Towards Speaker-specific Biomechanically-driven Articulatory Speech Synthesis

Victor Zappi[1], Arvind Vasudevan[2], Keyi Tang[2], Negar M. Harandi[2], Sidney Fels[2]

[1]*Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy*
[2]*Department of Electrical and Computer Engineering, University of British Columbia, Canada*
victor.zappi@gmail.com

Abstract

We describe our progress on speaker-specific biomechanically–driven articulatory speech synthesis. Starting from a modular biomechanical Functional Reference ANatomical Knowledge (FRANK) template, we register FRANK to medical image data of the subject whose voice we want to synthesize. The morphed FRANK model contains an airway mesh that changes shape as the simulated muscles activate to articulate different vocal sounds, based on inverse dynamics using dynamic MRI data. The acoustics of the biomechanical model is computed by means of a 2D finite difference time domain wave solver, coupled with vocal fold models, that allow the synthesis of the sounds produced by the simulated articulation. In this work, we present results on the effectiveness of the used morphing and inverse modeling techniques, as well as on the precision of the acoustic simulation.

Keywords: articulatory speech synthesis

1. Introduction

Speaker-specific, biomechanically-driven, articulatory speech synthesis continues to be a challenging research direction. In this paper, we present our approach to address some of the challenges such as (1) creating a functional biomechanical model of the speaker that can reproduce their speech motion, (2) simulation of glottal excitation based on airflow and tissue properties of vocal folds, and (3) simulation of aeroacoustics of the complex, dynamic 3D vocal tract.

2. Biomechanical Modeling

A modular biomechanical model of the head and neck – referred to as a Functional Reference ANatomical Knowledge (FRANK) template – has been implemented in the ArtiSynth simulation framework (Anderson, Harandi, et al. 2015).

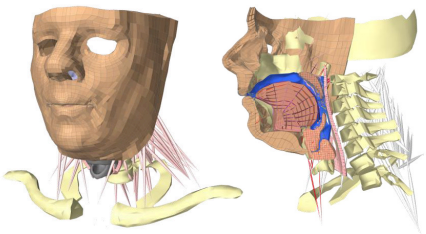

Figure 1: *FRANK model from the front (left), and cross-section (right). Airway mesh is shown in blue.*

As shown in Figure 1, FRANK consists of multiple components governed by a hybrid physical simulation technique that combines multibody physics with 3D finite element analysis. As a generic 3D model consisting of the soft-tissue and bones, FRANK has been used for simulating simple speech postures. This is done either by activating the embedded muscles, or by using an optimization technique to estimate those activations based on desired motion. Our current speaker-specific modelling methods are described in a three-step sequential procedure: **(1) Establishing the correspondences:** We use a modified version of the extrinsic Iterative Closest Point method to find the correspondences between the surface meshes in the template, and the ones we segment from the medical images of the speaker. We limit our deformation field to be As-Similar-As-Possible to maintain the morphology of the template. **(2) Anatomy transfer:** Subject to the correspondence constraints, we then transfer sub-groups of FRANK components to the speaker space using a topology-preserving deformation. In each sub-group, the mapping function preserves the spatial relationship between the source components and maintains their regularity. **(3) Functionality transfer:** Finally, we transfer the functional information of FRANK, such as configuration of the muscles and joints, to the registered speaker-specific meshes. Our methods are automatic, removing the need for additional manual efforts. Figure 2 shows a speaker-specific model obtained by registering FRANK to a Computed Tomography (CT) image volume. Three speech postures – for the vowels **/a/**, **/i/** and **/u/** – were simulated in response to three sets of muscle activations defined by Anderson, Harandi, et al. (2015).

3. Real-time Vocal Tract Solver

We use a 2D wave solver to compute the aero-acoustics of the postures simulated in FRANK. Based on a 2D finite difference time domain (FDTD) scheme, the solver runs on the GPU and allows for the fast simulation of pressure wave propagation within a chosen domain, achieving under specific condi-

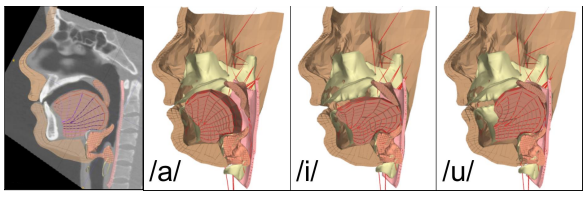

Figure 2: *Cross-section of a speaker-Specific model overlaid on the CT image (right), and after simulating the vowels (left).*

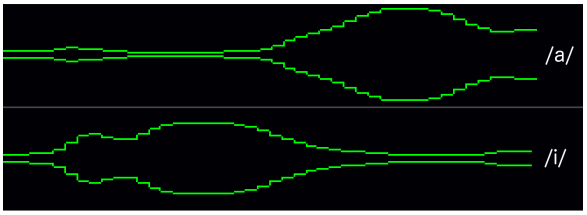

Figure 3: *2D contours extracted from an /a/ vowel posture (top) and an /i/ vowel posture, with a resolution of 0.56 mm.*

tions real-time performances (Zappi et al. 2016).

Once the biomechanical model is registered and the posture is obtained, the resulting airway mesh (Figure 1 on the right) is extracted as an *area function*. This works as a simplified representation of the vocal tract, which still maintains the fundamental acoustic properties of the original mesh, but allows for a less computationally expensive simulation. Finally, the area function is turned into a 2D contour (Figure 3), i.e., the final simulation domain where the FDTD scheme is applied.

The vocal sound resulting from the simulated posture is synthesized by exciting the domain with a glottal waveform, outputted by the vocal fold model described in the next section. The pressure variation at the mouth opening (the right end of the contours in Figure 3) is then sampled to produce an audio stream, that can also run in real-time. Other boundary conditions are enforced on the vocal tract; we use local reactive boundaries to characterise the vocal tract's walls with an absorptive behavior, while mouth radiation at the end of the tube is simulated by the employment of Perfectly Matched Layers.

We tested the precision of the overall acoustic simulation by computing the impulse responses of a set of standard area functions, measured on a real subject by Story (2008). The vocal fold model was detached from the system and the 2D contours were excited using a frequency-limited impulse. We compared the positions of the first 8 formants extracted from the impulse responses of the 2D vocal tracts with the ones obtained with a highly precise and computationally expensive 3D finite element model (Zappi et al. 2016). Results show a very good match with the 3D data, with most positional errors below 1%.

4. Vocal Fold Models

The 2D acoustic simulation is excited by a train of glottal pulses that feed into the laryngeal cavity of the resonant vocal tract (leftward entrace of the contour in Figure 3).The supraglottal pressure feedback from the vocal tract in conjunction with a defined subglottal pressure, drives the self-oscillating vocal fold vibration. We first choose to couple computationally cheaper lumped-element models to the solver, directly implemented on the GPU shader. This enables capturing of non-linear source-filter coupling while running at real-time simulation rates. We couple the system with the body-cover model (Story and Titze 1995), which divides the vocal fold into three lumped-mass elements in the coronal plane. This model allows us to represent the material property differences between the 'body' and 'cover' layers of the vocal folds.

To achieve high-quality articulatory speech synthesis, we require models that can accurately represent the specific vocal fold structure, driven by the aerodynamic pressures. We propose a novel 2D continuum model of the vocal folds that combines a FEM structural solver (Alipour et al. 2000) loosely coupled with a continuous 1D flow model (Anderson, Fels, et al. 2013). The model can be driven by both velocity and pressure boundary conditions, and is capable of robustly handling tube closures. Through this model we aim to achieve a balance between FEM structural models driven by lightweight Bernoulli-based fluid models that make generous steady-flow assumptions, and accurate but computationally expensive 2D Navier-Stokes simulations. This makes high quality speech synthesis possible at significantly lower computational costs, when coupled with the real-time GPU acoustic solver.

5. Conclusion

We continue to improve each component of the workflow to enable speaker-specific biomechanical modeling of articulatory speech synthesis. Research addressing each of the individual challenges posed by the longer term goal of articulatory speech synthesis improves the overall quality of articulatory speech enabling applications in communications, medicine and entertainment.

As the functionality of the FRANK model improves, so does our speaker-specific models. Using registration methods that work directly with MRI images (rather than surface meshes) would eliminate the burden of segmentation. We look to improve the generated voice quality through coupling our acoustic model with more detailed continuum vocal fold models and building more realistic vocal tract's contours.

6. Acknowledgements

This work is funded by Natural Sciences and Engineering Research Council of Canada (NSERC), NSERC-Collaborative Health Research Project (CHRP) and by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme.

7. References

Alipour, Fariborz, David A Berry, and Ingo R Titze (2000). "A finite-element model of vocal-fold vibration". In: *The Journal of the Acoustical Society of America* 108.6, pp. 3003–3012.

Anderson, Peter, Sidney Fels, and Sheldon Green (2013). "Implementation and validation of a 1D fluid model for collapsible channels". In: *Journal of biomechanical engineering* 135.11, p. 111006.

Anderson, Peter, Negar M Harandi, Scott Moisik, Ian Stavness, and Sidney Fels (2015). "A comprehensive 3D biomechanically-driven vocal tract model including inverse dynamics for speech research". In: *Interspeech 2015*. Dresden, Germany.

Story, Brad H (2008). "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002". In: *The Journal of the Acoustical Society of America* 123.1, pp. 327–335.

Story, Brad H and Ingo R Titze (1995). "Voice simulation with a body-cover model of the vocal folds". In: *The Journal of the Acoustical Society of America* 97.2, pp. 1249–1260.

Zappi, Victor, Arvind Vasuvedan, Andrew Allen, Nikunj Raghuvanshi, and Sidney Fels (2016). "Towards real-time two-dimensional wave propagation for articulatory speech synthesis". In: *Proceedings of Meetings on Acoustics 171ASA*. Vol. 26. [in press]. ASA.