar contacts in [s, u<sup>s</sup>, ʃ], missing central velar contacts in [s<sup>l</sup>, ʒ, a<sup>s</sup>, s<sup>l</sup>, a<sup>l</sup>, ʃ] and missing lateral alveolar contacts in [l<sup>l</sup>, i:]. The lateral asymmetry index,

$$\text{LAT} = \frac{(C_5 + \dots + C_8) - (C_1 + \dots + C_4)}{C_1 + \dots + C_8} \quad (10)$$

(Marchal and Espesser, 1987), was therefore calculated. The natural contact pattern was found to be shifted to the left, with  $\text{mean}(\text{LAT}) = -0.21$  and  $\text{std}(\text{LAT}) = 0.08$ . The tongue model was not able to replicate this lateral asymmetry fully, having  $\text{mean}(\text{LAT}) = -0.15$  and  $\text{std}(\text{LAT}) = 0.18$ . This is due to the fact that only the activation of the parameters was tuned, whereas lateral adjustments would call for changes to the parameter influence weights  $W_v$  as well.

For future lateral adjustment, the optopalatography system by Wrench et al. (1998) mentioned above would be ideal, as it provides real-time lateral information of both linguopalatal contact and distance. The information on the distance from the sensor to the closest tongue point can then

be mapped on the current model, letting it interpolate between the measured fleshpoints, and adjust the vertex weights to get a smoothly varying tongue shape passing through the measured points.

## 7. Kinematic parameter control

The main problem with MRI in speech production modelling is that the measurements are static and the articulations thus artificially sustained. The EPG data was used in the previous section to readjust the parameters from hyperarticulated to values for normal speech. Information is however still needed on the articulatory movement and the timing of different articulators to make the tongue model move in a realistic way.

Tongue movements are dynamic, and are modelled as such in biomechanical models, where an equation of motion is solved to determine the displacement of the model as a function of muscle activation and the effects of inertia. A simpler, and computationally faster, alternative is to focus on a kinematic description of the tongue, dealing with the trajectory of tongue movement, instead. The dynamics of the parts of the tongue is then disregarded for a phenomenological description of the movement of the tongue surface, often coupled to observed trajectory records in measurements with e.g. X-ray or EMA.

This section describes how the EMA data was used in this respect, as an attempt to control the movement of the tongue model with articulatory data. The Movetrack EMA system (Branderud, 1985) measures the movement of receiver coils placed in the midsagittal plane and as the articulatory parameters of the tongue are defined in that plane, the measurements of the coil movement can be used to control the tongue parameters.

This study uses the simplified assumption that each of the coils can be coupled to different articulatory parameters in the model. This is a simplification as the EMA measurements are of fleshpoints, coils physically attached to the tongue, whereas the MRI data represents a continuous shape, where it is not evident exactly which point corresponds to which in two different shapes.

Generally, the problem of relating the EMA data to the MRI based model should therefore be dealt with as an inversion problem, where the control parameters of the MRI model should be searched combinatorially to find the combination that makes the midsagittal tongue contour pass through the locations defined by the EMA coils (cf. Badin et al., 1997). However, it was found in the fricative study in (Engwall, 2000b) that the movement of the second coil  $T_2$  within the VCV sequences was along the front-back line corresponding to the activation line of TB (cf. Fig. 4) and that  $T_3$  moved more or less along the arching-flattening line of parameter TD (cf. Fig. 4). For this restricted corpus of fricatives in symmetric vowel context it can hence be a justifiable simplification to control TB by the measurements from  $T_2$  and TD by the measurements from  $T_3$ , taking the movement of the EMA coils as a measure for a region of the tongue rather than one specific fleshpoint. The reason for testing this simplified approach is that one discussed potential application of the tongue model is in a biofeedback loop for speech rehabilitation, displaying in real-time the patient's tongue movement measured with EMA. If the inversion technique were to be used, the model would not be able to respond in real-time, which would greatly diminish the use of the feedback. The important information for the patient is moreover about timing and direction of the articulatory movement, rather than the exact fleshpoint displacement, and the approximation would hence be acceptable in the biofeedback application.

It is noteworthy however, that the simplified assumption would not be valid with a more varied corpus, and Engwall (2002a) consequently used the inversion technique to define TB and TD when the corpus was the phonetically balanced sentences in the MOCHA-TIMIT database (Wrench and Hardcastle, 2000).

Another issue when combining the EMA and MRI measurements is the posture effect, which might lead to different articulations in supine and upright position. Tiede et al. (2000, p. 28) found that the articulatory trajectories in CV sequences were altered in supine position, to produce consistent acoustics, but that the “acoustically sensi-

tive targets involving narrow constrictions are produced with little variability between postures”. It is hence assumed in this study that the fricative articulations measured with supine MRI and upright EMA do correspond.

### 7.1. *Selecting electromagnetic articulography measures*

The EMA coils were placed on fleshpoints corresponding to the prototype definitions for the articulatory parameters. The data from the coil  $J$  on the lower incisor controls the jaw height parameter (JH) and the three coils on the tongue,  $T_1 - T_3$ , the movement of different parts of the tongue.

The measurements for each coil consist of its movement in the  $x$ - and  $y$ -directions in the midsagittal plane, considered separately, as functions of time,  $x = f_1(t)$  and  $y = f_2(t)$ . The jaw movement is almost exclusively vertical and the kinematics of the jaw height parameter JH is thus controlled by the vertical movement of coil  $J$ . Measurements in both  $x$ - and  $y$ -directions of  $T_1$  were used, to control the advancing-retraction (TA) and the raising-lowering (TT) of the tongue tip and blade, respectively. As mentioned above,  $T_2$  measurements were used to control the tongue body movement (TB) and  $T_3$  to control the tongue dorsum movement (TD). The widening and narrowing of the tongue blade, controlled by the tongue width parameter TW, is not directly measured by any of the EMA coils. TW is however highly correlated with the tongue dorsum movement, since the tongue blade is narrowed as the tongue volume grows in the velar region and widened as the tongue dorsum is lowered, because of volume conservation, but also because the tongue dorsum lowering coincides with anterior tongue-palate contact that spreads the tongue laterally. The sixth tongue parameter TW is thus also controlled by the  $T_3$  measurements, to ensure volume conservation. An empirical relation between the two parameters TD and TW was determined from the component analysis of the MRI data, such that TW is linearly proportional to the level of TD ( $TW = TW_0 + kTD$ ,  $k < 0$ ).

### 7.2. Scaling measurements to parameter activation

The EMA measures of the coil movements are given as the deviation in cm from the reference position, which was the same as for the MRI measurements, i.e. the tongue rested on the floor of the mouth with the tongue tip touching the lower incisor and with upper and lower incisors touching and horizontally adjusted to be in line. The parameters in the tongue model are however defined using the activation level of a prototype movement towards a target (cf. Section 3.3) and the EMA measures need hence to be transformed from distance measures to activation of the corresponding parameter.

The parameters for the tongue, TB, TD and TT, control the portion of the motion of the respective parts of the tongue that is not due to the jaw movement. The contribution of the jaw opening to the tongue coil deviation from the reference position was hence removed from the total deviation before the parameters TB, TD and TT were calculated from the measurements of the respective coils, as

$$T_{i,y}^{\text{relative}} = T_{i,y} - \text{JH}_{\text{infl}}^{T_i} \cdot J_y \quad (11)$$

where  $T_{i,y}$  is the measured deviation of tongue coil  $i$  in the  $y$ -direction,  $J_y$  is that of the jaw coil  $J$  and  $\text{JH}_{\text{infl}}^{T_i}$  is the influence of the jaw height at the position of coil  $T_i$  according to the MRI based tongue model.

The Euclidean deviations from the reference tongue shape were scaled to correspond to the parameter activation levels defined with the statistical analysis of MRI data in Section 3. The EMA data was moreover downsampled from 2000 Hz to 100 samples per second, corresponding to the frame rate of the KTH visual speech synthesis, giving six parameter activation functions  $A_i(t)$  that can be used as input to the tongue model, to replicate the measured VCV sequences, as an illustration of how the model could be used in a biofeedback loop.

### 7.3. Generating synthetical control sequences

In the more general text-to-visual speech synthesis application, rules for generation of articu-

latory trajectories from text are needed. Such rules need to take articulator timing and coarticulation into account (e.g. Cohen and Massaro, 1993; Pelachaud, 1991; Le Goff and Benoît, 1996; Masuko et al., 1998) when generating the parameter values for the synthesis. In doing so, it is commonly assumed that the articulators are controlled by a sequence of phonemic target gestures (cf. Cohen and Massaro, 1993). This assumption has been refuted in articulatory studies, such as Fujimura (1990), and e.g. Krakow (1999) claimed instead that the unit for articulatory organization is rather the syllable. The control of the articulator movement would hence require longer planning than controlling the articulators by static target positions for individual phonemes. In currently available audio-visual speech synthesis, where the visual speech is combined with a text-to-speech synthesis that is based on the principle of concatenation and coarticulation of phonemic segments, it is nevertheless natural to use the concept of phonemic target gestures.

For the VCV sequences discussed here, such synthetical parameter control sequences  $A_s(t)$ , as shown in Fig. 13, were generated to approximately replicate the EMA based measures (dashed lines in Fig. 13), considering

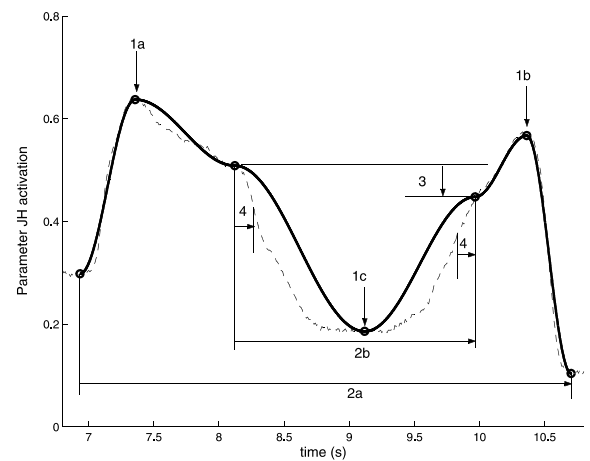


Fig. 13. Generation of the synthetic control sequence for JH (solid line), based on the prototypic measures of the vowels (1a and 1b), the fricative in that context (1c), the duration of the VCV sequence (2a) and the fricative (2b), the difference between onset and offset of the fricative (3) and the fricative parameter timing (4) measured in the natural EMA sequence (dashed line).

- (1) one prototypical parameter activation value for each vowel before (1a) and after the fricative (1b) and for each fricative in the appropriate vowel context (1c) as the targets in the control sequence,
- (2) the entire rise-fall sequence of the parameter control as six sinusoidal functions (two for each of the vowel parts and two for the fricative) with smooth transitions and requiring that the functions match the mean measured duration times for the VCV unit (2a) and the fricative (2b),
- (3) the relation between the measured onset and offset position for the articulators. Engwall (2000b) e.g. found that the tongue tip was lower at offset than at onset,
- (4) the relation between timing of different parameters. Engwall (2000b) e.g. found that the tongue tip motion precedes that of the jaw both at fricative onset and offset.

The synthetical parameter activation functions were used to synthesize all VCV unit combinations of the five fricatives and the three vowels. As a future development of the tongue model, it would be possible to base the control of all articulatory trajectories in the text-to-visual speech synthesis on a larger and more general EMA database, which would possibly improve the naturalness of the model's tongue movements.

## 8. Results and future work: a kinematic three-dimensional tongue model

Several different measurement sources have been used to create an articulatory 3D model. The shape and parameters are determined through statistical analysis of static MRI data, the parameter activation is based on the combination of MRI and EPG and the timing of the movements is determined from EMA data. Animated sequences of the tongue, controlled by EMA measurements, are available in the accompanying image files at [www.elsevier.nl/locate/specom](http://www.elsevier.nl/locate/specom).

The presented 3D MRI data on tongue shapes and the articulatory tongue model provide an important source of information on 3D articula-

tions, especially as the evaluated corpus is much larger than in earlier studies of 3D tongue shapes using MRI (e.g. 25 articulations in Badin et al., 2000). More importantly, it replicates, for another subject and another language, the findings in (Badin et al., 2000) that the 3D tongue shape can be linearly controlled using articulatory parameters defined in the midsagittal plane.

It should be noted that the presented 3D tongue model represents on-going work and that much remains to be done. More real-time EMA and EPG data is needed, in order to supplement the static and supine MRI data, to be able to model the kinematics and the upright speaking position correctly. The exact relation of measurements from the different datasets should in future, more exact development, be based on computations of tongue shape deformations (cf. Section 7 for a discussion on the inversion problem), rather than empirical readjustment. Real-time continuous data, such as from X-ray, of the reference subject would also be useful to evaluate how accurate the present model really is.

In the longer run, when more computational power and relevant approximations are available, a more physiologically oriented approach should be implemented. This is needed firstly to deal with the linguopalatal contact in a more exact manner and secondly to model the dynamics, rather than the kinematics of the tongue, as the dynamic nature of the tongue is crucial for an exact understanding of the speech production.

Ongoing work (Engwall, 2001) focuses on modelling the remaining parts of the vocal tract, so as to make it possible to evaluate the model regarding its acoustic properties and bringing it closer to the goal of a real-time multimodal 3D speech synthesizer. The tongue and inner structures have further been introduced in the KTH rule-based visual speech synthesis system (Beskow, 1995) to allow for text-to-intraoral-visual speech synthesis. The multimodal speech synthesis is based on the RULSYS text-to-speech rule synthesis framework (Carlson et al., 1982), where the orthographic text is transformed to strings of phonemes (for the audio output) and articulatory parameters for the face and tongue models (for the visual modality). The results of the EMA study

was taken into account as far as possible in the synthesis framework, including timing between articulator movement and parameter activation of the fricatives.

### Acknowledgements

This research was carried out at the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. Pierre Badin (Institut de la Communication Parlée, UPRESA CNRS 5009, INPG-Université Stendhal, Grenoble, France) and Christoph Segerbarth (INSERM U438, Centre Hospitalier Régional Universitaire Grenoble) assisted with the MRI acquisition, Elisabet Eir Cortes of the Department of Linguistics at Stockholm University assisted with the EMA data acquisition and Elisabet Eir Cortes and Peder Livijn (Department of Linguistics at Stockholm University) assisted with the EPG recordings. The three anonymous reviewers for *Speech Communication* are gratefully acknowledged for many helpful suggestions and constructive criticism.

### References

- Badin, P., Baricchi, E., Vilain, A., 1997. Determining tongue articulation: from discrete fleshpoints to continuous shadow. In: *Proceedings of Eurospeech97*, Vol. 1, pp. 47–50.
- Badin, P., Bailly, G., Raybaudi, M., Segebarth, C., 1998. A three-dimensional linear articulatory model based on MRI data. In: *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 249–254.
- Badin, P., Borel, P., Bailly, G., Révère, L., Baciú, M., Segebarth, C., 2000. Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images. In: *Proceedings of the 5th Seminar on Speech Production: Models and Data*, pp. 261–264.
- Beautemps, D., Badin, P., Bailly, G., 2001. Degrees of freedom in speech production: analysis of cineradio- and labio-films data for a reference subject, and articulatory-acoustic modeling. *J. Acoust. Soc. Am.*, 2165–2180.
- Beskow, J., 1995. Rule-based visual speech synthesis. In: *Proceedings of Eurospeech95*, pp. 299–302.
- Branderud, P., 1985. Movetrack—a movement tracking system. In: *Proceedings of the French-Swedish Symposium on Speech*, Grenoble, pp. 113–122.
- Carlson, R., Granström, B., Hunnicut, S., 1982. A multi-language text-to-speech module. In: *Proceedings of ICASSP—Paris*, Vol. 3, pp. 1604–1607.
- Cohen, M., Massaro, D., 1993. Modeling coarticulation in synthetic visual speech. In: *Thalman, N.M., Thalman, D. (Eds.), Models and Techniques in Computer Animation*. Springer-Verlag, Tokyo.
- Coker, C., Fujimura, O., 1966. Model for the specification of the vocal tract area function. *J. Acoust. Soc. Am.* 40, 1271.
- Cohen, M., Beskow, J., Massaro, D., 1998. Recent development in facial animation: an inside view. In: *Proceedings of AVSP98*, pp. 201–206.
- Dang, J., Honda, K., 1998. Speech production of vowel sequences using a physiological articulatory model. In: *Proceedings of ICSLP98*, Vol. 5, pp. 1767–1770.
- Dang, J., Honda, K., 2000. Improvement of a physiological articulatory model for synthesis of vowel sequences. In: *Proceedings of ICSLP2000*, Vol. 1, pp. 457–460.
- Engwall, O., 1999. Modeling of the vocal tract in three dimensions. In: *Proceedings of Eurospeech99*, pp. 113–116.
- Engwall, O., 2000a. Are static MRI data representative of dynamic speech? Results from a comparative study using MRI, EMA and EPG. In: *Proceedings of ICSLP2000*, Vol. 1, pp. 17–20.
- Engwall, O., 2000b. Dynamical aspects of coarticulation in Swedish fricatives—a combined EMA & EPG study. *TMH-Quarterly Status and Progress Report*, Vol. 4, KTH, Stockholm, pp. 49–73.
- Engwall, O., 2001. Synthesizing static vowels and dynamic sounds using a 3D vocal tract model. In: *Proceedings of 4th ISCA Tutorial and research workshop on Speech synthesis*.
- Engwall, O., 2002a. Evaluations of a system for concatenative articulatory visual synthesis. In: *Proceedings of ICSLP*.
- Engwall, O., 2000b. Tongue talking—studies in intraoral visual speech synthesis. Ph.D. thesis, KTH, Stockholm, Sweden.
- Engwall, O., Badin, P., 1999. Collecting and analysing two- and three-dimensional MRI data for Swedish. *TMH-Quarterly Status and Progress Report*, Vol. 3–4, KTH, Stockholm, pp. 11–38.
- Engwall, O., Badin, P., 2000. An MRI study of Swedish fricatives: coarticulatory effects. In: *Proceedings of the 5th Seminar on Speech Production: Models and Data*, pp. 297–300.
- Fant, G., 1965. Formants and cavities. In: *Proceedings of ICPHS'65*, pp. 120–140.
- Fujimura, O., 1990. Methods and goals of speech production research. *Lang. Speech* 33 (3), 195–258.
- Fujimura, O., Kakita, Y., 1979. Remarks on quantitative description of the lingual articulation. In: *Öhman, S., Lindblom, B. (Eds.), Frontiers of Speech Communication*. Academic Press, London, pp. 17–24.
- Harshman, R.A., Ladefoged, P., Goldstein, L., 1977. Factor analysis of tongue shapes. *J. Acoust. Soc. Am.* 62, 693–707.

- Honda, K., Hirai, H., Dang, J., 1994. A physiological model of speech production and the implication of tongue larynx interaction. In: *Proceedings of ICSLP94*, pp. 175–178.
- Jones, W., 1977. Electropalatograph hardware. *Phonetics Laboratory of University of Reading, Work in Progress, Vol. 1*, pp. 7–13.
- Kiritani, S., Miyawaki, K., Fujimura, O., 1976. A computational model of the tongue. *Res. Instit. Logoped. Phoniatr. Annual Bulletin, Vol. 10*, University of Tokyo, pp. 243–252.
- Krakow, R., 1999. Physiological organization of syllables, a review. *J. Phonetics* 27, 23–54.
- Le Goff, B., Benoît, C., 1996. A text-to-audiovisual speech synthesizer for French. In: *Proceedings of ICSLP96*, pp. 2163–2166.
- Lindblom, B., Sundberg, J., 1971. Acoustical consequences of lip, tongue and jaw movements. *J. Acoust. Soc. Am.* 50, 1166–1179.
- Madea, S., 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In: *Hardcastle, W., Marchal, A. (Eds.), Speech Production and Modelling*. Kluwer Academic Publishers, pp. 131–149.
- Maeda, S., 1988. Improved articulatory models. *J. Acoust. Soc. Am.* 84, S146.
- Marchal, A., Espesser, R., 1987. L'asymétrie des appuis linguo-palatins. In: *Communications aux 16e journées d'études sur la parole*, Hammamet.
- Masuko, T., Kobayashi, T., Tamura, M., Masubuchi, J., Tokuda, K., 1998. Text-to-visual speech synthesis based on parameter generation from HMM. In: *Proceedings of ICASSP98*, pp. 3745–3748.
- Mermelstein, P., 1973. Articulatory model for the study of speech production. *J. Acoust. Soc. Am.* 53, 1070–1082.
- Nguyen, N., 2000. A Matlab toolbox for the analysis of articulatory data in the production of speech. *Behav. Res. Meth. Instrum. Comput.* 32, 464–467.
- Payan, Y., Perrier, P., 1997. Synthesis of V–V sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech Commun.* 22, 185–205.
- Payan, Y., Perrier, P., Laboisière, R., 1995. Simulation of tongue shape variations in the sagittal plane based on a control by the equilibrium-point hypothesis. In: *Proceedings of ICPhS95, Vol. 2*, pp. 474–477.
- Pelachaud, C., 1991. Communication and coarticulation in facial animation. Ph.D. thesis, University of Pennsylvania.
- Perkell, J., 1974. A physiologically-oriented model of tongue activity in speech production. Ph.D. thesis, MIT, Cambridge, MA.
- Perrier, P., Perkell, J., Payan, Y., Zandipour, M., Guenther, F., Khalighi, A., 2000. Degrees of freedom of tongue movement in speech may be constrained by biomechanics. In: *Proceedings of ICSLP2000*, pp. 162–165.
- Rubin, P., Baer, T., Mermelstein, P., 1981. An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am.* 70, 321–328.
- Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., Browman, C., 1996. CASY and extensions to the task-dynamic model. In: *Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Producing Modeling—4th Speech Production Seminar*, pp. 125–128.
- Schwartz, J.-L., Boë, L.-J., 2000. Predicting palatal contacts from jaw and tongue commands: a new sensory model and its potential use in speech control. In: *Proceedings of the 5th Speech Production Seminar: Models and data*, pp. 257–260.
- Stone, M., 1990. A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data. *J. Acoust. Soc. Am.* 87, 2207–2217.
- Stone, M., Lundberg, A., 1996. Three-dimensional tongue surface shapes of English consonants and vowels. *J. Acoust. Soc. Am.* 99, 3728–3737.
- Tiede, M., Masaki, S., Wakumoto, M., Vatikiotis-Bateson, E., 1997. Magnetometer observation of articulation in sitting and supine conditions. *J. Acoust. Soc. Am.* 2, 3166.
- Tiede, M., Masaki, S., Vatikiotis-Bateson, E., 2000. Contrasts in speech articulation observed in sitting and supine condition. In: *Proceedings of the 5th Speech Production Seminar: Models and data*, pp. 25–28.
- Wilhelms-Tricarico, R., 1997. Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *J. Acoust. Soc. Am.* 97, 3085–3098.
- Wilhelms-Tricarico, R., 2000. Development of a tongue and mouth floor model for normalization and biomechanical modeling. In: *Proceedings of the 5th Speech Production Seminar: Models and data*, pp. 141–144.
- Wilhelms-Tricarico, R., Wu, C.-M., 1997. Mapping of muscle anatomy on 3-d mr images of the human tongue based on morphological landmark selection. *J. Acoust. Soc. Am.* 102 (5, Pt 2), 3163.
- Wrench, A., Hardcastle, W., 2000. A multichannel articulatory speech database and its application for automatic speech recognition. In: *Proceedings of the 5th Seminar on Speech Production: Models and Data*, pp. 305–308.
- Wrench, A., McIntosh, A., Hardcastle, W., 1998. Optopalatograph: real-time feedback of tongue movement in 3D. In: *Proceedings of ICSLP98*, pp. 1867–1870.