A comprehensive 3D biomechanically-driven vocal tract model including inverse dynamics for speech research

Peter Anderson\textsuperscript{1}, Negar M. Harandi\textsuperscript{2}, Scott Moisik\textsuperscript{2}, Ian Stavness\textsuperscript{3}, Sidney Fels\textsuperscript{3}

\textsuperscript{1}University of British Columbia
\textsuperscript{2}Max Planck Institute for Psycholinguistics
\textsuperscript{3}University of Saskatchewan

\begin{flushleft}
\texttt{peter.anderson@alumni.ubc.ca, negar.mohaghegh@gmail.com, Scott.Moisik@mpi.nl, stavness@gmail.com, ssfels@ece.ubc.ca}
\end{flushleft}

\section*{Abstract}

We introduce a biomechanical model of oropharyngeal structures that adds the soft-palate, pharynx, and larynx to our previous models of jaw, skull, hyoid, tongue, and face in a unified model. The model includes a comprehensive description of the upper airway musculature, using point-to-point muscles that may either be embedded within the deformable structures or operate externally. The airway is described by an air-right mesh that fits and deforms with the surrounding articulators, which enables dynamic coupling to our articulatory speech synthesizer. We demonstrate that the biomechanics, in conjunction with the skinning, supports a range from physically realistic to simplified vocal tract geometries to investigate different approaches to aeroacoustic modeling of vocal tract. Furthermore, our model supports inverse modeling to support investigation of plausible muscle activation patterns to generate speech.

\textbf{Index Terms:} biomechanical simulation, upper airway model, speech synthesis, inverse modeling.

\section{1. Introduction}

Speech production and perception are an essential part of daily life for most people, yet the mechanisms involved are extremely complicated and remain poorly understood. Speech research is usually performed experimentally on people; however, measurements may be highly limited to the measurements that can feasibly and ethically be performed on humans and within the instrumentation capabilities. Simulations may complement experiments to address many of these concerns, and provide a framework for testing hypotheses and making predictions that cannot be done experimentally.

Models of the human vocal tract (VT) initially focused on the function of individual components. For example, the face \cite{1}, the tongue \cite{2}, soft palate \cite{3,4,5}, and larynx \cite{6}. Recent progress has been made towards developing a comprehensive model of the VT, as seen in the face, tongue, skull, jaw, hyoid model of Stavness et al. \cite{7} and the tongue and epiglottis model by Pelletier et al. \cite{8}. The model we present in this paper is the most complete that the authors are aware of, including finite element (FE) models of the face, tongue, soft palate, pharynx, larynx, and supporting structures of bone and cartilage.

Speech simulation involves the simulation of sound production and propagation through the airway until it radiates into free space beyond the lips. The glottis typically provides the main sound source, though sound may also be generated by other sources as in stop consonants and fricatives \cite{9}. Ishizaka & Flanagan \cite{10} modeled the glottis as a system of two masses connected by springs, which has since been challenged and elaborated \cite{11,12}. Most sound propagation models attempt to reduce the airway to a 1D area function and calculate sound if propagating though a tube of varying cross-section \cite{13,14}. The more sophisticated 1D models include the sub-glottal tract, a split airway to include the nasal passage, resonator cavities, sound sources at locations of (modeled) turbulence, and sound modification by the walls \cite{13}. More recent efforts have simulated sound propagation using the 3D wave equation \cite{15} or the Navier-Stokes equation \cite{16}.

In our simulation toolkit Artisynth, we have developed a complete biomechanical model of the upper airway structures, which is coupled to an acoustics synthesizer and may be used to calculate inverse mechanics (given a motion, find the muscle activations) as well as forward mechanics (find the motions caused by muscle activations). In this paper, we present this model and demonstrate how it may be used towards numerous applications in speech production and perception.

\section{2. Methods}

\subsection{2.1. Model Design}

Our VT model is composed of a deformable FE bodies, rigid bodies, surface skin meshes, and point-to-point muscles. A mid-sagittal view of the VT is shown in Fig. 1, and the components are summarized in Table 1. The laryngeal cartilages (not listed in Table 1) include the: epiglottis, criocoid, thyroid, cuneiform (left and right) and arytenoid (left and right). Except for the face, all components are symmetrical with respect to the mid-sagittal plane. The model is implemented in Artisynth; we refer the interested reader to Lloyd et al. \cite{17} for details of the forward and inverse mechanics in Artisynth.

The face, tongue, jaw, skull, and hyoid have previously been composed as a unified model \cite{21,7}, and the larynx along with its supporting structures have also been simulated independently \cite{6}; however, here we compose them into a single model including the soft palate and pharynx.

The pharynx geometry was designed by fitting spline surfaces to the approximate contour of the pharynx from cone beam Computed Tomography (CT) data. The constrictor muscles were contoured to smoothly extend toward their respective insertion regions relative to the jaw and cartilage structures of the larynx. The pharynx surface was then smoothly fitted with a quadrangle surface mesh, non-rigidly registered to landmarks of the other VT components, and extruded to form a hex-dominant mesh that fits and deforms with the surrounding articulators, which enables dynamic coupling to our articulatory speech synthesizer. We demonstrate that the biomechanics, in conjunction with the skinning, supports a range from physically realistic to simplified vocal tract geometries to investigate different approaches to aeroacoustic modeling of vocal tract. Furthermore, our model supports inverse modeling to support investigation of plausible muscle activation patterns to generate speech.

\section{3. Results}

Our model supports inverse modeling that elucidates the function of individual components. For example, the face cannot be done experimentally. Simulations may complement experiments to address many of these concerns, and provide a framework for testing hypotheses and making predictions that cannot be done experimentally.

\section{4. Discussion}

Speech simulation involves the simulation of sound production and propagation through the airway until it radiates into free space beyond the lips. The glottis typically provides the main sound source, though sound may also be generated by other sources as in stop consonants and fricatives \cite{9}. Ishizaka & Flanagan \cite{10} modeled the glottis as a system of two masses connected by springs, which has since been challenged and elaborated \cite{11,12}. Most sound propagation models attempt to reduce the airway to a 1D area function and calculate sound if propagating through a tube of varying cross-section \cite{13,14}. The more sophisticated 1D models include the sub-glottal tract, a split airway to include the nasal passage, resonator cavities, sound sources at locations of (modeled) turbulence, and sound modification by the walls \cite{13}. More recent efforts have simulated sound propagation using the 3D wave equation \cite{15} or the Navier-Stokes equation \cite{16}.

In our simulation toolkit Artisynth, we have developed a complete biomechanical model of the upper airway structures, which is coupled to an acoustics synthesizer and may be used to calculate inverse mechanics (given a motion, find the muscle activations) as well as forward mechanics (find the motions caused by muscle activations). In this paper, we present this model and demonstrate how it may be used towards numerous applications in speech production and perception.

\section{5. Conclusion}

Speech production and perception are essential components of daily life for most people; however, the mechanisms involved are extremely complicated and remain poorly understood. Speech research is usually performed experimentally on people; however, measurements may be highly limited to the measurements that can feasibly and ethically be performed on humans and within the instrumentation capabilities. Simulations may complement experiments to address many of these concerns, and provide a framework for testing hypotheses and making predictions that cannot be done experimentally.

Models of the human vocal tract (VT) initially focused on the function of individual components. For example, the face \cite{1}, the tongue \cite{2}, soft palate \cite{3,4,5}, and larynx \cite{6}. Recent progress has been made towards developing a comprehensive model of the VT, as seen in the face, tongue, skull, jaw, hyoid model of Stavness et al. \cite{7} and the tongue and epiglottis model by Pelletier et al. \cite{8}. The model we present in this paper is the most complete that the authors are aware of, including finite element (FE) models of the face, tongue, soft palate, pharynx, larynx, and supporting structures of bone and cartilage.

Speech simulation involves the simulation of sound production and propagation through the airway until it radiates into free space beyond the lips. The glottis typically provides the main sound source, though sound may also be generated by other sources as in stop consonants and fricatives \cite{9}. Ishizaka & Flanagan \cite{10} modeled the glottis as a system of two masses connected by springs, which has since been challenged and elaborated \cite{11,12}. Most sound propagation models attempt to reduce the airway to a 1D area function and calculate sound if propagating through a tube of varying cross-section \cite{13,14}. The more sophisticated 1D models include the sub-glottal tract, a split airway to include the nasal passage, resonator cavities, sound sources at locations of (modeled) turbulence, and sound modification by the walls \cite{13}. More recent efforts have simulated sound propagation using the 3D wave equation \cite{15} or the Navier-Stokes equation \cite{16}.

In our simulation toolkit Artisynth, we have developed a complete biomechanical model of the upper airway structures, which is coupled to an acoustics synthesizer and may be used to calculate inverse mechanics (given a motion, find the muscle activations) as well as forward mechanics (find the motions caused by muscle activations). In this paper, we present this model and demonstrate how it may be used towards numerous applications in speech production and perception.

\section{6. Acknowledgments}

We acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Government of British Columbia.

\section{7. References}

\begin{thebibliography}{9}
\bibitem{1} Pelteret et al. \cite{8}.
\bibitem{2} Pelteret et al. \cite{8}.
\bibitem{3} Pelteret et al. \cite{8}.
\bibitem{4} Pelteret et al. \cite{8}.
\bibitem{5} Pelteret et al. \cite{8}.
\bibitem{6} Pelteret et al. \cite{8}.
\bibitem{7} Pelteret et al. \cite{8}.
\bibitem{8} Pelteret et al. \cite{8}.
\bibitem{9} Pelteret et al. \cite{8}.
\bibitem{10} Pelteret et al. \cite{8}.
\bibitem{11} Pelteret et al. \cite{8}.
\bibitem{12} Pelteret et al. \cite{8}.
\bibitem{13} Pelteret et al. \cite{8}.
\bibitem{14} Pelteret et al. \cite{8}.
\bibitem{15} Pelteret et al. \cite{8}.
\bibitem{16} Pelteret et al. \cite{8}.
\bibitem{17} Pelteret et al. \cite{8}.
\end{thebibliography}
The tongue and palate is shown in Fig. 2. As an anatomical reference. The fitting of the pharynx with correct the uvula for an upright posture (using Gray’s Anatomy) was significantly altered in geometry to smoothly connect to the pharynx model and the musculature described by Gick et al. [19], was significantly altered in geometry to smoothly connect to the pharynx model and correct the uvula for an upright posture (using Gray’s Anatomy [24] as an anatomical reference). The fitting of the pharynx with the tongue and palate is shown in Fig. 2.

<table>
<thead>
<tr>
<th>Component</th>
<th>Type</th>
<th>DOF</th>
<th>Mesh</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>face</td>
<td>FE</td>
<td>8720</td>
<td>hex</td>
<td>[1]</td>
</tr>
<tr>
<td>tongue</td>
<td>FE</td>
<td>948</td>
<td>hex</td>
<td>[2]</td>
</tr>
<tr>
<td>skull,jaw,hyoid</td>
<td>Rigid</td>
<td>6 (x3)</td>
<td>tri</td>
<td>[18]</td>
</tr>
<tr>
<td>soft palate</td>
<td>FE</td>
<td>735</td>
<td>tet</td>
<td>[5, 19]</td>
</tr>
<tr>
<td>pharynx</td>
<td>FE</td>
<td>2436</td>
<td>hex</td>
<td>N/A</td>
</tr>
<tr>
<td>larynx</td>
<td>FE</td>
<td>3034</td>
<td>hex</td>
<td>[6]</td>
</tr>
<tr>
<td>larynx cartilage</td>
<td>Rigid</td>
<td>6 (x7)</td>
<td>tri</td>
<td>[6]</td>
</tr>
<tr>
<td>airway</td>
<td>skin</td>
<td>N/A</td>
<td>tri</td>
<td>[20]</td>
</tr>
</tbody>
</table>

Table 1: Summary of components. Includes component name, component type (either FE models, rigid body, or deformable skin surface), degrees of freedom (DOF), mesh type (hexahedral-dominant, tetrahedral, or triangular surface mesh) and references to source publications.

The superior portion of the pharynx has a thickness of 2.85 mm [22] while the inferior portion has a thickness of 4.0 mm [23]. The soft palate geometry, though originally derived from the geometry of Chen et al. [5] and employing the internal musculature described by Gick et al. [19], was significantly altered in geometry to smoothly connect to the pharynx model and correct the uvula for an upright posture (using Gray’s Anatomy [24] as an anatomical reference). The fitting of the pharynx with the tongue and palate is shown in Fig. 2.

The muscles of the upper airway are approximated as point-to-point hill-type muscles that may attach to a rigid body or to a fixed point in space. Muscles may also pass through a FE body and exert a local force on that body at attachment locations, or they may be embedded in the FE body as transversely-isotropic material properties. There are 11 tongue muscles, 5 palate muscles, 11 face muscles, 6 pharyngeal muscles, and 46 external muscles. These muscle groupings are approximate, as some muscles may pass through multiple bodies (for example, the palatopharyngeus muscle).

The components of the upper airway are coupled by various means. The skull is anchored in space, and provides the primary anchoring for the model. The nodes of the face, palate, and pharynx that connect with the skull are chosen to be attached to the skull, and thus be immobile. An FE model may also attach to a dynamic rigid body, as occurs between the face-jaw, jaw-tongue, pharynx-thyroid, and the larynx with cartilage components. An FE model may attach to another FE model, as occurs with regions between the palate-pharynx, pharynx-tongue and tongue-face. Components may also be connected by point-to-point muscles, as happens, for example, between the crycoid and thyroid.

2.2. Biomechanically Driven Acoustics

Articulatory speech synthesizers generate sound based on the biomechanics of speech. Vibration of the vocal folds under the expiratory pressure of the lungs acts as the input to the system and the VT constitutes a filter where sound frequencies are shaped. We use the two-mass glottal model proposed by [10] coupled with the 1D linearised Navier-Stokes equation described by van den Doel & Ascher [14].

The airway was modelled as a deformable air-tight mesh, referred to as a skin, which is coupled to the articulators as described in [20]. Each point on the skin is attached to one or more master components, which is either a FE node on the face, tongue, soft-palate, pharynx, or larynx model, or a 6-DOF frame such as jaw and hyoid rigid bodies. The position of each skin vertex is calculated as a weighted sum of contributions from each master component. To provide two-way coupling between the skinned mesh and articulators, the forces acting on the skin points are also propagated back to their dynamic masters (as required for fluid-solid interaction).

For the purpose of examining how the acoustic results depend on the initial airway mesh, we define two geometries: an exact airway and a CT-derived airway. The exact airway was manually extracted from the surface geometries of the relevant components and stitched together in Blender, hence it conforms precisely to the surrounding geometries. The procedure was similar to the semi-automated method described by Widing & Ekeberg [25]. The CT-derived airway was extracted from CT data and manually smoothed in MeshMixer to approximately fit the oral airway, but neglects the cavity anterior to the epiglottis, the nasal passage, and does not extend laterally to the teeth or cheeks. The CT-derived airway, being a single tube, yields an obvious area function at the expense of fitting the geometry precisely while the exact airway provides a precise fit but the definition of the area function is less simple. The exact airway is shown in Fig. 3a and the CT-derived airway in Fig. 3b.

A centerline is used to find the cut-planes and hence cross-sectional areas that drive the acoustics model. It is deformed by the same principles as the skin mesh, being attached to master components and deforming as they deform. The initial centerline, which is chosen by hand, is shown in Fig. 3. For any given cutplane, the CT-derived airway yields a single closed polygon representing the cross-sectional area (CSA), which is used to calculate the CSA. However, the exact airway may have complex cross-sections without an obvious CSA. Thus we examine...
two techniques. First, line-of-sight CSA calculates the CSA as the region that is visible to the centerline vertex defining that cutplane. Second, contour CSA keeps the entire polygon containing the centerline vertex while subtracting out of the area any “islands” that may intersect the area such as may be caused by the uvula or epiglottis. The contour CSA gives a precise CSA, but at the risk of including cavities of the airway that have little involvement in the airflow.

Figure 3: Images demonstrating the two airway skin-meshes used to drive the acoustics: a) shows the exact airway that closely fits the VT geometry, while b) shows the tube-like geometry that is suited to an area-function approach. The centerline is shown in green.

3. Experiments and Results

We now demonstrate the functionality of our VT model with forward and inverse simulation scenarios.

3.1. Biomechanically Driven Acoustics

We defined 3 sets of muscle activations, corresponding to the vowels /a/, /I/ and /u/. Below we list the muscles involved followed by the percent activation in parenthesis.

/a/ Formed using tongue and jaw muscles. Tongue: genioglossus middle (0.08), anterior (0.12); hyoglossus (0.15); verticalis (0.05). Jaw opening: anterior and posterior digastric, sternohyoid (all at 0.1) [26].

/I/ Formed with tongue, face, and jaw muscles. Tongue: genioglossus posterior (0.5), middle (0.02), anterior (0.02); superior longitudinal (0.05); interior longitudinal (0.05); mylohyoid (0.10). Face: risorius (0.05) and zygomaticus (0.05). Jaw closing: temporalis (anterior, middle, posterior), masseter, median pterygoid (all at 0.01) [26].

/u/ Formed using tongue and face muscles. Tongue: styloglossus (0.2). Face: orbicularis oris middle ring (0.35) [21].

The model is allowed to settle under gravity for 0.1 s, at which point the muscles are linearly increased to their maximum level at 0.3 s, and again the model settles without further changes until 0.4 s at which point the acoustic results are calculated.

Under these muscle activations, the upper airway structures deform. The exact and CT-derived airway meshes are then used to calculate area functions which are fed into the acoustics model. The area functions are shown in Fig. 4, and the resulting formants in Fig. 5.

The area function plots of Fig. 4 demonstrate that the CT-derived tube and the exact tube have dramatically different area functions, though usually similar in trend. Likewise, line-of-sight CSA calculation and contour CSA calculation also result in major differences at some locations. The differing area functions lead to differing acoustics as well, which is reflected in Fig. 5.

3.2. Inverse Modeling

To demonstrate the inverse modeling capabilities, we seek to re-create /a/ by supplying the inverse simulation with 1 target point on the jaw and 5 target points on the tongue, all on the mid-sagittal plane. The motion of these target points was captured by the forward simulation, and the inverse simulation seeks to re-create the activations used to attain the posture.

The forward simulation activations are compared with the inverse-predicted activations in Fig. 6. We see that the inverse simulation re-creates the original activations with good fidelity. Supplying additional target points, including points off of the mid-sagittal plane, would further improve the quality of the inverse prediction. The posture predicted by the inverse simulation is the same as shown in Fig. 4a.

4. Discussion

Most existing articulatory synthesizers use a mid-sagittal representation of the vocal tract articulators [27]. This 2D representation is common, since it can be readily derived from mid-
airway geometry, and were unstable at locations of airway clo-

calculations depend on the centerline staying entirely within the

derived airway; in contrast, both line-of-sight and contour CSA

area function is a simple and stable procedure with the CT-

putational expense.

that have small influence on the acoustics, thus reducing com-

the same underlying dynamic articulations. For 3D fluid simu-

tailed to the specific acoustics synthesizer, while maintaining

phonetic contexts (and particularly dynamic contexts) for which

3D VT synthesis is the most promising direction for achiev-

is a more physiological approach, and we feel that pursuing

sired acoustics. While more challenging, the fully 3D approach

naturally by the model, rather than be tuned to achieve the de-

significant challenge of moving to a 3D biomechanically-driven

airway model is that the airway width now must be produced

desired airway CSA for each articulatory configuration. A sig-

ics, i.e. the width of the airway tube is prescribed to achieve the

derived airway.

The forward simulation takes approximately 2.6s per time

step on an 8-core 2.67GHz PC with 12 GB RAM, and thus sim-

ulating 0.4 s (80 time steps) requires 3.5 minutes. In contrast,

the inverse simulation takes about 7.85 s per time step, and thus

simulating 0.4 s requires 10.5 minutes.

4.1. Limitations and future work

To further improve the simulation quality, the soft palate

should be remeshed to have a hexahedral mesh, and the tongue

remeshed to have finer elements. In future work, we aim to

demonstrate comprehensive motions including pharyngeal and

soft palate muscles.

We intend to include all the features of the Birkholz model

[13] in our 1D acoustics model, and also to investigate 3D mod-

eels. We also aim to generate sound by driving our larynx model

with a fluid model, then moving beyond a mass-spring sound

source model.

5. Conclusions

In this paper, we have presented our biomechanical model of the

human upper airway, including 5 FE bodies representing soft

tissue, 10 rigid bodies representing bone and cartilage, point-to-

point muscles to describe the musculature of the upper airway,

and airway meshes that deform in keeping with the surrounding

structures. This model is capable of sophisticated speech mod-

eling, including biomechanically-driven speech synthesis and

inverse simulation to predict muscle activations given a kinem-

atic motion target.

Using this model, we have simulated the articulation of the

vowels /al, /el, and /u/ driven by muscle activations, and com-

pared the acoustic results considering exact or idealized airway
descriptions. This comparison highlights issues that arise with

1D acoustics models as we strive towards biomechanically-

driven acoustics: an airway that precisely conforms to the VT

gometry does not have a unique area function, whereas a tube-

like airway geometry that readily yields an area function does

not completely represent the underlying structures.

We have also performed inverse simulations, in which we

provide the kinematic motion desired for the model, and sim-

ulate muscle activations needed to attain that motion. For the

purpose of this demonstration we use motion targets for which

we know the muscle activations, but in general inverse simul-

ation provides a powerful tool to determine which muscles are

used to achieve at motion.

6. Acknowledgements

The authors would like to thank Alan Hannam, Bill Pearson

and Bryan Gick for their assistance with the airway geometry

and muscle design; and Antonio Sanchez for assisting with the

inverse simulation.
7. References


