INTERDISCIPLINARY APPROACHES FOR ADVANCING ARTICULATORY SPEECH THEORY AND SYNTHESIS

Sidney Fels[1] and Bryan Gick[2, 3]

[1] Dept. of Electrical and Computer Eng., University of British Columbia, 2356 Main Mall, Vancouver, BC, Canada, V6T1Z4
[2] Dept. of Linguistics, University of British Columbia, 2613 West Mall, Vancouver, BC, Canada, V6T1Z4
[3] Haskins Laboratories, 300 George St., New Haven, CT, USA 06511

1. INTRODUCTION

Articulatory synthesis research has long been dominated by frequency domain and concatenate sample-based speech synthesis techniques. While successful in some domains (e.g., voice-based databases), these techniques still cannot produce natural looking and sounding speech from text for an arbitrary speaker. Natural looking and sounding speech technology is one of the next major milestones in voice-based interaction for natural user interfaces. Through a team of interdisciplinary international researchers, we have been steadily working towards creating the necessary platform to overcome basic problems in synthetic speech production. Our approach is three fold: 1. Create a Functional Reference for Anatomical Knowledge (FRANK) template model for use in studying articulatory synthesis, 2. Develop new methods for voice synthesis that couple to the platform, and 3. Use the biomechanically driven articulatory synthesizer to explore and develop theories of speech production. Our team draws from electrical, computer and mechanical engineering, linguistics, computer science, dentistry, radiology and surgery to assemble the needed components for advancing the state of the art.

2. The FRANK Model

We have created a modular biomechanical model of the head and neck referred to as a Functional Reference ANatomical Knowledge (FRANK) template [3]. It consists of multiple components governed by a hybrid physical simulation technique that combines multibody physics with 3D finite element analysis. Figure 1 shows an illustration of some of the components of FRANK. Components for FRANK have evolved as we have completed different experiments on speech [2][3][4][6][9][10][11][12][19][20][21][14], swallowing [15] and mastication [13] By combining the models from these studies and tailoring them to fit well together, we generate a generic 3D biomechanical model of the muscles and bones of the human head and neck suitable for modeling the biomechanics of speech.

FRANK is built using a biomechanical modeling toolkit called ArtiSynth [6][18] that has been specifically designed to support modeling of human anatomy and biomechanics. ArtiSynth is an open-source[a] platform that allows for mixed multibody and FEM simulation, supporting both contact and constraints. This enables efficient simulation of a large number of connected dynamic hard and soft tissues such as those involved in the upper airway.

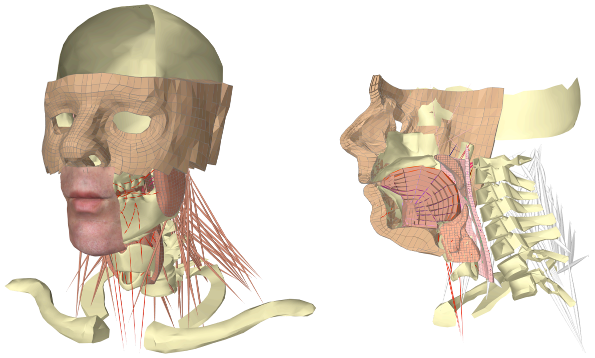

Figure 1: Overview of FRANK: oblique view showing external structures and midsagittal view showing some of the FEM soft tissues. FRANK includes an airway surface mesh connecting to the tissue models to provide an air-tight tube suitable for acoustic simulation.

When used in an inverse simulation setting, kinematic data can drive FRANK, as the model attempts to estimate the muscle activations needed to control actions when speaking. The template model uses a combination of FEM, rigid body, and spring-like components, allowing its functional resolution to be tailored to the needs of specific applications by trading off between speed, complexity and accuracy. By making FRANK modular, FEM elements can be introduced where high fidelity is required, while simpler, reduced degree-of-freedom models can be used where low-fidelity approaches provide sufficient scaffolding or accuracy, such as with line based muscles or rigid body bone models.

3. Acoustic Simulation

FRANK supports modeling of the 3D biomechanics of the upper airway. To be consistent with our approach to modular methods for simulation, we are developing multiple synthesis modules to combine with FRANK. Currently, we support 1D source filter models [17] with calculations of the acoustic transfer function of the vocal tract using JASS [23]. For the glottis, we are comparing the classic two mass model [16] with higher degree of freedom models of the vocal folds [22][5]. We are also investigating a 2D finite difference approach [24] based on [1]. Thus far, our comparison illustrates how mapping the 3d structure of the

[a] ArtiSynth is available for research use at http://www.artisynth.org.

vocal tract to lower dimensional representations for acoustic simulation results in poor estimates of frequency response. However, low dimensional representations of frequency response can be tuned to give good performance; though, the connection to physical geometry is compromised.

4. Speech Research Using FRANK

Model simulation using ArtiSynth coupled with speech production data show that the body offers only a small inventory of reliable, biomechanically robust actions for speech sounds, and that these postures have special biomechanical properties, whether viewed at the lips, the tongue, the larynx, the oropharyngeal isthmus, or the velopharyngeal port [2][9][10][12][20]. These findings align with research in neurophysiology and computation of motor control in which movement is built upon inventories of semi-closed neuromuscular structures, or "modules". We have developed a modular theory for speech (e.g., [8][7]) whereby these modules emerge through use as part of a learner's strategy to optimize the biomechanics of speech production, building on phylogenetically encoded motor structures [19]. This model provides robust and highly predictive results with no brain, no experience, and no anatomically pre-defined body parts/articulators (e.g., "lips", "jaw", etc.). The result is a theory of embodied phonetics built on an inventory of emergent, highly specialized and reliable functionally defined body structures that are harnessed for communication, each of which is used to serve a specific phonetic function (e.g., lip rounding, vocal fold approximation).

5. DISCUSSION AND CONCLUSIONS

Our large interdisciplinary team continues to create a platform for speech research using a 3D biomechanical articulatory speech synthesizer. By including team members with expertise in all areas related to speech we can improve each aspect and corroborate results using assumptions made to help capture the complexity of speech production. To date, our FRANK model incorporates over seven years of different modeling efforts and improvements of physics simulation techniques. As the effort is open source, we invite researchers to participate in this effort to continually add modules and improvements to the platform.

REFERENCES

[1] Allen A. and Raghuvanshi, (2015) N., Aerophones in Flatland: Interactive Wave Simulation of Wind Instruments, ACM ToGS (SIGGRAPH), 34(4).

[2] Anderson, P., Fels, S., Stavness, I. and Gick, B. (2016). Intrinsic and extrinsic portions of soft palate muscles in VP and oropharyngeal constriction: A 3D modeling study. *Can. Ac*.

[3] Anderson, P., Harandi, N. M., Moisik, S., Stavness, I. and Fels, S. (2015). A comprehensive 3D biomechanically-driven vocal tract model including inverse dynamics for speech research. *Interspeech 2015*. Dresden, Germany.

[4] Derrick, D., Stavness, I. & Gick. B.(2015).Two phonological segments, one motor event: Evidence for speech-motor disparity from Eng flap production. *J. Ac. Soc. Am.* 137, 1493-1502.

[5] Fariborz, A., DBerry, D. and Titze, I. (2000). A finite-element model of vocal-fold vibration, *JASA* 108.6: 3003-3012.

[6] Fels, S., Gick, B., Jaeger, C., Vogt, F. and Wilson, I. (2003). User-centered design for an open source 3-D articulatory synthesizer. In M. J. Solé, D. Recasens & J. Romero (eds.) *Proc. XVth Int. Congr. Phonet. Sci.*, Barcelona, Spain. 179-184.

[7] Gick, B and Stavness, I. (2013). Modularizing speech. *Front. Psych.: Cog. Sci*. 4, 977.

[8] Gick, B. (2016). Ecologizing dimensionality: prospects for a modular theory of speech production. *Ecol. Psy*. 28(3), 176-181.

[9] Gick, B., Allen, B., Stavness, I. (in press) Speaking tongues are actively braced. *J. Speech, Lang. & Hear. Res*.

[10] Gick, B., Anderson, P., Chen, H., Chiu, C., Kwon, H. B., Stavness, I., Tsou, S., Fels, S. (2014). Speech function of the oropharyngeal isthmus: a modelling study. *Comp. Meth. in Biomech. and Biomed. Eng.: Imag. & Visualization*, 2, 217-222.

[11] Gick, B., Stavness, I. and Chiu, C. (2013). Coarticulation in a whole event model of speech production. Proc *JASA* 19. 060207, 4.

[12] Gick, B., Stavness, I., Chiu, C. and Fels, S. (2011). Categorical variation in lip posture is determined by quantal biomechanical-articulatory relations. Can. Acoust. 39. 178-179.

[13] Hannam, A.G., Stavness, I., Lloyd, J.E., Fels, S.S., Curtis, D. and Miller, A. (2010). A comparison of simulated jaw dynamics in models of segmental mandibular resection versus resection with alloplastic reconstruction, JProsthetic Dentistry, 104(3):191-8.

[14] Harandi, N., Woo, J., Stavness, I., Stone, M., Fels, S. and Abugharbieh, R. (2015). Subject-Specific Biomechanical Modelling of the Oropharynx: Towards Speech Production, May:1-11, CMBBE:I&V.

[15] Ho, A., Affoo, R., Nicosia, M., Inamoto, Y., Eichii, S, Green S., and Fels, S. (2016). Computer Simulation of a Liquid Bolus for Studying Hyposalivation from Dynamic 320-slice CT images, Proc. of the Dysphagia Research Society (DRS2016), Feb 25-27.

[16] Ishizaka K and Flanagan, JL (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell system technical journal*, vol. 51, no. 6, pp. 1233–1268.

[17] Kelly JL. and Lochbaum, CC, (1962). "Speech synthesis," in *Proc. Fourth Int. Congr. Acoust.*, Paper G42, pp. 1-4.

[18] Lloyd, J. E., Stavness, I. and Fels, S. (2012). Artisynth: A fast interactive biomechanical modeling toolkit... In Soft tissue biomech modeling for comp assisted surgery:355-394. Springer.

[19] Mayer, C., Roewer-Despres, F., Stavness, I. and Gick, B. (in press). Does swallowing bootstrap speech learning? *Can. Ac.*

[20] Moisik, S. and Gick, B. (2013). The quantal larynx revisited. *J. Ac. Soc. Am. – Proc. Meet. Acoust.* 19. 060163, 8 pp.

[21] Stavness, I., Gick, B., Derrick, D. and Fels, S. (2012). Biomechanical modeling of English /r/ variants.*JASA*.131:355-360.

[22] Story, B, and Titze. I. (1995). Voice simulation with a body-cover model of the vocal folds, *JASA* 97.2:1249-1260.

[23] van den Doel K. and U. Ascher, U., (2008) "Real-time numerical solution of webster's equation on a nonuniform grid," IEEE Trans. ASLP,16(6):1163–1172.

[24] Zappi, V. and Vasudevan, A. and Fels, S., Towards real-time two-dimensional wave propagation for articulatory speech synthesis, JASA: 139: 2010-2010 (2016).

ACKNOWLEDGEMENTS

The work is funded by NSERC, CIHR (CPG-121024), Autodesk Research and nVidia. We thank: ArtiSynth team, Oral, Pharyngeal and Laryngeal modeling team, Swallowing Research Group, Speech Synthesis team, and the Parametric Human Project.