Using Linguopalatal Contact Patterns to Tune a 3D Tongue Model

Olov Engwall

Centre for Speech Technology (CTT), Kungliga Tekniska Högskolan (KTH), Sweden
olov@speech.kth.se

Abstract
The six articulatory parameters of a three-dimensional tongue model were adjusted to replicate linguopalatal contact patterns measured with Electropalatography (EPG). The tongue model is based on artificially sustained articulations measured with MRI and the EPG data provides one possibility to tune the parameters to dynamic speech. A 3D model was generated of the palate and the electrode distribution, allowing the synthetic contact patterns to be calculated. The tongue parameters were then adjusted to minimise the deviation from the natural contact patterns. Substantial reduction of the false and missing electrode contacts was made in the tuning and the synthetic linguopalatal contact pattern is shown to replicate the total characteristics of the natural patterns rather well. The remaining error is often due to lateral asymmetry or central-to-edge contact variations.

1. Introduction
A new 3D tongue model [1] was recently developed within the KTH 3D Vocal Tract project [2]. The model was based on Magnetic Resonance Imaging (MRI) of a reference subject producing 43 artificially sustained articulations of Swedish and was able to replicate the tongue shapes in the corpus with an overall RMS reconstruction error of 1.2 mm.

In a comparative study [3] of coarticulation measured by static MRI and combined dynamical electromagnetic articulography (EMA) and electropalatography (EPG), it was however found that the artificially sustained articulations were hyperarticulated. The conclusion of the study was hence that the static MRI data needed to be complemented with dynamical data, in order to generate a model representative of dynamic speech.

Following this conclusion, the parameters of the 3D tongue model were tuned in this study, using data on the linguopalatal contact, collected with EPG. The concept of using three-dimensional models of the tongue and palate to determine synthetic linguopalatal contact pattern that can be compared to natural EPG data has been proposed earlier in [4] and [5], but no actual training of 3D tongue models has yet been presented.

2. Data acquisition
The reference subject was the same, 28 year-old male native speaker of Swedish, used for the tongue model [1].

The corpus consisted of a subset of the MRI corpus [6], such that it included the 25 articulations with clear linguopalatal contact. The vowels /æ, i, y, u/ were acquired in isolation, whereas the consonants /s, c, ʃ, t, k, ɫ/ were collected in VCV context with V = /a, i, u/. The fricatives were acquired simultaneously with EMA data, allowing articulatory measures such as the jaw height and midsagittal tongue contour points to be determined at the same time, whereas the rest of the corpus was acquired with EPG only. The linguopalatal contact data was collected using a Reading 62 electrode EPG system (cf. [7] for details).

3. EPG analysis
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.

3.1. The 3D EPG palatal model
The palatal shape was generated as a three-dimensional reconstruction from MR Images of the subject’s dental cast [6]. 62 artificial EPG electrodes were placed on the synthetic palate according to their placement on the real EPG palate. The electrode coordinates were calculated by superposing a two-dimensional 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 two-dimensional representation.

The three-dimensional model of the EPG palate allows for displays of the contact patterns in 3D, and more importantly, it can be used to calculate synthetic contact patterns using the 3D tongue model.

3.1.1. Palate and teeth models for visual speech synthesis
New models of the palate and upper and lower teeth were generated for the KTH 3D Vocal Tract project [2] at the same time. The model of the hard structures was based on contours extracted from MR Images of the subject’s dental cast, as for the synthetic palate described above. However, in order to allow real-time display of the model, 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, that were 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. 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 model was made symmetrical, discarding the right part of the model and making it a reflection of the left.

This process simplified the teeth-palatal model to about 800 vertices and 1400 polygons, which is about twice the amount of vertices and polygons needed for the tongue model. The teeth-palate model is rendered using the KTH Visual Speech System [8], as shown in Fig. 1.

3.2. Determining the synthetic EPG patterns

The synthetic EPG patterns were calculated by checking if the point on the tongue surface closest to the synthetic electrode was touching (or even penetrated) the original, asymmetric, palate surface model.

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 synthetic palate. Fig. 2 gives an example of the divergence between the natural and the (tuned) synthetic data. The initial divergence and the type of error vary greatly over the corpus (cf. Fig. 3), not only between consonants, but also between contexts for the same consonant. There is a weak tendency of articulations with much linguopalatal contact having larger errors (e.g. /l/ has larger error than /K, /l/ and /stl/ more than /amal/). Some contextual influence on the amount and type of error was also found, with the main error being too many alveolar or palatal contacts for consonants in /la/ context, too few alveolar contacts in /la/ context and too few contacts in total for front consonants in /la/ context. The relation between closed vowel context and too little alveolar contact did not generalise to vowels, however. /i, /e/ both had high initial errors, but in /l/ it is due to too few alveolar contacts and in /la/ too many.

In total over all articulations, with maximally 1550 contacts, 370 missing and 230 false contacts were registered (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. The two types of error were considered as equally wrong and the following tuning consisted in minimising the sum of the two errors, rather than aiming for an equal error rate of missing and false contacts.

4. Parameter tuning

The tongue model has six articulatory parameters: jaw height (JH), tongue body (TB), tongue dorsum (TD), tongue tip (TT), tongue advance (TA) and tongue width (TY) that control the tongue shape using linear deformations (cf. [1] for details). For every articulation, the difference between the natural and synthetic EPG patterns was minimised with a combinatorial search varying each of the six parameters in the interval $par_{\text{JJ}} \leq 0.4$; $par_{\text{TT}} + 0.4$, where $par_{\text{JJ}}$ was the parameter value determined by the component analysis of the MRI data. The combination of parameter values closest to $par_{\text{JJ}}$ that resulted in the least deviation from the natural pattern was chosen as the new parameter values. Only the articulatory parameters were tuned; the weights that determine the parameter influence on each vertex in the mesh and the axes of activation were maintained from the MRI based model. The vertex weights thus remain common for all articulations and each articulation was tuned individually by varying its parameter values in successively smaller steps.
4.1. Modifying synthetic EPG patterns

In order to provide a means for manual tuning of the synthetic EPG patterns to replicate details of an articulation that was not necessarily handled by the automatic parameter tuning, e.g. the laminal contact in /l/, a console was developed for manual parameter changes using scrollbars. The interface includes the displays of the deviation in contact pattern (Fig. 2) and of the tongue and palate contours so that the influence of the parameter change on the tongue contour is reflected, hence permitting to adjust the parameters based on both the contact pattern and the tongue contour. The results presented here have not been tuned manually.

4.2. Results

Fig. 3 shows the correspondence between the natural and the modified synthetic patterns and the improvement from the initial synthetic patterns. The total number of deviating contacts is lowered substantially for all articulations but /ufol/. The total error over the whole corpus is more than halved (183 missing + 101 wrong = 284 vs. 370+230=600), with a mean error of 18%. The characteristics of the natural contact pattern are quite well replicated by the tuned model, if measured by the centre of gravity index, COG, displayed in Fig. 4, where \( COG = (8 \times R_1 + 7 \times R_2 + ... + R_7) / (R_1 + ... + R_7) \) and \( R_i \) denotes the number of contacts in row \( i \) [9]. Even if the total number of errors is as high as 18%, the overall contact patterns are thus nevertheless correct. This is explained by the fact that switching the correct, but different, values of two electrodes in the same row counts as two contact pattern errors, while the COG index is unaffected, as for example for /l/ in all vowel contexts. The improvement compared to the initial pattern is evidenced also in the COG index.

4.2.1. Remaining error

The distribution of the remaining error was investigated for common features. Fig. 5 shows large remaining errors for /l/, /s/, /i/, /y/ and /θ, uθ, oθ/. These phonemes have laminal alveolar contact with central free electrodes, and this is apparently hard to accomplish in the current model. The number of false contacts is inferior to the number of missing for all articulations but /asa, usu, oσu/ that have mainly false velar contacts. This is due to a general problem for articulations with a low tongue body and a high tongue tip, that have larger errors than articulations with either both high tongue body and tip (e.g. /i/u/) or only high tongue body (e.g. /θ, uθ/).

The most common error was lateral shifts, e.g. false right velar contacts (/h, usu, fθl/), missing central velar contacts (/tw, c, aθa, tσa, aθl/) and missing lateral alveolar contacts (/lθl, iθ/). The lateral asymmetry index, \( LAT = \left| (C_2^+ + ... + C_θ^+ + C_2^- + ... + C_θ^-) / (C_2^- + ... + C_θ^-) \right| \) [10], was hence calculated. The natural contact pattern was found to be shifted to the left (cf. Fig. 6), with mean(LAT)=0.21 and std(LAT)=0.08. The asymmetric tongue model was not able to replicate this lateral asymmetry fully, having mean(LAT)=0.15 and std(LAT)=0.18. This is due to the fact that only the activation of the parameters was tuned in this study, whereas lateral adjustments would call for changes to the weights applied to each vertex as well. As this would however also have an uncontrolled influence on articulations without any palatal contact, the weight adjustment was excluded in the tuning.
5. Conclusion

To the author’s knowledge this study is the first where the three-dimensional information in the EPG pattern has been exploited for modeling. The results show that parameter tuning can improve and adjust the palatal contact of the static model to replicate that observed in running speech more closely. It was further shown that the remaining error in the model is mainly lateral, such that the centre of gravity of the contact pattern is in accordance with the natural. The EPG patterns suggest two main weaknesses in the tongue model that should be considered: raising the tip independently of the tongue body and lateral variations.

Using EPG patterns as the basis for the evaluation of the tongue model, optimal parameter values can be determined for articulations with linguopalatal contact that were not included in the component analysis of the MRI data, i.e. voiced stops and fricatives.

A detection and correction algorithm for tongue points that penetrate the hard palate and teeth boundaries has recently been implemented to improve the naturalness of the contact between the tongue and the palate.

6. Acknowledgement

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. Elisabet Eir Cortes and Peder Livijn of the Department of Linguistics at Stockholm University assisted with the EPG data acquisition.

7. References