Speaker adaptation of a three-dimensional tongue model

Olov Engwall

Centre for Speech Technology, KTH, Stockholm, Sweden

10.21437/Interspeech.2004-196

Abstract

Magnetic Resonance Images of nine subjects have been collected to determine scaling factors that can adapt a 3D tongue model to new subjects. The aim is to define few and simple measures that will allow for an automatic, but accurate, scaling of the model. The scaling should be automatic in order to be useful in an application for articulation training, in which the model must replicate the user’s articulators without involving the user in a complicated speaker adaptation. It should further be accurate enough to allow for correct acoustic-to-articulatory inversion. The evaluation shows that the defined scaling technique is able to estimate a tongue shape that was not included in the training with an accuracy of 1.5 mm in the midsagittal plane and 1.7 mm for the whole 3D tongue, based on four articulatory measures.

1. Introduction

1.1. ARTUR - the ARticulation TUtoR

We are planning to use the KTH 3D face and vocal tract models to create an automatic articulation tutor, who can help its users, e.g. hearing-impaired children, second language learners or speech therapy patients, to correct deviant articulations. The goal is to provide relevant visual feedback on the place and manner of articulation by contrasting the user’s own articulation with a correct one. This involves visuo-acoustic-to-articulatory inversion (i.e. the articulation should be determined from the speech signal and constraints imposed by computer vision analysis of facial features in video images) which requires that the articulatory model is scaled to correspond to the user. The application puts two, somewhat conflicting, conditions on the scaling. Firstly that it has to be accurate enough for the articulatory inversion to be successful, secondly that it should be possible to do automatically from images of the face, so that the user does not need to be concerned with a time consuming speaker adaptation. We will use a database of images of the face and 3D Magnetic Resonance Images (MRI) of the vocal tract to establish relations between facial and vocal tract features, allowing us to do the scaling from facial images.

The scaling of the 3D tongue model is however a non-trivial problem in itself (i.e. even if the scaling is based on images of the vocal tract rather than of the face) as the second constraint implies that the articulatory scaling factors should be easily established automatically (i.e. as few factors as possible should be used and the ones used should not require manual alignment or measurements) and that the same type of scaling should be valid for all users (i.e. the factors should capture the universal dimensions along which the individual variations occur). Moreover, speaker adaptation of articulatory models has certainly been performed previously, but for two-dimensional models of the midsagittal plane [1, 2, 3, 4], whereas our application requires three-dimensional adaptation. This paper therefore presents, and then evaluates, scaling factors suited to adapt a 3D tongue model, based on one reference subject, to other speakers, whose vocal tract dimensions were determined with MRI. Even if the scaling here is based directly on MR images, the aim is to define factors that have the potential to be set from facial data only.

1.2. Speaker adaptation of articulatory models

Adaptation of a vocal tract model to another speaker can be made at many different levels of detail, from normalization of the entire vocal tract length, over ratio scaling of two tube models, to more detailed scaling of different articulators.

Mathieu & Laprie [1] adapted Meada’s two-dimensional articulatory model [5] to a new speaker using one sagittal MR image each of eleven French vowels uttered by the speaker. Meada’s model includes two scale factors, which modify the longitudinal dimension of the vocal tract contours in the pharynx and oral cavity, respectively, with an interpolation in the zone in-between. The two factors were set manually to achieve the best overall fit, based on an overall subjective correspondence, between the model and the MR images. The exterior vocal tract contour was then set to the mean exterior contour in the MR images. Ouni & Laprie [6] later instead determined the two scale factors from formant measures.

Boë et al. [2] also used the Maeda model and applied variable scale factors for the pharynx and the oral cavity to simulate vocal tract growth in infants. Goldstein [3] had the same goal of studying children’s vocal tracts, but instead set the geometrical variables in the Mermelstein model [7] based on data from the medical literature. Goldstein notably defined factors establishing the radius of the tongue body arc, the length of the line between the end points of the tongue blade arc and the tip and the distance between the two arcs. Goldstein’s approach is closer to the one proposed here in that it is able to alter not only the tongue size, but the shape...
2. The 3D tongue model and subjects

The tongue model is based on a 3D MRI database of one 27-year-old male reference subject of Swedish [9]. The corpus included 13 Swedish vowels in isolation and 10 consonants in three symmetric VCV contexts with V=[a, i, o].

Nine new subjects, five male and four female, were then imaged while producing a neutral vowel [7]. The subjects were between 25 and 48 years of age and of varying body mass and in order to avoid unwanted head movements, the head was fixed with solid foam cushions.

A 1.5 T MRI system with fast gradients and a quadrature neck coil was used in the acquisition. The slice thickness and the interslice center-to-center distance were both 4 mm. The resolution was either 0.976563 mm/pixel (two subjects), 1.0549 (in the coronal images for eight subjects), or 2.10938 mm/pixel (in the sagittal images for three subjects).

The scaling is based on four independent factors that influence either different parts of the tongue or scales it in different directions using a linear estimation. The four factors are: tongue Dorsum Size (tDS), tongue Body Size (tBS), tongue Length (tL) and Tongue Width (tW). The effect of each scale factor sf can be written as:

\[
\text{Length (tL)} = AM \cdot \text{tL} + AM_{DS} \cdot \text{tDS} + AM_{BS} \cdot \text{tBS} + ... \\
\text{Width (tW)} = AM_{tL} \cdot \text{tL} + AM_{tW} \cdot \text{tW}
\]  

Thus, once the weight vectors \( \mathbf{w}_{sf} \) have been determined, an unknown subject tongue can be replicated using only the articulatory measures.

As all scale factors are linearly independent, each weight vector can be determined independently as:

\[
\mathbf{w}_{sf} = \frac{AM_{sf}}{\text{Length (tL)}} (S_{sf} - \mathbf{R}_{sf})
\]
where $\mathbf{AM}_{sf}$ is the vector with the articulatory measure for $s_f$ for each subject, $\mathbf{S}_{sf}$ is the matrix containing all the subjects’ tongue vertex points in the region of influence and $\mathbf{RS}_{sf}$ is the matrix with the vertex points of the reference tongue in the same region.

The three factors $t_{DS}$, $t_{BS}$ and $t_{L}$ use one articulatory measure each in the midsagittal plane, whereas $t_{W}$ is based on a width measure, cf. Fig. 2. The three first factors can hence be used for midsagittal scaling of the model, but they also influence the tongue shape outside the midsagittal plane. It can be noted that the length factor $t_{L}$ is similar to the “blade” scaling in [3] and that two factors ($t_{BS}$ and $t_{DS}$) had to be used for the tongue body, since a single scale factor was not able to reconstruct the variation in the subjects’ tongue shape shown in Fig. 2 – the size in the two regions are not directly coupled and scaling the radius of a circular arc, as in the Mermelstein model [7] is not suitable for the subjects of this study (no one scaling factor can describe the difference between e.g. subjects 1, 2 and 4 in Fig. 2, who have a large deviation from the reference tongue in one region each).

The tongue model was constructed based on a semi-polar grid, and this same grid is used to determine the scale factors, cf Fig. 2. One factor, $t_{DS}$, is used to scale the tongue at gridlines 1-11, and another, $t_{BS}$, to scale the tongue at gridlines 13-20, with a blending between the two at gridline 12.

The weights $\overline{w}_{t_{DS}}$, $\overline{w}_{t_{BS}}$ and $\overline{w}_{t_{L}}$ are set based on the reconstruction from the sagittal images and each factors is determined by an articulatory measure to determine the size in the relevant region. $AM_{t_{DS}}$ is the mean distance from the grid centre to the tongue dorsum (at gridlines 9-11), $AM_{t_{BS}}$ is the mean distance from the grid centre to the tongue blade (at gridline 14-16) and $AM_{t_{L}}$ is the distance from gridline 16 to the tongue tip, cf. Fig. 2a). The mean, rather than the maximum, deviation was chosen for $AM_{t_{DS}}$ and $AM_{t_{BS}}$, as it was judged that a mean articulatory value will be easier to estimate from images of the face.

Using the scale factor weights, determined from Eq. 2 and the articulatory measures above, the subjects’ tongue shapes can be replicated. Fig. 3 shows the midsagittal contours of three of the subjects. Note that the tongue contour was in each case estimated based on scale factor weights trained on the other subjects, meaning that the reconstruction is made with only the three articulatory measures as input information on the subject.

The distribution of the midsagittal error in Fig. 4 further shows that while the error is larger in the upper pharynx it is low in the oral cavity, which is positive as the articulation tutor is more dependent on a high accuracy in the oral cavity where the places of articulation lie closer.

The weight $\overline{w}_{t_{W}}$ is set based on the reconstruction from the coronal images, with the weight factor $t_{W}$ determined by the distance between the left and right side of the tongue at gridline 14, cf. Fig. 2b). Fig. 3 shows the actual and estimated coronal contours of three of the subjects.

5. Evaluation

The reconstruction was evaluated using the root mean squared (RMS) error between the vertices in the scaled model and the surface spanned by the contours extracted from the MR images. The reconstruction error was calculated for the the scaled model compared to the midsagittal contour, the sagittal images and the coronal images. In each case, the error was defined as the three-dimensional Euclidean distance between the model’s vertices and the tongue shape measured in the images. The sagittal error measures hence also includes lateral differences between the goal tongue and the model whereas the coronal also includes differences in the front-back dimension. The two errors are however calculated over partially different vertices, since the sagittal and coronal images have different regions of good accuracy, as discussed in section 3. The sagittal error is hence calculated over the vertices on the tongue surface, excluding those on the lateral edges and the coronal error is calculated over all vertices in front of gridline 11 (cf. Fig. 2).

The midsagittal RMS error in the reconstruction (i.e. if the subject tongue was included in the training) was 1.3 mm, the sagittal 1.7 mm and the coronal 1.7 mm. This is a higher level of error than the reconstruction of tongue shapes for the reference subject (for whom the sagittal error was 1.3 mm and the lateral 1.2 mm). Considering that the image resolution was lower for the new subjects and that much fewer configurations (9 subjects compared to 43 articulations) were used in the training, the error however seems acceptable. Table 1 shows that the estimation of the tongue shape for a new subject is not worse if the subject tongue was not included in the training of the scale coefficients. This is crucial, as it means that unknown tongue shapes can be estimated, with the same level of accuracy, based on scale coefficients established earlier with a training database and the four artic-
ulatory measures. Table 1 also shows that the results of the estimation differs substantially between subjects, with larger errors for some. As the scaling is to be used in visual resynthesis of the subjects tongue movements this level of error might nevertheless be quite acceptable if the overall resemblance with the subjects tongue shape is good, as illustrated by Fig. 3. It can also be noted that the correlation between the midsagittal and the sagittal error is very weak (0.20); a good estimation of the midsagittal contour is no guarantee for a good overall resemblance with the tongue shape.

There was no difference in the scaling accuracy between men (subjects 1-5) and women (6-9), indicating that the same scale factors can be used regardless of sex. To test this conclusion further, the men’s and women’s tongues were also scaled based on the sub-databases of only the other men and women, respectively, again giving the same level of accuracy.

6. Conclusions & future work

This study shows that once a set of scaling weights have been determined from a training database, the tongue shape of a new subject can be estimated using one lateral and three sagittal measures. The next step is to determine whether these measures can be estimated from facial data, so that the model tongue can be adapted based on the subject’s face.

Future work using the scaled model for acoustic-to-articulatory inversion will show if the accuracy achieved with these scale factors is high enough to determine the place of articulation. As a comparison, Hiroya & Honda [10] achieved a midsagittal RMS error of 1.65 mm on the horizontal and vertical positions of eight EMA coils estimated from speech acoustics of an unknown speaker using an HMM-based model of the link between acoustic and articulatory data.

7. Acknowledgements

This research was carried out at the Laboratoire de Phonologie, Université Libre de Bruxelles (ULB), during the author’s post-doctoral stay funded by the Wenner-Gren foundation. All MRI data was collected at the Unité de Résonance Magnétique, Hôpital Erasme, ULB, following an invitation from Didier Demolin. The data of subject 9 was collected separately as part of a National Institute of Health grant (R0 DC022114) to Harvey Sussman, University of Texas at Austin, and Björn Lindblom, Stockholm University for articulatory studies and development of the APEX speech production model. The experimental skills and experience of Thierry Metens, Vincent Denolin, Didier Demolin and Alain Soquet in gathering the data are gratefully acknowledged.

8. References


<table>
<thead>
<tr>
<th>Subject</th>
<th>Total</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midsag.</td>
<td>1.5</td>
<td>0.8</td>
<td>1.7</td>
<td>2.2</td>
<td>1.4</td>
<td>1.2</td>
<td>1.9</td>
<td>1.0</td>
<td>1.6</td>
<td>1.2</td>
</tr>
<tr>
<td>Sagittal</td>
<td>1.7</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>1.2</td>
<td>1.5</td>
<td>1.6</td>
<td>1.5</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Coronal</td>
<td>1.7</td>
<td>1.8</td>
<td>1.6</td>
<td>1.5</td>
<td>1.8</td>
<td>2.0</td>
<td>1.4</td>
<td>1.5</td>
<td>1.7</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 1: RMS error in mm when estimating the tongue shape of a subject based on four articulatory measures.

Figure 4: The midsagittal reconstruction error as a function of gridline.