Lund University, Dept. of Linguistics
Working Papers 49 (2001), 30–33

Modeling the Talking Tongue

Olov Engwall
CTT, Centre for Speech Technology, KTH
olov@speech.kth.se

Abstract

Electromagnetic articulography (EMA) and Electropalatography (EPG) data has been used to adjust the shape and the movement of a three-dimensional tongue model to follow the articulations of the reference subject. The EMA data was used to control the dynamics of the parameters of a three-dimensional tongue model, whereas the EPG data was used to tune the six articulatory parameters to replicate the subject's linguopalatal contact patterns. A 3D model has been generated of the palate and the jaw and an algorithm for handling boundary collisions between the tongue and the inner structures has been implemented.

1 Introduction

A new 3D tongue model (Engwall 2000a) was recently developed within the KTH 3D Vocal Tract project (Engwall 1999). The model was based on Magnetic Resonance Imaging (MRI) of a reference subject producing 43 artificially sustained articulations of Swedish and it was able to replicate the tongue shapes in the corpus with an overall RMS reconstruction error of 1.2 mm.

The artificially sustained articulations were however found to be hyperarticulated (Engwall 2000b) and dynamical measurements with electromagnetic articulography (EMA) and electropalatography (EPG) are hence used to tune the model to the articulations used in running speech. This paper summarises the recent steps taken in order to make the model replicate the tongue movements of the reference subject.

2 Data acquisition

The reference subject was the same, 28 year-old male native speaker of Swedish, used for the tongue model and the corpus consisted of two subsets of the MRI corpus (Engwall & Badin 1999). The EPG corpus included the 25 articulations with clear linguopalatal contact: the vowels /eː, iː, yː, ɯː/ acquired in isolation and the consonants /s, ɕ, ʂ, ʃ, t, k, l/ collected in VCV context with V = /a, ɪ, ʊ/. The EMA corpus included only the five fricatives /f, s, ɕ, ʂ, ʃ/ in the same VCV context as above.

The linguopalatal contact data was collected using a Reading 62 electrode EPG system and the articulographic data using the Movetrack electromagnetic measurement system (Branderud 1985) developed at the Department of Linguistics, University of Stockholm. Six receiver coils were used in the acquisition, of which the four indicated in Figure 1 are of importance here. The coil on the lower incisor, J, measures the jaw height and the three coils T1-T3 the movement of different parts of the tongue. T1 monitors the tongue tip and T2-T3 the movement along the two main articulatory axes of the tongue (*front-back* and *flattening-arching*). The remaining two coils were used to measure the protrusion and to serve as reference to adjust for head movements (refer to Engwall 2000c for details on the EMA-EPG acquisition).

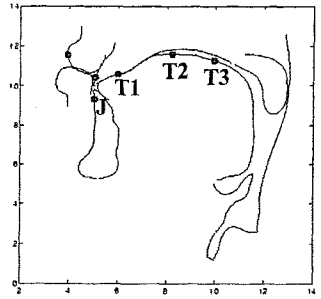


Figure 1. Data from the receiver coils at the jaw and the tongue surface were used to control the six articulatory parameters.

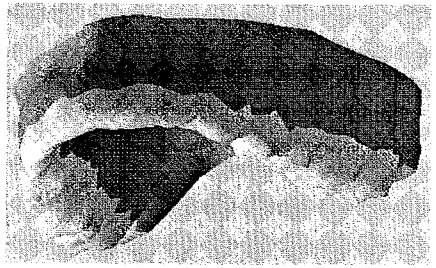


Figure 2. The dental casts were also used to create models of the palate and the jaw (not shown) for the KTH 3D Vocal Tract model.

3 EPG evaluation: using contact patterns for parameter tuning

A palate model was generated as a three-dimensional reconstruction from MR Images of the subject's dental cast (Engwall & Badin 1999). Firstly, it was used to create a new model (Figure 2) of the hard palate, teeth and upper gums for the vocal tract model (a model of the jaw, with lower teeth and gums, was created in the same way), rendered in the KTH visual speech synthesis environment (Beskow 1995). Secondly, 62 artificial EPG electrodes were placed on the synthetic palate according to their placement on the real EPG palate. This three-dimensional model of the EPG palate was used in combination with the 3D tongue model to calculate synthetic contact patterns.

Based on the difference between the natural and synthetic EPG patterns, the six parameters of the tongue model; jaw height (JH), tongue body (TB), tongue dorsum (TD), tongue tip (TT), tongue advance (TA) and tongue width (TY) were tuned. The tuning was done with a combinatory search varying each of the six parameters until the combination of parameter values that resulted in the least deviation from the natural pattern was found. Figure 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 /ʊʃʊ/. The total error over the whole corpus is more than halved (183 missing + 101 wrong = 284 vs. 370 + 230 = 600), with a reduction of the mean error from 38% to 18%. The characteristics of the natural contact pattern are moreover quite well replicated by the tuned model, if comparing the total contact in each row, rather than electrode for electrode.

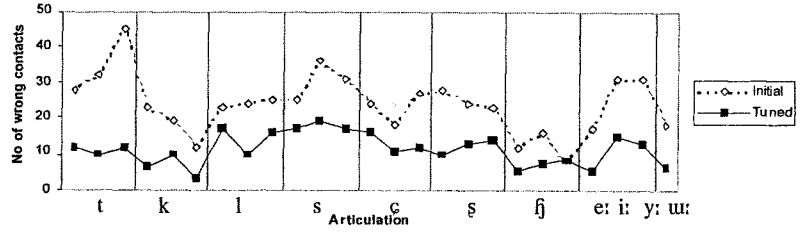


Figure 3. The total number of contact error (false and missing contacts) for the initial and t uned synthetic articulations. Context from left to right for the consonants: /a, ɪ, ʊ/.

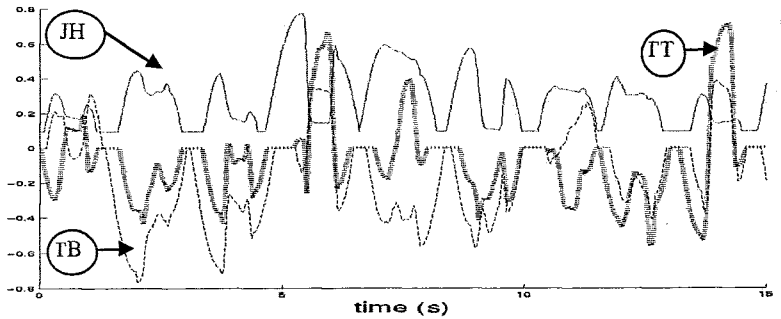


Figure 4. An idealized control sequence for the parameters JawHeight (solid, JH), Tongue Body (dashed, TB) and Tongue Tip (thick dotted, TT) in a series of VCV sequences.

4 EMA evaluation: using measures for dynamical parameter control

The EMA data from the four coils in Figure 1, was used so that J controls the jaw height (JH) and the three coils on the tongue, T1-T3, that of different parts of the tongue. The measurements of T1 control the raising (TT) and advancing (TA) of the tongue tip, those of T2 the tongue body (TB) and those of T3 the tongue dorsum movement (TD).

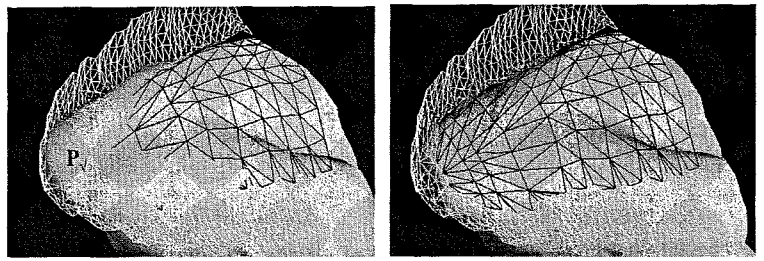
The EMA measures of the coil movements are given as the deviation in cm from the reference position, which is a neutral tongue position with upper and lower incisors touching. The parameters in the tongue model are however defined using the activation A ($-1<A<1$) of a prototype vertex movement towards a target vertex and a weight vector W_i, stating the influence of the parameter on all vertices i of the mesh. The EMA measures were hence transformed from distance to activation level of the corresponding parameter.

Based on the output from the four coils, idealized control sequences were generated, where the movements between the VCV sequences were faded out to a threshold value and where the transitions in the sequences were filtered in a recursive binomial smoothing, avoiding too abrupt changes. The result is control sequences of the type shown in Figure 4, that can be used to control the tongue model. Short movies of the talking tongue will be presented at the meeting and the movies are also available at the project page *http://www.speech.kth.se/multimodal/vocaltract.html*. The results of the EMA evaluation are taken into account as far as possible when incorporating the tongue model and its parameters into the KTH visual speech synthesis system (Beskow 1999; Carlson et al. 1982), including timing between articulator movement and parameter activation of the measured fricatives.

5 Handling boundary collisions

The combination of a dynamic tongue and models of the palate and the jaw, makes a method for handling contacts between surfaces necessary. A detection and correction algorithm has hence been introduced in the model to avoid impossible situations, where the tongue intersects 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 by placing these points on the boundary surface instead, similiar to the method proposed by Cohen et al. (1998). The maximal error in the correction is 1.4 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 outcome of the correction method is exemplified in Figure 5.



a) before correction *b) after correction*

Figure 5. The palate and tongue surface before and after correction for large activations of the tongue body and tip parameters. The points P_v violating the boundary condition are mapped onto the outline of the palate and teeth.

6 Discussion

The presented, dynamic 3D tongue model is an example of a successful combination of different measurement sources to replicate the speech production of one reference subject. The results of the EPG evaluation show that parameter tuning can improve and adjust the palatal contact of the static model to come closer to that observed in running speech. The remaining error in the EPG patterns suggests two main weaknesses in the tongue model: raising the tip independently of the tongue body and lateral variations.

7 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 EMA-EPG data acquisition.

References

Beskow, J. 1995. Rule-based visual speech synthesis. *Proc Eurospeech '95,* 299-302.

Branderud, P. 1985. Movetrack - a movement tracking system. *Proc of the French-Swedish Symposium on Speech,* Grenoble, 113-122.

Carlson, R. Granström, G., Hunnicut, S. 1982. A multi-language text-to-speech module. *Proc ICASSP-Paris*, vol. 3, 1604-1607.

Cohen, M., Beskow, J. and Massaro, D. 1998. Recent development in facial animation: an inside view. *Proc AVSP'98*, 201-206.

Engwall, O. 1999. Vocal tract modeling in 3D. *TMH-QPSR* 1-2, 31-38.

Engwall, O. 2000a. A 3D tongue model based on MRI data. *Proc ICSLP2000*, vol. III: 901-904.

Engwall, O. 2000b. Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA. *Proc ICSLP2000*, vol. I:17-20.

Engwall, O. 2000c. Dynamical aspects of coarticulation in Swedish fricatives: A combined EMA & EPG study. *TMH-QPSR* 4, 49-73.

Engwall, O., Badin, P. 1999. Collecting and analysing two- and three-dimensional MRI data for Swedish. *TMH-QPSR* 3-4: 11-38.