Towards Articulatory Speech Synthesis with a Dynamic 3D Finite Element Tongue Model

Kees van den Doel[1] **, Florian Vogt**[2] **, R. Elliot English**[1] **, Sidney Fels**[2]

[1] Department of Computer Science,

[2]Department of Electrical and Computer Engineering,
University of British Columbia
Canada

kvdoel@cs.ubc.ca,fvogt@ece.ubc.ca,ssfels@ece.ubc.ca

Abstract. *We describe work towards articulatory speech synthesis driven by realistic 3D tissue and bone models. The vocal tract shape is modeled using a fast 3D finite element method (FEM) of a muscle-activated human tongue in conjunction with fixed rigid models of jaw, hyoid and palate connected to a deformable mesh representing the airway. Actuation of the tissue model deforms the airway providing a time-varying acoustic tube which is used for the synthesis of sound. We describe our initial validation of our models geometrically using magnetic resonance images and acoustically using articulatory configurations.*

1. Introduction

Different anatomical substructures of the vocal tract, such as the tongue, jaw, hyoid, larynx, lips, and face, have been modeled using both parametric and physically-based dynamic models. We mention for example Rosa and Pereira (2004); Dang and Honda (2001). Many of these models are very complex and they are often developed independently of other structures. The aero-acoustical processes that involve the interaction of these anatomical elements with airflow and pressure waves leading to speech production have also been studied (Svancara et al. (2004); Sinder et al. (June, 1998)). We are working toward integrating these models in a common platform, ArtiSynth (Fels et al. (2005)), for speech synthesis.

Articulatory speech synthesis to date has been mainly driven by 2D geometrical parameters that are closely connected to articulatory configurations that are relevant to speech sounds. We make no attempt to review the vast literature on this, but mention for example the ASY system and its extensions (Rubin et al. (1996, 1981)). Attempts to drive speech synthesis by faithfully simulated motion of 3D tissue models faces the problem of missing components, as it is usually prohibitively difficult, both technically and organizationally, to obtain models of all relevant anatomical parts that define the 3D airway in which acoustical phenomena occur.

In this report we describe a possible solution to this problem for the creation of speech sounds driven by motion of a muscle-activated human tongue model, described

earlier in Gerard et al. (2006); Vogt et al. (2006). Apart from the tongue, we have three other anatomical parts in place: the jaw, the hyoid, and the palate. These parts are currently modeled as static fixed meshes and only serve to constrain the tongue motion within realistic bounds and to constrain the airway. We expect to integrate our dynamic jaw and hyoid model (Stavness et al. (2006)) with the tongue and palate in the very near future.

For the acoustical modeling we insert and attach a separate airway model, modeled as a deformable mesh, to the anatomical models. This mesh represents the volume of air present in the vocal tract and is in principle completely determined by all the anatomical parts. The airway is registered partly with the tongue and partly with the palate. Unregistered parts of the airway mesh function as placeholders for missing anatomical parts such as cheeks and lips. This allows modeling of speech phenomena with an incomplete model of the vocal tract, focusing attention on the structure at hand, in our case the tongue. The airway model effectively acts as an extrapolator and regulator function of the vocal tract area function. As new anatomical components are added, the airway mesh becomes more accurate.

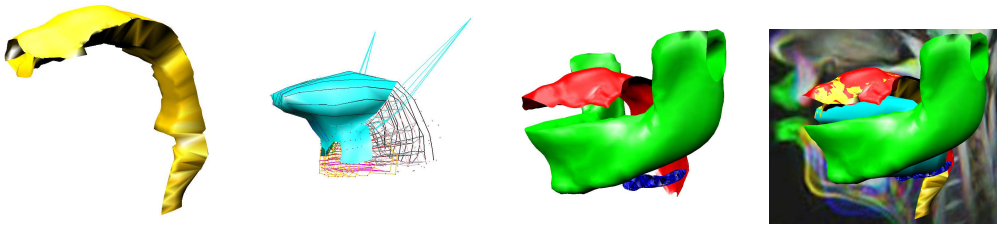
The aero-acoustical modeling may be done using only the resultant airway mesh. This can be computed independently from the anatomical models once a configuration has been determined and does not need to be aware of the tongue. However, our approach includes dynamical interaction between the tongue and air flow in the vocal tract/airway as pressure can be computed on the airway wall and transmitted to the tongue model. We hope to tune our models in the near future to be able to model complex interactions for speech phenomena involving oscillations of the articulators such as various trills.

The remainder of this paper is organized as follows. In Section 2 we describe the 3D biomechanical tongue, palate and jaw models used and describe collision detection and resolution between the structures in motion. In Section 3 we describe the 3D airway and how it is connected to the tongue and palate. We describe the initial, manual, registration process and the automated final locking mechanism. We describe how pressure forces can be transmitted through the airway mesh onto the tongue. At this point in our investigation, we have completed combining the tongue, jaw, palate and airway and can drive the configuration with tongue muscles. We are following a process to validate our models geometrically using magnetic resonance images and acoustically using articulatory configurations as described in Section 4.

2. Biomechanical Tongue Model

The biomechanical finite element tongue model that we use is described in Gerard et al. (2006). The model, depicted in Fig. 1, contains 946 nodes and 740 hexahedral elements. The tongue model is activated by 11 physiological representative muscle groups. We have ported this tongue model to ArtiSynth and were able to achieve interactive simulation rates by using a quasi linear stiffness-warping scheme combined with an implicit numerical integration method to permit large time steps without numerical instability, at the expense of some small loss of accuracy (Vogt et al. (2006)).

For the implementation of the of muscle activations, we designate specific collections of finite elements as muscles, each associated with a force activation level. Certain FEM edges within each muscle are selected to act as “fibres”, along which a uniform



(a) (b) (c) (d)

Figure 1. The airway (a) is connected dynamically to the muscle activated finite element tongue model (b) and palate (c red). The palate, jaw, and hyoid (c) constrain the tongue. The complete model (d) is registered using magnetic resonance images.

contractile force is exerted. More details on the motivation for the muscle model and the selection of the muscles and fibres are given in Gerard et al. (2006). The tongue muscles are attached by fixing certain FEM nodes to the jaw, hyoid, and skull bones, which we have modeled as fixed rigid bodies. In the future we expect to connect the tongue to our dynamic jaw and hyoid model (Stavness et al. (2006); Langenbach and Hannam (1999)), which consists of a fixed rigid skull, floating rigid mandible, two temporo-mandibular joints, eighteen muscles, and multiple bite points.

The tongue motions are further constrained by the jaw, the hyoid, and the palate which are fixed static meshes depicted in Fig. 1. Collision detection and response is necessary to prevent the tongue from penetrating these components. Our approach is to find the intersections with the tongue's surface mesh and then project the penetrating vertices onto the rigid body's mesh. We chose this technique over a force based technique because it does not require parameter tuning to control penetration depth and produces more stable results. Despite the increased complexity of the code it is still very efficient, using far less cpu time than the FEM computations themselves.

To find the penetrating vertices we first determine all of the triangle-triangle intersections with the tongue using the fast triangle-triangle intersection algorithm described in Möller and Trumbore (1997). To reduce the number of intersection tests required we apply an oriented bounding box (OBB) tree (Gottschalk et al. (1996)). The intersections are found by recursively testing each face from the tongue's surface mesh with the OBB trees from the palate and jaw. From the resulting set of triangle-triangle intersections, the vertices outlining the intersections are found by iterating through the intersecting edges and tracking whether or not the ends of each edge are penetrating the mesh. A coloring algorithm is then used in conjunction with this set of vertices to find the vertices penetrating the rigid body.

The final step is to iterate through each of the penetrating vertices and project them onto the set of penetrating faces from the rigid body. In the majority of cases this yields excellent results. However, it is not a complete solution. For example when the tip of the tongue is stretched out, no vertices penetrate but edge or face intersections occur, allowing a significant amount of penetration. We are currently looking at solutions to this problem which include also projecting penetrating faces and edges, in addition to the vertices, onto the colliding body.

3. Airway Modeling

To produce speech, aero-acoustical phenomena that occur in the vocal tract airway have to be modeled. In principle, the airway is determined implicitly by its adjacent anatomical components, but as some of these components may not yet have been modeled, or may be of limited relevance, we have developed a stand-alone version of the vocal tract airway. Such explicit airway modeling is also described in Yehia and Tiede (1997); Honda et al. (2004).

Our airway consists of a mesh-based surface model depicted in Fig. 1a. The mesh is structured as a number of cross-sections along the length of the tube, which implicitly defines a center line. This structure permits the fast calculation of the area function, which allows the acoustics to be modeled in a cylindrically symmetrical tube. The airway is deformable and will change shape in concert with the anatomical components which surround it.

The wave propagation through the vocal tract is modeled using the linearized Navier-Stokes equations which we solve numerically in real-time on a 1D grid using an implicit-explicit Euler scheme (Doel and Ascher (2006)). The method remains stable when small constrictions in the airway generate strong damping. An advantage of this approach over the well-known Kelly-Lochbaum (Kelly and Lochbaum (1962)) tube segment filter model is that the airway can be stretched continuously (when pursing the lips for example) which is not possible with the classical Kelly-Lochbaum model which requires a fixed grid size.

The vocal chords are modeled using the Ishizaka-Flanagan two-mass model (Ishizaka and Flanagan (1972); Sondhi and Schroeter (1987)). This model computes pressure oscillations as well as glottal flow and is dynamically driven by lung pressure and tension parameters in the vocal chords. The vocal chord model is coupled to the discretized acoustics equation in the vocal tract. Noise is injected at the narrowest constriction and at the glottis according to the model described in Sondhi and Schroeter (1987). The resulting model is capable of producing vowels as well as fricatives.

For a given palate mesh and tongue mesh the airway is positioned first manually into a reasonable position with respect to these anatomical parts. For this purpose we use a combination of Amira (Amira (2006)) and ArtiSynth. After the initial approximate registration phase, we make the final connection. For each vertex $\boldsymbol{v}$ on the airway mesh we search for the closest point $\boldsymbol{q}$ on the tongue or palate mesh. If this distance is smaller than a cutoff of about $5mm$ a connection is made and the airway vertex is moved to $\boldsymbol{q}$. We maintain a list of registered vertices and their corresponding points on the tongue or palate, which we store in terms of the barycentric coordinates with respect to the corresponding face. At every time-step in the simulation, whenever the tongue moves the corresponding airway vertices move also. Additionally, since the FEM model representing the tongue is such a coarse approximation, we use curved point-normal triangles (Vlachos et al. (2001)) to interpolate the tongue's mesh, improving the resulting shape of the airway mesh. Airway vertices registered to the palate and remaining unregistered vertices representing the missing anatomical parts remain at a fixed location relative to the palate at runtime. In Fig. 2 we show the meshes before and after registration.

At every time-step during the simulation the area function is recomputed. The

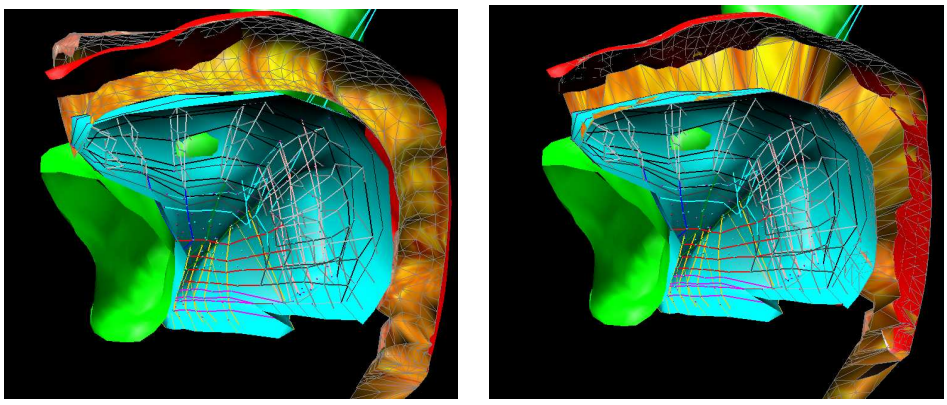


Figure 2. The jaw, tongue, palate, and airway meshes are depicted after manual positioning (left) and after final registration (middle).

air pressure is available at every point along the centerline of the 1D tube model and is extrapolated back to the airway faces. The air velocity u along the airway determines the pressure P through the steady state solution of the Navier-Stokes equation (Bernoulli's law)

$$P = P_l + \frac{\rho}{2} u_g (1 - \left(\frac{A_g}{A}\right)^2),$$

where P_l is the lung pressure, ρ is the air density, u_g is the glottal velocity (obtained from the Ishizaka-Flanagan model), A_g is the area at the glottis and A is the area of the section at which P is computed. Because of the sectioned structure of the airway mesh each triangular face connects two planar cross-sections. The pressure P on a face is taken to be the average of the pressures $P1$ and $P2$ taken at the centerline points defined by the sections. The section pressures $P1$ and $P2$ are obtained from the 1D grid used for the solution of the acoustics equation. If so desired, a finer grid could be used for the flow modeling as done in Sinder et al. (June, 1998). Once the pressure P on an airway mesh face is known, the force exerted on a vertex $\boldsymbol{v}$ of the face is given by $A(P - P_0)\boldsymbol{n}/3$, where A is the area of the face, P_0 is the atmospheric pressure, and $\boldsymbol{n}$ is the normal to the face. The total force on a vertex $\boldsymbol{v}$ is obtained by summing contributions of all faces containing $\boldsymbol{v}$. An airway vertex to which a force $\boldsymbol{F}^{\text{airway}}$ is applied and that is connected to a point $\boldsymbol{q}$ on the tongue, transmits a force $\boldsymbol{F}_i^{\text{tongue}}$ to each vertex $\boldsymbol{w}_i$ of the tongue face that $\boldsymbol{q}$ belongs to according to $\boldsymbol{F}_i^{\text{tongue}} = \lambda(\boldsymbol{q})_i \boldsymbol{F}^{\text{airway}}$, where $\lambda(\boldsymbol{q})_i$ is the i'th barycentric coordinate of $\boldsymbol{q}$ with respect to the triangular tongue face.

4. Speech Synthesis Results

Thus far we have only been able to perform preliminary testing and validation of the models by generating tongue poses corresponding to the vowels /ə/, /ʊ/, /ɑ/, and /ɪ/ (using IPA notation). For the first three we determined muscle activations by matching the 2D cross-section of the tongue shape to magnetic resonance images shown in Fig. 4. The three red, green and blue channels correspond to /ə/, /ʊ/, and /ɑ/. The fourth vowel was tuned manually to match the formants of /ɪ/. The tongue activations are shown in table 1. The resulting tongue muscle activations were then used to deform the airway and create

area functions which were used for acoustical simulation. In Fig. 3 we plot the formants and the area functions and compare the F1 and F2 values to some canonical values for each phoneme. We use the centre of the vowel region as the canonical target for F1 and F2.

phoneme	gga	ggm	ggp	sty	tr	il	sl	F1	F2
/ə/	.32	.75	1.3	.64	6	.43	.75	470	1510
/ʊ/	.65	0	0	1.4	1	.22	0	510	1359
/ɑ/	1.4	0	0	0	0	1	1	670	1280
/ɪ/	0	0	4	2.25	8.6	0	.54	300	1920

Table 1. Non zero muscle activations (in N) for each phoneme, and their first two formants (in Hz). The muscles and their abbreviations are: Genioglossus (anterior, medium, posterior) [gga ggm ggp], Geniohyoid [gh], Hyoglossus [hg], Styloglossus [sty], Superior longitudinal [sl], Inferior longitudinal [il], Transversalis [tr], Verticalis [vert] and Mylohyoid [mh].

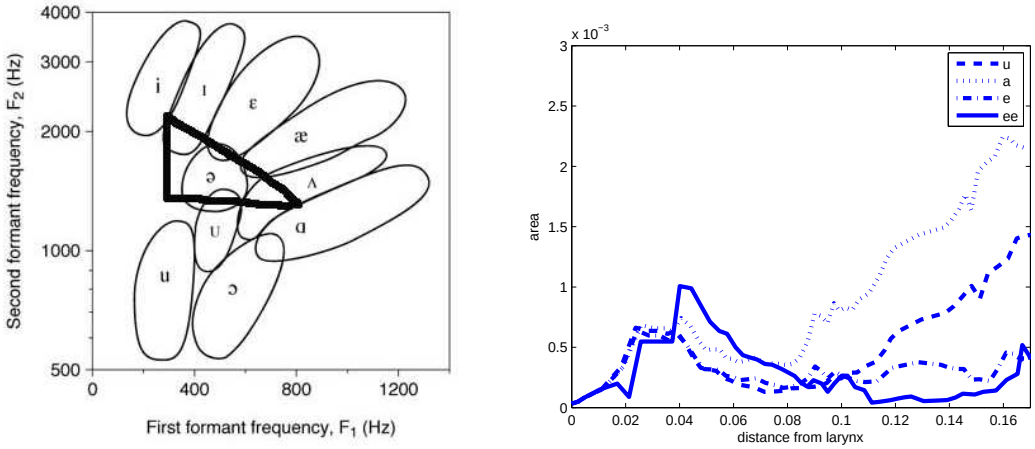


Figure 3. First versus second formant for four vowels and their area (in m^3) functions (x-axis (in m) starts at 0 towards the larynx). The vowels and their first two formants (obtained/canonical target) are: e=/ə/ (470-1510/450-1700,), u=/u/ (510-1359/500-1200), a=/ɑ/ (670-1280/1000-1300), and ee=/ɪ/ (300-1920/200-2700). The triangle indicates vowels that can be reached with the current model. Figure from http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/filter.html.

Because of several limitations in our current model, the range of vowels is rather restricted as can be seen in Fig. 3. We believe these restrictions can be overcome by improving the current model. Issues to improve on are: 1) the lack of a lip model which prevents us from lowering the second formant in particular for the back vowel /ʊ/; 2) the lack of a movable jaw prevents us from forming the open vowels such as /ɑ/ with higher first formant; 3) the second formant of the front vowel relies on a very accurate model of the area function for which our current coarse tongue model is not sufficiently accurate as can be see, for example, in the difficulty in forming a small closure near the mouth for /ɪ/;

4) the lack of a geometric model for the teeth and cheeks (modeled by the unregistered sides of the airway) prohibits the calculation of an accurate area function. Finally, we intend to use 3D data to match our models to rather than 2D data.

In the near future we plan to address these issues and improve the model. In addition, we also hope to address the more ambitious goal of utilizing the dynamics of the tongue and the airflow to synthesize sounds which depend crucially on complex tongue motion such as an alveolar trill (/r/,rolled "r").

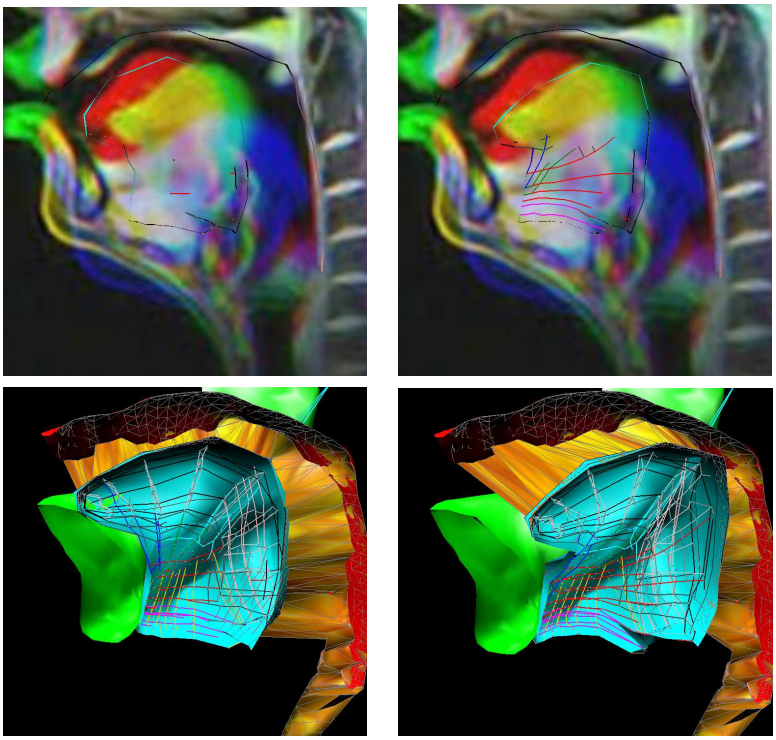


Figure 4. Configurations producing the vowels ɘ (left) and ʊ (right). We use the top figures to find activation parameters to configure the tongue according to magnetic resonance images (R:/ɘ/,G:/ʊ/,B:/ɑ/).

5. Conclusions

We have presented our progress towards integrating a 3D tongue model with a static jaw and palate model. We are at the point of validating this combination geometrically and acoustically so that we understand the tradeoffs between model fidelity and speech synthesis quality. This understanding is necessary before we continue to add more complexity to our models including activating the jaw muscles and adding lips and cheeks as part of a complete 3D vocal tract model for articulatory speech synthesis.

Acknowledgements

We would like to thank P. Perrrier, Y. Payan for the contribution of the tongue model and Mark Tiede for providing the MR images and Veronica Orvalho for help with the mesh editing.

References

Amira. Amira 3D Data Visualization Software - Project Homepage, http://www.amiravis.com, 2006.

Dang, J. and Honda, K. A physiological articulatory model for simulating speech production process. *JASJ*, 22(6):415–425, 2001.

Doel, K. v. d. and Ascher, U. Staggered grid discretization for the Webster equation. *in preparation*, 2006.

Fels, S., Vogt, F., van den Doel, K., Lloyd, J., and Guenter, O. Artisynth: Towards realizing an extensible, portable 3d articulatory speech synthesizer. In *International Workshop on Auditory Visual Speech Processing*, pages 119–124, July 2005.

Gerard, J., Perrier, P., and Payan, Y. *3D biomechanical tongue modelling to study speech production*, pages 85–102. Psychology Press: New York, USA., 2006.

Gottschalk, S., Lin, M. C., and Manocha, D. OBBTree: A hierarchical structure for rapid interference detection. *Computer Graphics*, 30:171–180, 1996.

Honda, K., Takemoto, H., Kitamura, T., and Fujita, S. Exploring human speech production mechanisms by mri. *IEICE Info Sys*, E87-D:1050–1058, 2004.

Ishizaka, K. and Flanagan, J. L. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Systems Technical Journal*, 51:1233–1268, 1972.

Kelly, K. L. and Lochbaum, C. C. Speech Synthesis. In *Proc. Fourth ICA*, 1962.

Langenbach, G. and Hannam, A. The role of passive muscle tensions in a three-dimensional dynamic model of the human jaw. *Arch Oral Bio*, 44:557–573, 1999.

Möller, T. and Trumbore, B. Fast, minimum storage ray-triangle intersection. *Journal of Graphics Tools*, 2(1):21–28, 1997.

Rosa, M. O. and Pereira, J. Towards full-scale three dimensional larynx simulation. In *Proc ICVPB*, 2004.

Rubin, P., Saltzman, E., GoldStein, L., McGowan, R., Tiede, M., and C, B. Casy and extensions to the task-dynamic model. In *Proc 4th Sp Prod Sem*, pages 125–128, 1996.

Rubin, P. E., Baer, T., and Mermelstein, P. An articulatory synthesizer for perceptual research. *JASA*, 70:321–328, 1981.

Sinder, D. J., Krane, M. H., and Flanagan, J. L. Synthesis of fricative sounds using an aeroacoustic noise generation model. In *Proc. ASA Meet.*, June, 1998.

Sondhi, M. M. and Schroeter, J. A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer. *IEEE Trans on ASSP*, 35(7):955–967, 1987.

Stavness, I., Hannam, A. G., Lloyd, J. E., and Fels, S. An integrated, dynamic jaw and laryngeal model constructed from ct data. *Springer LNCS 4072*, pages 169–177, 2006.

Svancara, P., Horacek, J., and Pesek, L. Numerical modeling of production of czech vowel /a/ based on FE model of vocal tract. In *Proc ICVPB*, 2004.

Vlachos, A., Peters, J., Boyd, C., and Mitchell, J. L. Curved PN triangles. In *Symp. on Int. 3D Graph*, pages 159–166, 2001.

Vogt, F., Lloyd, J. E., Buchaillard, S., Perrier, P., Chabanas, M., Payan, Y., and Fels, S. S. Investigation of efficient 3d finite element modeling of a muscle-activated tongue. *Springer LNCS 4072*, pages 19–28, 2006.

Yehia, H. C. and Tiede, M. A parametric three-dimensional model of the vocal-tract based on mri data. In *Proc ICASSP*, pages 1619–1625, 1997.