Biomechanical Tongue Models: An Approach to Studying Inter-speaker Variability

Ralf Winkler, Susanne Fuchs, Pascal Perrier, Mark Tiede

Biomechanical Tongue Models: An Approach to Studying Inter-speaker Variability

Ralf Winkler [1], *Susanne Fuchs* [1], *Pascal Perrier* [2], *Mark Tiede* [3, 4]

[1] ZAS, Berlin, Germany
[2] DPC/GIPSA-lab, Grenoble-INP, CNRS, Grenoble, France
[3] Haskins Labs, New Haven, CT, USA
[4] R.L.E.-MIT, Boston, MA, USA

winkler|fuchs@zas.gwz-berlin.de, Pascal.Perrier@gipsa-lab.grenoble-inp.fr, tiede@haskins.yale.edu

Abstract

Speakers of a given language vary with respect to their acoustics, articulation, and motor commands. This variation is driven by a variety of influences, such as emotional states, communicative interaction, and individual properties of the vocal tract. In this work we focus on the latter. First, we build speaker-specific biomechanical tongue models. Second, we discuss the impact of the relative position of the bending in the vocal tract on the basis of extensive simulations with two different models. We focus on /i,a,u/ by defining target regions in the acoustic space, and discuss the corresponding speaker-specific articulatory and motor command variability observed.
Index Terms: Biomechanical tongue models, speaker-specific articulation, variability

1. Introduction

Similar to ideas recently developed in the domain of embodied cognition we suppose that the realization of abstract categories like phonemes is constrained by the physical properties of the speech production apparatus. In order to understand the realization of phonemes, we try to understand the functioning of speech production mechanisms and their corresponding patterns of speech motor control. To do so, over the last decade we have carried out a number of kinematic studies focusing on control strategies in the production of different phonemes, the interaction of the tongue with its enclosing vocal tract boundaries, and the impact of speech rate on the temporal organization of inter-articulatory coordination. Moreover, we have compared the experimental data with simulations of a physically realistic tongue model in order to test our predictions concerning the underlying speech motor control mechanisms. In the current step of our work, we have built speaker-specific biomechanical tongue models to explain speaker-specific articulatory behavior on the basis of differences in vocal tract shape.

We consider physically realistic models to be primarily biomechanical, although geometrical models for instance can also involve a physically realistic description of the tongue and the surrounding vocal tract in the mid-sagittal plane. However, what is different in biomechanical models is that they include a description of the macro muscle fibers which are recruited to produce a certain speech movement. One might critically ask the questions: What kind of additional information can biomechanical models provide in comparison to geometrical models? Why is it worth using biomechanical models given their high computational cost?

First of all, we think the choice of the model crucially depends on the task to be fulfilled and one should choose a model appropriate to the research question. Biomechanical tongue models have the advantage that muscular activity, muscular forces, stiffness and muscular co-contraction underlying a given speech movement can be studied. Hence, they provide a window into the motor command level and speech motor control. Geometrical models may also follow biomechanical principles if they are based on large amounts of experimental data, but the movement between different tongue positions usually consists of a linear interpolation and does not follow physical principles. To study speaker-specific behavior, as is the aim of this paper, it is possible to build either speaker-specific geometrical or biomechanical models [1]. The crucial test, however, is to map the parameter space of one speaker onto another. In geometrical models the physical realism of this mapping is unknown. We think that for such a task biomechanical models provide a more elegant solution since macro fibers are mapped onto the tongue surface/volume and all articulatory movements and degrees of freedom are a result of muscle activation, not on the articulatory parameter space of some reference speaker. This reason and the fact that biomechanical models support the investigation of the relations between variability of the motor commands, of the articulation and of the acoustics motivated us to build these models.

2. Modeling methodology

The following processing steps describe how speaker-specific models are constructed, and are grouped by the overall aims of modeling speaker-specific articulation and speaker-specific acoustics. This section ends with an overview of how all the components act together.

2.1. General tasks

2.1.1. Image data acquisition

Speakers were recorded by means of Magnetic Resonance Imaging (MRI) with a Philips Gyroscan T10-NT Powertrack 1000 scanner. The image matrix was composed of 256 x 256 pixels, each with a spatial resolution of 1 mm in the y-direction and 1.4 mm in the x-direction. Data were originally collected for 10 isolated vowels to study inter-speaker acoustic and articulatory variability [2]. For each vowel three 18 slice series of 3.6 mm thick parallel sections with a 0.4 mm distance between slices were recorded. The three slice series differ in their respective slice orientation (transverse, coronal, oblique), chosen to be approximately orthogonal to the tract airway over some portion of its length.

2.1.2. Airway segmentation

Within each MR image the airway was segmented from the surrounding tissues manually. We used the itk-SNAP (version 2.1.4) software [3] for segmentation. Several rules were

established to ensure reliable segmentation suitable for speaker-specific articulatory and acoustic modeling. The biomechanical tongue model already includes standard teeth and standard lips. For that reason segmentation of the air channel terminates at the incisors. For acoustic purposes, the front tube for the lip region used during area function calculation is invariant. Another convention was to always exclude the epiglottis from segmenting the tongue, because the tongue contour is handled separately in the articulatory model. In cases where the epiglottis was located in the middle of the airway, the tissue of the epiglottis was ignored and the contour of the tongue body and the pharyngeal walls determined the cross-sectional area. If the epiglottis touched the tongue body, and hence tissues were hard to separate, tissue of the epiglottis was ignored on the basis of a smooth contour of the tongue back. For acoustic modeling the area function in the region of the epiglottis was systematically decreased (see Sec. 2.3.3). A last convention regards the uvula, which is not part of the biomechanical tongue model. In order to allow for comparable distance function values in the corresponding region, the uvula was excluded during image segmentation as well.

2.1.3. Spatial reconstruction

Airway segmentation was accomplished by producing binary images of each MR image with one value for segmenting a pixel as air surrounded by tissue, and a second value otherwise. The resulting contours were sub-sampled with 100 equally spaced points. The single contours were combined to form a vocal tract by re-orientating each contour of the three series of 18 slices according to their original spatial orientation. The derived 3D wire-frame model of the relevant vocal tract airway constitutes the basis for speaker-specific modeling.

2.2. Biomechanical modeling

The biomechanical component of the model is based on the 2D tongue model [4]. This model has been shown capable of accounting for some important kinematic characteristics of speech articulation such as velocity profiles, trajectory shapes [4], or relations between trajectory curvature and speed [5]. It mainly consists of a deformable Finite Element Mesh (FEM) embedded in rigid vocal tract walls in the mid-sagittal plane. The geometry of the mesh has been specifically designed to facilitate anatomical implementation of the muscles within the tongue. The model is controlled according to the λ-model [6], which generates muscle force as a function of the difference between a centrally specified threshold length λ and the actual muscle length. Two basic hypotheses underlie the design of the speaker-specific models: (1) the general anatomical arrangements accounted for by the mesh geometry is common to all human beings, (2) variations across speakers in muscle lengths and muscle orientations within the tongue are the results of global variations of the head morphology such as variations in larynx height, mandibule ramus length, head size, and mid-sagittal palate shape. Hence anatomical landmarks have been selected that enable characterization of each speaker using these morphological parameters: the mid-sagittal palatal contours and tongue contours at rest, the lower and upper limits of the tongue insertions on the mental spine (P1 and P2), and the styloid process (P3) were the landmarks chosen.

2.2.1. Creating the mid-sagittal vocal tract contour

The individual mid-sagittal palatal and tongue contours were extracted from MRI data for the subjects at tongue rest position and used to align the coronal and transverse contours of the vocal tract resulting from the airway segmentation (Section 2.1.2). After determining a global left and a global right extremum of the vocal tract, the mid-sagittal plane was established as the vertical plane passing through the center of these extrema. Then, adjustment of the mid-sagittal position was done by hand for each speaker interactively to ensure that the plane bisected all the contours.

2.2.2. Determination of anatomical landmarks

P1, P2 and P3 were measured from a high-definition mid-sagittal view of the speaker's head recorded with anatomical MRI. P1 and P2 were determined on the basis of grey level changes in the mental spine region. P3 was determined using the internal contour of the sphenoid bone in the mid-sagittal plane. It was located at 1/3 of the global length of this contour from the bottom.

2.2.3. Model creation and adaptation

Given these anatomical landmarks matching the original generic biomechanical model to speaker-specific anatomy was straightforward. First the upper contour of the tongue model was projected onto the mid-sagittal tongue contour measured for the subject. This contour corresponds to the upper limit of the new speaker-specific FEM tongue mesh. Second, the distribution of the nodes along this new upper contour was done proportionally to the distribution of the nodes in the original model, selected for anatomical appropriateness. Third, the lower and upper attachment points of the new tongue mesh on the mandible were assigned to points P1 and P2. Then, the distribution of the nodes within the mesh was obtained by deforming the original mesh linearly from the nodes on the upper contour to the insertion nodes of the mesh into the mental spine. Finally, the extremity of the external Styloglossus fiber was attached to point P3. This matching procedure fully determines the geometry of the new mesh and consequently muscle arrangement within the new speaker-specific tongue model. It preserves the original topology of the mesh while accounting for the speaker-specific morphology.

2.3. Acoustic modeling

In order to determine acoustics from articulation, the 2D distance function resulting from the biomechanical tongue model has to be converted to its corresponding area function. In our approach this is accomplished by determining the distance function based on a pre-defined grid and subsequent reconstruction of the area according to the coefficients associated to the respective grid line.

2.3.1. Grid orientation

Different kinds of grids have been suggested for reconstructing area functions from vocal tract distance functions, following at least two constraints which are important for determining acoustics: grid lines should ideally be perpendicular to the vocal tract midline and area values should not change abruptly. In our approach we use the grid proposed in [4] and adapt it to the vocal tract morphology of a particular speaker. The applied grid additionally allows the precise determination of coefficients necessary for the calculation of the area function, since the grid line orientation roughly follows the orientation of the slices during MRI acquisition.

2.3.2. Determining α coefficients

For the purpose of area function calculation we adapted a modified version of the α-β-model proposed in [7]. Essentially α-coefficients have different values depending on the vocal tract region (i.e. lips, palate, etc.), but α values also vary within one region for different speakers. For that reason we determine α-values for each speaker as follows: 3D wireframe representations of the three cardinal vowels as well as the neutral vowel are used to relate the distance values to the area values of a speaker. We further determine adequate thresholds corresponding to small and large distance values, leading to two speaker-specific α- values per contour.

2.3.3. Determining area functions and acoustics

Area values are calculated based on the mid-sagittal distances by applying the corresponding α coefficient of large values, if a distance is larger than the pre-determined threshold for large sagittal distances. If a distance is less than the pre-determined threshold for small distances, α-coefficient of small values were applied. For distances in between, α-coefficients were linearly interpolated before calculation of an area value. The segmentation of the airway in the image volumes ends in the region of the incisors. To control for variation in the lip region the area values at the lips are modeled by a tube of length 10 mm and 19.5 mm diameter for open vowels and 20 mm and 8 mm respectively for closed vowels. Determining acoustics from area functions does not involve any speaker-specific adaptation. In our approach formants are simply computed by coupling an acoustic analog of the vocal tract with the reconstructed area functions.

2.4. Simulating tongue movements

In order to simulate a tongue movement, motor commands are interpolated starting at values resulting from tongue rest position and reaching the specified target motor command. Within this modeling framework we are capable of relating every tongue target position represented as set of six muscle activations to the resulting tongue shape and the corresponding first two formant values.

3. Studying speaker-specific variability

In the following section we present a first application of the individual biomechanical models. Two speaker-specific models were used to investigate the impact of the location of the bending vocal tract on the inter-speaker variability in the articulatory and motor command space.

3.1.1. Motivation

Acoustic models of speech production suppose that the length of the vocal tract, and the length and location of the constriction defining the resonance cavities are crucial parameters for the description of the spectral properties of vowels. The location of the vocal tract bend seems to play a negligible role with respect to the acoustics, assuming the length of the tract is kept constant [8]. However, the relative location of the bend within the vocal tract is crucial with respect to determining the vertical and horizontal dimensions of the tract and is likely to influence the degrees of freedom of articulatory motion in the respective directions. Moreover, the location of the bend probably also affects the biomechanical properties of the tongue, since the tongue is soft tissue and deforms to its surrounding structures. Hence, the location of the vocal tract bend may be one source of inter-speaker articulatory variation and may also be reflected in the motor command level.

We hypothesize that the location of the vocal tract bend and the corresponding relation between horizontal and vertical portions of the vocal tract affect the degrees of freedom in the articulatory domain. For instance, a speaker with a relatively long vertical vocal tract dimension and a short horizontal dimension (i.e. a rather anterior location of the bending vocal tract) should have a relatively large degree of freedom moving vertically up and down, but should be constrained in the horizontal direction, moving the tongue front and back.

3.1.2. Method

In order to study inter-speaker variability two individual biomechanical tongue models were produced. The two distinctive speakers used for building the models were selected from a MRI database of 9 French speakers. One female speaker (AV) has a vocal tract with approximately equal lengths of the pharynx (vertical dimension) and the palate (horizontal dimension from the velum to the incisors). The male speaker (CS) has, in comparison to AV, a longer pharynx and a slightly shorter palate. Thus, the two speakers represent vocal tract morphologies typically found in female and male speakers, and they show some differences in the relative location of the vocal tract bend. They also show some differences in the shape and steepness of the palate. These differences in vocal tract shape have already been quantified by means of a Principal Component Analysis in [9].

For each model we ran a series of 8500 simulations in order to account for the complete vowel space. To run the simulations, motor commands of six major tongue muscles were randomly varied with respect to the values at rest position.

Assuming that auditory targets are the primary goals for vowel production, we kept the acoustic variability across speakers constant. Three formant ellipses were defined in the F1/F2 plane for the three corner vowels. Prior to the definition of the formant ellipses, the area functions were manipulated to match the acoustic vowel space of the two speakers. In a first step the length of the epilaryngeal tube of speaker CS was shortened to match that of speaker AV. Secondly, the original area values in the laryngeal region were fixed to a physiologically realistic value of 0.4 cm^2. In the region of the epiglottis the area values were decreased by 50% for both speakers to avoid an artificially large cavity resulting from the excised epiglottis in the tongue model. Finally, the area functions were re-scaled to a vocal tract length of 17 cm, since we wanted to disentangle the impact of vocal tract length and the location of vocal tract bending. For each acoustic vowel target 60 simulations per speaker were randomly selected whose respective first two formant values are equally distributed over the corresponding ellipse area. To summarize, we define the same acoustic target regions for the three corner vowels and investigate the corresponding speaker-specific articulatory variability underlying these acoustic goal regions. Moreover, we obtain some indication of the variability of the motor commands.

3.1.3. Results

Figure 1 shows the articulatory results for model CS and model AV. Dispersion ellipses are plotted at the constriction location for /i/, for /u/ and for /a/.

For /i/ the main diagonal of the dispersion ellipses follows the palatal contour in both models. The shape of the palate, which differs among the two models, is therefore crucial for the production of /i/. Model AV shows greater variability in the horizontal dimension (as would be predicted) for the most posterior dispersion ellipse, since this is at a location where the

palate starts to be straight. The main difference in terms of motor commands between both models is the variability in Styloglossus activation. Model AV shows large variation of Styloglossus activity in comparison to model CS.

For /u/ and /a/ the articulatory results confirm our hypothesis. Model AV always shows a larger variability in the horizontal direction than model CS, but it is more constrained in the vertical direction than model CS. In /u/ model AV shows a greater variability in Hyoglossus activation than model CS and for /a/ the variability of Hyoglossus and Styloglossus are greater for model CS than model AV.

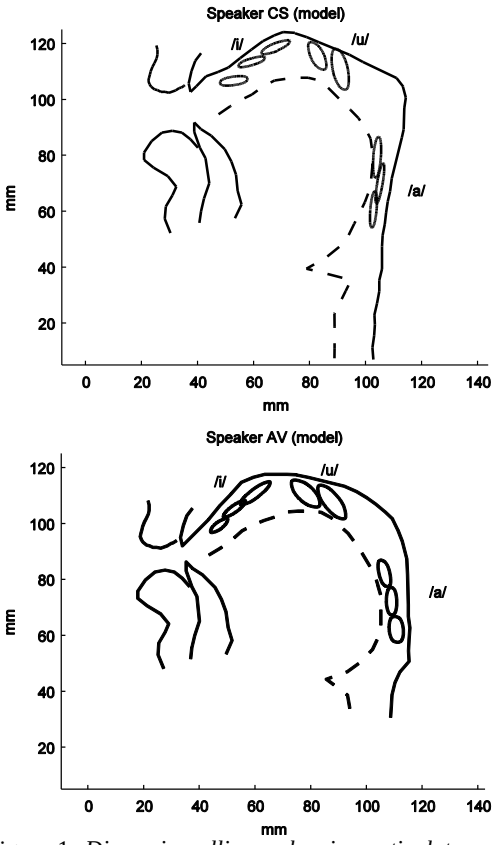

Figure 1: *Dispersion ellipses showing articulatory results for model CS on top and model AV at the bottom. Dashed tongue contour corresponds to rest position.*

The linkage between articulatory positions and muscle activation is not straightforward, because tongue positioning is the result of the activation of several muscles simultaneously, which makes the linkage highly nonlinear. However, certain combinations of muscle activation lead to the characteristic constrictions for vowels and so far we consider these muscles to be important for the study of variability. For /a/ a constriction is formed in the pharyngeal region. This constriction is produced by a combination of Hyoglossus (down and back) and Styloglossus (up and back) activation. Both muscle activations display a larger variability in the CS model than in the AV model, but the AV model also shows smaller λ-values, and then larger activation levels. For /i/ the palatal constriction is primarily accomplished by a combination of Posterior Genioglossus (GGP, up and front) and Styloglossus activation. Model AV allows more variability in Styloglossus activation than model CS. Note however that this is not the case for GGP. This reflects the very precise requirement for a small constriction area located in the front palatal region, where the force of GGP (front direction) is directed. For /u/ the velar constriction results primarily from the activation of Styloglossus and GGP (similar to /i/, but with a different balance), in combination with the activation of Hyoglossus, which together move the constriction location backward, as in the production of /a/. Our data do not show any difference in λ-activations across subjects. There is also a trend in both subjects for greater variability for /u/ as compared to /a/ and /i/. This may reflect the fact that the requirements for tongue control accuracy are less for /u/ than for /i/ or /a/ due to the rounded lips.

4. Conclusion

Speaker-specific biomechanical models have been designed to study articulatory and motor control variability. This approach has been shown to give interesting insights into individual articulatory strategies, and can be used to relate them to individual morphological characteristics. Future work will focus on interpreting this data in relation to the achievement of perceptual goals.

5. Acknowledgements

This work was supported by the German Research Council to the SPEECHart project (Grant Nr. FU 791/1-1) and by the French-German University to the PILIOS project. It was initiated thanks to the Christian Benoit Award given to the second author in 2007.

6. References

[1] Fuchs, S., "Benoit Project: Speaker specific vowel articulation", Online: http://benoit.susannefuchs.org/tutorial3/en.html.

[2] Apostol, L., "Étude et simulation des caractéristiques individuelles des locuteurs par modélisation du processus de production de la parole", Unp. PhD thesis, INP Grenoble, France, 2001.

[3] Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C. and Gerig, G., "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability", Neuroimage, 31(3):1116-28, 2006.

[4] Perrier, P., Payan, Y., Zandipour, M. and Perkell, J., "Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study", JASA, 114(3):1582-1599, 2003.

[5] Perrier, P. and Fuchs, S., "Speed-curvature relations in speech production challenge the one-third power law", Journal of Neurophysiology, 100:1171-1183, 2008.

[6] Feldman, A.G., "Once More on the Equilibrium-Point Hypothesis (Lambda-Model) for Motor Control", J Mot Behav., 18(1):17-54, 1986.

[7] Perrier, P., Boë, L.J., and Sock, R., "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients", J Speech Hear Res., 35:53–67, 1992.

[8] Sondhi, M.M., "Resonances of a bent vocal tract", JASA, 79(4):1113-1116, 1986.

[9] Fuchs, S., Winkler, R. and Perrier, P., "Do Speakers' Vocal Tract Geometries Shape their Articulatory Vowel Space?", in Proc. ISSP Strasbourg, pp. 333-336, 2008.