Extensible infrastructure for a 3D face and vocal-tract model

Florian Vogt[†‡], **S. Sidney Fels**[†], **Bryan Gick**[†*], **Carol Jaeger**[†], **Ian Wilson**[†]

[†] University of British Columbia, Canada
[‡] ATR, HIS Laboratory, Japan
[*] Haskins Laboratory, USA

ABSTRACT

We describe an architecture for a combined 3D face and vocal tract animation simulator for articulatory speech synthesis. The architecture provides five main modules: 1. a simulator engine, 2. a 3D geometry module 3. a graphical user interface (GUI) module, 4 a synthesis engine and 5. a numerics engine. Elements of the model are specified using nodes placed hierarchically in a scene graph. Traversal of the nodes in the scene graph by the simulator engine creates the animation and drives the articulatory synthesis.

Part of the motivation for the structure of the architecture is the recognition that many researchers have done extensive research on separate aspects of the problems of vocal tract and face modelling in addition to speech synthesis based on articulation. Our architecture is meant to facilitate combining models of different structures and levels of detail from different research groups easily providing a testbed for articulatory based speech research and production. Our ultimate aim is to have a fully functioning 3D vocal tract model that uses aeroacoustic models to produce speech.

1 INTRODUCTION

Currently, many researchers in articulatory speech synthesis are working on different substructure models of the vocal tract, for e.g. tongue, larynx, lips, and face. Each of these are very complex and are often developed independently of other structures. Since the integration of separately developed models is very time consuming, comparisons between different modeling approaches have almost never been performed. Our research focusses on creating an articulatory speech syntheizer that allows easy integration of different parts so that different vocal tract models can be easily compared both geometrically and acoustically.

Our speech synthesis project aims to accomplish the following: (1) to formulate a novel framework for articulatory speech modeling, (2) to implement a core simulator for dynamic 3D vocal tract models to synthesize speech, (3) to create a data-driven modeling and validation integration facility.

This core simulator implements all framework modules in order to create a physically based dynamic 3D face and vocal tract model. The starting point of the core simulator is the mass-spring model based facial animation system of [13]. In addition, vocal tract shapes can be driven from image data sources.

The long term goal of this project is to develop an articulatory speech synthesizer which can form the basis for an open collaborative system to produce natural sounding speech. Since it is not clear which modeling approach achieves the long term goal, we provide a simulation framework for developing and evaluating different modeling approaches. Starting with a core simulator, some functionality is provided which does not properly solve the long term goal in one step. However, the modular framework design extends the core simulator with, e.g., a full aerodynamic simulation to model the air flow though the 3D vocal tract model with aeroacoustics methods. This and other extensions to the core simulator, which can be provided by members of the speech research community, tackle the long term goal.

2 RELATED WORK

A large body of work in articulatory speech synthesis can be roughly broken down into vocal tract structure models (including source), acoustic production and articulator data extraction. In our system we plan to support many of the works that have appeared in the literature. Our default configuration will come with a fully functional articulatory speech synthesizer. For this section, relevant work for developing the default configuration is discussed.

2.1 VOCAL TRACT STRUCTURE MODELS

There are several representations for the vocal tract geometry. These have been either 2D models, 3D models, and parametric representations. The excitation provided by the larynx has also been modeled in various ways. The main modeling techniques available for modeling the geometry include: spring-mass models, finite element models, boundary element models and parametric models.

Traditional 1D source models approximate the vocal tract shape as a sequence of about 10 to 15 tubes, each with a different cross-sectional area [5]. From

ISBN 1-876346-48-5

this model, the area functions of the vocal tract are used to calculate its transfer function [7] for synthesis. Many researchers use 2D shape models to determine the acoustic properties of the vocal tract, including [16, 4]. [16] defined six articulatory parameters for identifying the mid-sagittal cross sections of the vocal tract. From these six parameters, a set of tube areas are calculated giving the transfer function of the vocal tract. This articulatory synthesis model is effectively a sequence of 2D segments, and ignores many important properties of the vocal tract, such as a parallel nasal tract, the non-linear effects of radiation through the vocal tract walls, and the 3D shape of the vocal tract. Further, of particular relevance to our work is research on 2D articulator geometry [14, 20]. 3D models have been recently used to describe the vocal tract model shape and the dynamics of the articulators. E.g., [2, 6] modeled the tongue using mass-spring models, and represented the remaining vocal tract with 3D geometry. Ultimately, these parameters are still used to determine an area function so that speech can be synthesized. [22] also modeled 3D tongue movement using finite element models based on MRI reconstructions.

Much research has been done on various modeling aspects of the vocal tract fricatives [11], vocal tract dynamics or tongue motion [19, 6], nonlinear time-variant Webster equation (quasi 1D wave) [7, 9], and transmission line models [5].

Our focus at this point is to extract the 3D geometry of the vocal tract to form a geometric mesh. The mesh will attach to fixed reference points of the skull. Underneath the mesh will be a spring-mass model for the tissue along the lines of [13, 12]. The mesh will be partitioned into functional components including: lips, tongue, cheeks, pharynx, soft palate, hard palate, and hyoid. The skeletal structures include: teeth, jaw, skull, hyoid bone and other bony structures. By providing an interface to our system based on the geometric mesh compatible with spring-mass models such as [13] we will be able to easily provide articulatory speech synthesis that includes faces.

2.2 GLOTTAL MODEL

We plan to provide typical glottal models common in the research in the default synthesizer configuration. Overall, much speech research focuses on glottis models as a source. The most influential work models the glottis as an externally driven two mass-spring system [10] which we intend to use. Other airflow models use volume velocity representations since they translate easily into area functions for 1D models.

Instead of modeling, there are other ways to represent the glottis in a simulator: (1) recorded data from EGG or high speed video analysis [17] can be used as a real source, (2) analytical waveforms, such as [15] or switched noise/pulse train [1]. Overall, there are a variety of glottal representations, and it is hard to say which glottis model is generally better for vowel and consonant production. The proposed framework allows for the investigation of the tradeoffs between the different glottal models in a complete vocal tract model context. Our default simulator will provide for simple acoustic waveform excitation [15] as well as the driven two mass-spring system of [10].

2.3 ACOUSTIC OUTPUT

Modeling acoustic phenomena from air flowing through the vocal tract requires integrating results from aerodynamics with acoustical properties of a soft walled, time varying tube that has multiple, moving obstructions, some of which vibrate. One of the main goals for this research is to allow researchers to look at different ways of modeling speech output from the vocal tract. Thus, we intend to provide considerable flexibility for introducing different aeroacoustic techniques such as [21, 18]. However, for the default configuration, we will supply a first-order approximation [5, 16] for sound production.

2.4 DATA DRIVEN MODELS

Finally, an important component in articulatory modeling is the inclusion of measured data from medical imaging and other vocal tract instrumentation. This is one of the key advantages of an articulartory synthesis approach since we have an actual speech producing mechanism we can use for our reference as researchers have shown [25, 23]. Thus, it is important to allow integration of data directly into simulations for the verification and validation of models.

Since vocal tracts can not be captured in their spatial, temporally required resolution for speech with a single medical image source, a variety of formats are desirable for support. A combination of visual tracking data, magnetic tracking data, video, ultrasound, X-ray and MRI data, as well as acoustic and artificial data from animation software are potential data sources.

3 VOCAL TRACT SIMULATION FRAMEWORK

Designing large-scale software systems such as an articulartory speech synthesizer fits well within the scope of classical software engineering. Given the various methods, this simulator easily becomes a large scale project, which requires new techniques and representations in order to become an accepted frame work in the speech community. One important step in the design process is the development of representations for data and methods. These can be structured in a layered, object oriented, or hierarchical software model. At this stage, we plan to use a *scene graph* framework for representing our models. Additionally, as researchers' needs are quite varied, typically a user-centred design approach

ISBN 1-876346-48-5

is suited for developing the interface [8]

Scene graphs are mainly directed tree structures which contain heterogeneous nodes, and which include data structures and method handling. They provide a metaphor of the geometric structure for nodes. If a directed graph is not sufficient for the representation of a problem then the directed graph can be replaced with a general graph with some loss in efficiency. A Scene graph appears to be a suitable representation of the physical simulation of the vocal tract, even if its not very common. For this project, Inventor [24] is chosen as a scene graph library mainly since it is well established in the graphics community . Our system architecture needs to be designed to support a variety of cross-compatible modeling techniques and several modeling representations. During the design process we decided that a geometric mesh will be used to represent the common modeling layer. Further, our design is modular in order to accommodate the different requirements of the targeted application areas. Modular systems are easier to adopt by others for their purposes. The architecture consists of five central modules shown in Figure 1.

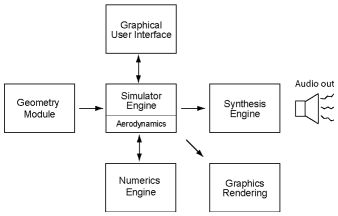

Figure 1: Simulation framework.

3.1 SIMULATION ENGINE

All render processes run infinite loops by traversal of the nodes in the scene graph using the simulator engine. Each loop cycle generates a state vector to create animations, to calculate the updates for the next time step, and to drive the articulatory synthesis.

Anatomical structures in models are represented by nodes placed hierarchically in a scene graph. E.g., a node hierarchy might consist of the face and vocal tract as a top-layer. These nodes consist of lips, skull, teeth, tongue, cheeks, pharynx, and nasal tract in the intermediate layer. Each of this nodes are represented by muscles, tissues, bones, and ligaments in the bottom layer. Attributes and methods of each structure are defined on a suitable level in order work sufficiently.

The core simulation engine is built on a high-level graphics library Inventor [24], which supports scene graph structures for graphics and animations. Created from data and analytic methods are nodes for specifying graphical models, including shapes, cameras, lights, properties, transformation, engines, selection, view, and so forth. Inventor allows the extension of its class structure to build nodes for physically based simulations and further handle multiple looping constructs for graphics, physical simulations and sound.

3.2 GEOMETRY MODULE

The geometry module is another key component in the architecture handles the import of various data sources. This module allows the manipulation of the state space for animation data by: (1) a base pose geometry import, (2) key frame animation for articulatory gestures, and (3) validation through geometric data analysis. Additional data originates from video cameras, ultrasound scanners, Magnetic Resonance Imaging Scanners (MRI), visual trackers and sound files.

3.3 GRAPHICAL USER INTERFACE (GUI)

Many of the users that the simulation architecture is intended for, are not primarily programmers. This is one property which is shared with computer game engines. Therefore, it is important to follow human computer interface and software engineering principles during the design and implementation phase. Design paradigms which improve system acceptance are a flexible GUI interface, portability, script-ability, modularity, and minimal dependencies for reducing complexity. Therefore, the core simulation module is separated from the user interface.

3.4 SPEECH SYNTHESIS ENGINE

In general, a speech synthesis engine outputs acoustic speech based on the input, which is a state vector of the shape. There is a wide range of synthesis techniques discussed in § 2.3.Three are implemented with the intent that they will validate the flexibility and modularity of the engine design, as well as provide a basis for the other contributions of the work.

3.5 NUMERICS ENGINE

Most physical-based simulations have the numerical processes as a bottle neck. The numerics engine separates the numeric algorithms from the simulation engine to allow for flexibly in the implementation. This allows for central numerical optimization and alternative implementations that is, implicit versus explicit Euler integration.

4 SUMMARY

Research in articulatory speech synthesis offers many promising approaches, however, a method of evaluating inherent tradeoffs between research developments is still missing. Moreover, this tradeoff evaluation method is the key component to make successful advances in the field. We hope that our proposed extensible infrastructure provides the glue necessary for making these advances. Three main contributions to the speech research community expected are: (1) the creation of a novel framework which combines existing

ISBN 1-876346-48-5

and future approaches for articulatory speech synthesis and facial animation into one system; (2) the implementation of a core simulator, which validates the framework and its components; and (3) a data-driven modeling component, which augments and verifies the modeling techniques and parameters with "real data" for perceptual testing methods.

The benefit of our software engineering approach for constructing an open-source 3D articulatory speech synthesizer is that the research community will be able to validate and verify their work in a common framework. As part of our process, we are soliciting contributions from colleagues to ensure that the system will meet needs. The advances possible from this approach are based on two main functions that will be provided: (1) simulating deformable models of vocal tract anatomy to output an animated geometric airspace and (2) synthesizing the acoustic speech signal from animated geometry. Thus, components can be added in to the overall model and tested without the need for each researcher to implement a complete vocal tract simulator. This development is fueled by inexpensive computing, new modeling methods, and improved medical imaging techniques. The anticipated benefits beyond the speech research field include applications in communication, medicine, education and entertainment. As Beckman illustrates [3], great advances are possible when bridges between basic science and synthesis exist, thus, we hope that the time is right for building this bridge.

REFERENCES

[1] B. S. Atal and J. R. Remde. A new model of lpc excitation for producing natural-sounding speech at low bit rates. In *ICASSP*, pages 614–617, 1982.

[2] P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth. A three-dimensional linear articulatory model based on mri data. In *ICSLP*, pages 14–20, 1998.

[3] M. E. Beckman. *Progress in Speech Synthesis*, pages 185–209. Springer, 1997.

[4] C. H. Coker. A model of articulatory dynamics and control. In *Proc. of IEEE*, volume 64, pages 452–460, 1976.

[5] H. K. Dunn. The calculation of vowel resonances, and an electrical vocal tract. *JASA*, 22:740–753, 1950.

[6] O. Engwall. Modeling of the vocal tract in three dimensions. In *Eurospeech*, pages 113–116, 1999.

[7] G. Fant. *Acoustic Theory of Speech Production, 2nd ed.* S'Grovenhage, Mouton, 1970.

[8] S. S. Fels, F. Vogt, B. Gick, C. Jaeger, and I. Wilson. User-centered design for an open source 3d atriculatory syntheesizer. In *ICPhS*, 2003.

[9] J. L. Flanagan. *Linear Prediction of Speech.* Speech analysis, synthesis and perception, Berlin, 1972.

[10] J. L. Flanagan and K. Ishizaka. Automatic generation of voiceless excitation in a vocal cord-vocal tract speech synthesizer. *ASSP*, 24:163–170, 1976.

[11] P. J. Jackson. *Characterisation of plosive, fricative and aspiration components in speech production.* PhD thesis, Univ. of Southhampton, 2000.

[12] K. Kähler, J. Haber, H. Yamauchi, and H.-P. Seidel. Head shop: Generating animated head models with anatomical structure. In *SIGGRAPH*, pages 55–64, 2002.

[13] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *SIGGRAPH*, pages 55–62, 1995.

[14] P. Mermelstein. Articulatory model for the study of speech production. *JASA*, 53:1071–1082, 1973.

[15] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *JASA*, 49(2):583–590, 1971.

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

[17] K.-I. Sakakibara, T. Konishi, K. Kondo, E. Z. Murano, M. Kumada, H. Imagawa, and S. Niimi. Vocal fold and false vocal fold vibrations and synthesis of khoomei. In *ICMC*, pages 135–138, 2001.

[18] C. H. Shadle, A. Barney, and P. Davies. Fluid flow in a dynamic mechanical model of the vocal folds and tract. i. measurements and theory. *JASA*, 105:444–455, 1999.

[19] K. Shirai and M. Honda. Estimation of articulatory motion from speech waves and its application for automatic recognition. In *Spoken Language Generation and Understanding*, pages 87–99. D. Reidel Publishing Company, 1980.

[20] Y. Shiraki and M. Honda. LPC speech coding based on variable-length segment quantization. *ASSP*, 36:1437–1444, 1988.

[21] D. J. Sinder. *Speech Synthesis Using an Aeroacoustic Fricative Model.* PhD thesis, Rutgers Univ., NJ, 1999.

[22] M. Stone. Toward a model of three-dimensional tongue movement. *Phonetics*, 19:309–320, 1991.

[23] F. Vogt, G. McCaig, M. A. Ali, and S. Fels. Tongue 'n' groove. In *NIME02*, pages 60–64, 2002.

[24] J. Wernecke. *Inventor Mentor.* Addison-Wesley, 1994.

[25] H. C. Yehia and M. Tiede. A parametric three-dimensional model of the vocal-tract based on mri data. pages 1619–1625, 1997.