From real-time MRI to 3D tongue movements

Olov Engwall

Centre for Speech Technology, KTH, Stockholm, Sweden
olov@speech.kth.se

Abstract

Real-time Magnetic Resonance Imaging (MRI) at 9 images/s of the midsagittal plane is used as input to a three-dimensional tongue model, previously generated based on sustained articulations imaged with static MRI. The aim is two-fold, firstly to use articulatory inversion to extrapole the midsagittal tongue movements to three-dimensional movements, secondly to determine the accuracy of the tongue model in replicating the real-time midsagittal tongue shapes. The evaluation of the inversion shows that the real-time midsagittal contour is reproduced with acceptable accuracy. This means that the 3D model can be used to represent real-time articulations, even though the artificially sustained articulations on which it was based were hyperarticulated and had a backward displacement of the tongue.

1. Introduction

Magnetic Resonance Imaging (MRI) is well suited to study the three-dimensional shape of the vocal tract and tongue during sustained articulations. However, as the acquisition times have been very long, MRI has firstly not been able to provide data on tongue movements, and has secondly required the subject to sustain the articulation for a prolonged time, typically over 30 seconds. The artificial sustaining may cause the articulations to be both hyperarticulated compared to a normally sustained articulation and more difficult to hold for the subject, inducing a backward movement of the tongue in the pharynx [1].

Real-time data is thus needed, in order both to provide data on articulator movements and to assure that the sustained articulations correspond to, or can be adapted to, real-time articulations. The real-time data can be either continuous or pointwise, necessitating different strategies to generate 3D tongue movements.

1.1. Pointwise real-time data

Electromagnetic articulography (EMA) or X-ray microbeam provide an easy method to control the real-time movements and shape of a tongue model [2, 3]. However, in themselves they only provide information on the oral part of the tongue, where the measurement coils are attached, whereas the deviation between real-time and sustained articulations noted in [1] occurs in the pharynx.

Pointwise data can still be used to define the entire tongue contour if it is combined with a reliable articulatory inversion method, e.g. [2, 4, 5], in which the model’s tongue shape and position are optimised to be in accordance with the data on coil positions, to remain articulatorily possible and assuring that the tongue volume is held constant.

Such an inversion method provide good results for the tongue shape in the oral part, but because of the lack of measurement points in the pharynx it can not guarantee that a correct pharyngeal tongue shape is achieved. Moreover, it is not even certain that the correct tongue shape can be achieved with a model based on artificially sustained articulations. In [6] it was noted that when real-time EMA data of a set of sentences was used to control a tongue model based on static 3D MRI data, the tongue root was much less mobile than in corresponding X-ray movies of the same sentences. As the EMA data could not give direct information on the tongue movement in the pharynx, the pharyngeal tongue movements were determined based on the correlation between oral and pharyngeal tongue shape in the MRI data. Since the tongue was displaced backwards for all articulations in the MRI corpus [1] the result was unnaturally small tongue root movements when they were generated from the combination of EMA coil movement in the oral cavity and the above mentioned correlation.

1.2. Continuous real-time data

The alternative to extrapolating the midsagittal contour from a few measurement points is to acquire real-time data with a method that captures the entire tongue contour, e.g. X-rays, ultrasound or real-time MRI. X-rays has the advantage that the time resolution is high, but ultrasound or real-time MRI is nevertheless preferable, as the speaker is not subjected to harmful radiation.

The acquisition frequency of real-time MRI depends firstly on a compromise between time and spatial resolution: Demolin et al. [7] collected 5 images/s with a thickness of 6 mm (spatial resolution: 3.9 mm × 1.95 mm) and Mady et al. [8] 8 images/s with a thickness of 10 mm. Secondly, it fortunately changes with improvements of MRI scanners and acquisition set-up, which means that the possible time resolution increases continuously.

In this paper midsagittal MR images collected at 9 images/s are used to control a 3D tongue model by means of an
inversion method that determines the articulatory parameters needed in order to extrapolate the midsagittal contour to a 3D tongue shape.

2. Model and data acquisition

2.1. 3D MRI data and tongue model

The 3D tongue model [3], shown in Fig. 1(a), is based on a three-dimensional MRI database of one male reference speaker of Swedish. The corpus included the 13 Swedish vowels [i, y, o, e, a; a, u, ã, e, æ, æ, ø, ø] in isolation and the 10 consonants [p, t, k, l, m, s, f, c, g, j] in three symmetric VCV contexts with V=[a, i, o]. Due to the acquisition time of 43 seconds to scan the 54 slices of each articulation, all phonemes were artificially sustained, and the VC and CV transitions were made immediately prior to and after the scan, respectively.

Using a statistical analysis the articulatory parameters jaw height (JH), tongue dorsum raise (TD), body raise (TB), tip raise (TT), tip advance (TA) and tongue width (TW) were defined. The parameters, with the exception of TW, are defined in the midsagittal plane, but influence the entire 3D shape through a weighted deformation of the model’s polygon mesh. The weights were set based on the relationship between the midsagittal contour and the position of every vertex point in the polygon mesh for all the articulations in the MRI corpus. Midsagittal data, such as from real-time MRI, can consequently be used to resynthesize 3D articulations.

2.2. Real-time MRI data

For the real-time data the same reference subject produced the vowel sequence [a e i u y o æ ø] and the consonants C=[p, t, k, l, m, s, f, c, g, j] in a sequence of three VCV words each: [aCa tCt oCt]. The data was registered at 9 images/s with a TSE Zoom sequence in which a single 6 mm thick midsagittal T1-weighted image, cf. Fig. 1(b), was acquired. The images are 128x128 pixels with a spacial resolution of 2.34375 mm. A quadrature neck coil was used as receiver coil and the scanner was a 1.5 T MRI system with fast gradients (CompactPlus, PowerTrak 6000, 20 mT/m and 100 mT/m ms⁻¹ maximum amplitude and slew rate, Philips Medical Systems, Best, The Netherlands) [7].

The tongue contour was extracted in each image using an automatic boundary detection, based on thresholding of the image, followed by manual correction of the spline control points found in the boundary detection. The contour was then resampled to have the same number of equally spaced points for each image. In each image, the jaw was further positioned manually. In total, 910 tongue shape configurations were extracted and subjected to the articulatory inversion described in section 3.

The palatal contour was used firstly to ensure that the subject maintained the same orientation of the head and secondly to roto-translate the extracted contours to the reference frame of the 3D model.

Figure 1: The tongue and jaw model in the frame of a synthetic face and a real-time midsagittal MR image.

3. Inversion procedure

The articulatory parameters of the tongue model, and hence the tongue shape, are recovered from the midsagittal contours of the tongue and jaw using an inversion procedure. The method is similar to the one in [5], where the midsagittal contour of the tongue was reconstructed from three EMA coils on the tongue and one on the lower incisor. In this paper, the contour is determined from 20 midsagittal points and the position of the maxilla.

The parameter JH, controlling the jaw height, is estimated directly from the jaw position as the normalised value of the vertical distance between the upper and the lower incisor. The parameters TB, TD, TT and TA were determined in two steps, the first step providing an approximative fit quickly, hence speeding up the optimisation in the second step. The rough fit between the model and the measured midsagittal contour was achieved by successively setting the articulatory parameters to values that bring the model closer to the goal contour. After removing the influence of JH, TB was first increased (or decreased) as long as the model’s tongue blade was lower (higher) than the tongue blade of the goal contour. TD, TT and TA were then successively set, in the same manner as TB, to approximatively reach the goal contour in the dorsum and tip regions, respectively. This successive parameter adjustment provides a rather good approximative correspondence with the goal contour, cf. Fig. 2, as TB and TD were defined in the model as the most important components of the tongue body after the contribution of JH has been removed, TT is the principal variable of the tongue tip after the influence of JH, TB and TD has been removed and the Tongue Advance parameter TA controls the remaining variability when the contribution of JH, TB, TD and TT has been removed.

A better fit is then achieved by simultaneous adjusting the parameters using the Matlab® optimisation function fgoalattain, to minimise the Euclidean distance be-
Figure 2: The two step fitting of the model’s midsagittal contour to the tongue in the MR image of [s] (solid), using first a successive parameter adjustment (dotted) and then a simultaneous optimisation (dashed).

between the tongue model’s midsagittal contour and the midsagittal contour in the MR images.

Once the articulatory parameters are set based on the midsagittal contours, 3D tongue movements can be resynthesised with the model. The resulting animations of the 3D tongue and the corresponding real-time MRI movies can be found on http://www.speech.kth.se/multimodal/mri.

4. Results

The inversion was evaluated through an objective global measure and an analysis of articulatory details.

4.1. Global achievement

Table 1 lists the root mean squared (RMS) error for the entire tongue, separate parts and different sub-corpora. The total RMS error of 0.14 cm can be compared to the sagittal RMS error of 0.13 cm for the 3D model’s reconstruction of the articulations in the corpus [3], and the midsagittal RMS error of 0.09 cm, 0.11 cm and 0.17 cm attained by Badin et al. [9] when their midsagittal articulatory model was based on 1222, 20 and 8 configurations extracted from X-ray images. The total RMS error attained here is higher than in [9] but the accuracy is nevertheless quite acceptable considering that it is lower than the spatial resolution in the real-time MR images and of about the same level as the accuracy of the 3D model and further that the resynthesis is made with a 3D model based on 43 configurations from one occasion and real-time data from another, whereas the reconstruction in [9] was made with a model based on carefully chosen midsagittal images collected at the same occasion as the configurations to replicate. A flaw in the reconstruction is that whereas it is good for the blade and tip, it is much less successful for the dorsum and pharyngeal parts. Fig. 3 however shows that the maximal reconstruction error occurring in the upper pharynx immediately below the velum corresponds to a maximum for the remaining standard deviation in the mid-

sagittal tongue contour for the 3D corpus in [3]. This indicates that there is no qualitative difference between the reconstruction of the real-time tongue contours and that of the static 3D shapes (at least in the midsagittal plane) and the error is thus not an effect of incompatibility between artificially sustained and real-time articulations.

4.2. Articulatory analysis

The correspondence between the model and the measured tongue shape was further investigated based on articulatory features. Over all, the correspondence was rather good, with the most striking differences being deviations in the tongue root position.

Eventhough the model’s tongue root follows the real tongue movements much better than when its movements were generated from EMA data of oral tongue movements [6], the lowest midsagittal point on the tongue model is still much less mobile than in the measurements, cf. Fig. 2. This artefact is due to the fact that when the tongue root was displaced backward in the static articulation it came in contact with the epiglottis. The movement of the tongue root was hence hidden, introducing the false notion that the lowest tongue points do not move at all into the statistical modelling.

Table 1 indicates that the error is much lower for vowels than for the VCVs, and this is evident also when the tongue contour is considered, with the tongue shape for the vowels being more neutral and therefore easier to replicate. The same phenomenon occurs for the different types of VCVs:

Table 1: RMS error in mm for the entire tongue and different parts.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Vowels</th>
<th>VCVs</th>
<th>Fric.</th>
<th>Stops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>1.4</td>
<td>1.0</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Pharynx</td>
<td>1.6</td>
<td>1.1</td>
<td>1.7</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>Dorsum</td>
<td>1.7</td>
<td>1.3</td>
<td>1.7</td>
<td>1.7</td>
<td>1.8</td>
</tr>
<tr>
<td>Blade &amp; tip</td>
<td>1.0</td>
<td>0.8</td>
<td>1.0</td>
<td>1.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Figure 3: The RMS error along the midsagittal tongue contour for the resynthesis of real-time movements (solid) and the remaining standard deviation in the 3D model of static tongue positions (dashed) [3].
labials, for which the constriction is not achieved with the tongue ([p, m, f]); front consonants using the tongue blade ([t, n, s, c]) or the tip ([s, l, r] to achieve the constriction and back consonants ([k, ʃ]). Fig. 4 shows the difference in RMS error for the four different VCV groups and from this comparison it is clear that even though the pattern in the RMS error is the same for all groups, the error is substantially larger in the velar region for the velar consonant VCVs, and somewhat larger at the tongue tip for the VCVs with dental and alveolar consonants. The RMS error is hence increased at the place of articulation, where the tongue differs the most from a neutral tongue. Logically, the RMS errors for bilabials, having no lingual constriction, and vowels, having a lesser degree of constriction, are lower.

5. Conclusions

This paper shows that real-time midsagittal MR images taken at 9 images/s can be used to resynthesise three-dimensional tongue movements by the means of an inversion procedure to extract articulatory parameters. This is a quite unsurprising finding, as X-ray images have been employed in the same fashion previously by eg. Badin et al. [10] and the extrapolation from X-ray images and real-time MR images is qualitatively the same. However, as the use of X-rays for speech production research can be questioned ethically, real-time MRI provides a promising alternative, even if the acquisition frequency is still too low for realistic speech production modelling of articulatory movements.

Further, and more importantly, it shows that the real-time articulations can be achieved, with reasonable accuracy, by the model based on artificially sustained articulations, even though these were hyperarticulated and had the tongue displaced backwards in the pharynx. The 3D model can consequently be used for correct resynthesis of the tongue movements in running speech.

6. Acknowledgements

This research was carried out at Laboratoire de Phonologie, Université Libre de Bruxelles (ULB), during the author’s post-doctoral stay funded by the Wenner-Gren foundation. Thierry Metens, Unité de résonance magnétique, hôpital Erasme, ULB is gratefully acknowledged for the image acquisition.

7. References


