User-centred Design for an Open-source 3-D Articulatory Synthesizer

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

[†] University of British Columbia, Canada
[‡] ATR, HIS Laboratory, Japan
[*] Haskins Laboratory, USA
Email: ssfels@ece.ubc.ca, fvogt@ece.ubc.ca, gick@interchange.ubc.ca,
carolj@ece.ubc.ca, ilwilson@interchange.ubc.ca

ABSTRACT

We describe our research direction to build a 3-D articulatory speech synthesizer. We are employing a user-centred design approach to develop the synthesizer. The end result will be added to the open-source research community. The system infrastructure we are planning is expected to allow for easy integration of current and future research results in vocal tract modelling and vocal sound production. In this paper, we identify the tasks that the speech synthesizer will support and the interface for using the speech synthesizer. Further, we hope that this paper will provide a call for participation from fellow researchers to contribute their results to this project.

1 INTRODUCTION

Research in articulatory speech synthesis started as far back as von Kempelen's synthesizer in 1791 [21] and has continued through to present-day techniques using MRI and ultrasound-based models to drive computer models of vocal tract parts. The difficulty we see is that the research techniques and tools to date have been mostly isolated from each other. Thus, while one group of researchers may have excellent lip models, another may have good tongue models; there is no common platform for each group to test and validate their contribution in the context of a fully functioning articulatory synthesizer. We are working towards creating a complete, three-dimensional (3-D) vocal tract modelling platform for researchers. Our direction is to employ a user-centred design approach to build an open-source oriented 3-D articulatory speech synthesizer. As the model is user-centred, we hope speech synthesis researchers will participate and send us their recommendations and design criteria to build a modular, flexible 3-D articulatory speech synthesizer. We call this project the Articulatory Speech Synthesizer (ArtSS).

Our approach in this paper is to start the process by enumerating a hierarchical decomposition of the tasks necessary for a useful 3-D articulatory speech synthesizer. This list is a good starting point for colleagues to change and expand to fit their own needs. We have derived an initial task list by looking at the broad range of research contributions in the field. Our first attempt to validate the synthesizer design is to elucidate a scenario based on a stereotypical sequence of steps a researcher may go through. We anticipate developing a number of typical scenarios to validate the flexibility of the synthesizer design to accommodate different researchers' needs.

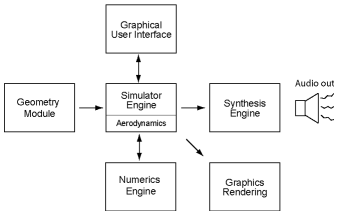

Figure 1: Example of Main Windows of ArtSS (prototype)

At this point in our research, we have begun to develop an articulatory speech synthesis software architecture to support the tasks we see. Our architecture is structured to provide modularity and flexibility so researchers can easily add and remove parts relevant to their own work. Likewise, many complexities related to numerical algorithms, graphical visualization and user interfaces are separated from the main tasks of modelling, synthesis and data extraction. Our progress to date in constructing this system is described in [20].

2 RELATED WORK

As we are taking a user-centred approach in designing the platform, this paper is both a call for participation in helping with the design of the interface and system as well as a status report on the construction.

We are seeking the help of researchers to let us know their needs beyond results published in the literature. Functionally, the speech synthesizer is anticipated to provide the following:

1. a modular technique for specifying 2-D and 3-D vocal tract geometry and dynamics, based on extensible mechanisms such as Scene Graphs;
2. an extensible mechanism for producing speech from vocal tract geometry, including support for traditional, area-function based, approaches and aeroacoustic approaches;
3. a flexible and extensible technique for controlling vocal tract geometry and dynamics interactively, from a variety of sources such as medical image data;
4. an extensible mechanism for modelling vocal tract dynamics including various soft tissue models such as finite element method models (FEM), boundary element method models (BEM) and spring-mass models.

The basic functions above, coupled to an easy-to-use interface should support a multitude of researcher tasks. At this point, we are not expecting to provide real-time capabilities, however, the modular approach we envision should make using our tool effective for investigating new modelling and synthesis techniques. Promising techniques can then be tested for eventual use in a real-time environment. Ultimately, our goal is to create an infrastructure that will provide a common toolset and 3-D articulatory speech synthesizer for the research community. We hope that the common infrastructure will allow researchers to focus on the complexities of their particular area and be able to try their contributions within the larger context without the overhead of re-implementing many parts of a 3-D articulatory speech synthesizer. In the end, collectively, we will make articulatory synthesis a competitive technique both for applications and for the study of speech production and understanding.

3 RESEARCH AND APPLICATION TASKS FOR ARTICULATORY SYNTHESIS

Depending upon researchers' objectives, there are many approaches, techniques and tools that are used in articulatory speech synthesis. Finding a common tool set that encompasses all needs is complicated. In 1981, Rubin et al. developed an approach for an articulatory synthesis model and made the model available for researchers [13]. This synthesis method is essentially a 2-D articulatory synthesis platform, limiting the types of sounds that are possible. With the development of new imaging techniques and modelling techniques, it is now possible to shift to a 3-D model of the vocal tract. We can still use a source-excited area-function based filter model for speech synthesis, however, this limits the types of sounds that can be produced. Our approach lays a foundation for both 2-D and 3-D vocal tract models. With a 3-D geometry, we can also support other techniques for sound production such as aeroacoustic models.

At this stage we have developed an initial hierarchical task decomposition through a survey of a representative set of researchers' work. Illustrative citations of some of the many research results investigated include: [1, 2, 3, 5, 4, 6, 7, 8, 9, 10, 12, 11, 14, 13, 15, 16, 17, 18, 19, 20, 21, 23, 24] The researchers involved in this symposium [22] include many of the people active in the field and thus can provide further insight into the functions needed in the articulatory speech synthesizer. While not exhaustive at this point, the survey provides a picture of some of the tasks that will need to be supported by the 3-D articulatory synthesizer. We are also seeking to use the 3-D articulatory synthesizer in our own research for text-to-speech synthesis, gesture-controlled speech and computational linguistic studies. Our own research goals also aid us in task decomposition.

We have identified several primary tasks that are performed by speech researchers using or exploring articulatory speech synthesis. These include:

1. Import and integrate new 2-D or 3-D models of vocal tract parts
   (a) mport static geometry of vocal tract parts
   (b) Import dynamic models
   (c) "Glue" new model into existing methods; e.g., attach a new lip model to a vocal tract
2. Import and integrate new excitation models
   (a) Import time domain glottal waveforms
   (b) Import 2-D or 3-D vocal fold models
   (c) Integrate excitation with vocal tract model and investigate both independent and dependent sources
3. Analyze new vocal tract models in articulatory speech synthesizer
   (a) Measure deformation of new model over time
   (b) Adjust parameters of new model
   (c) Adjust parameters of infrastructure to accommodate new model, i.e. change integration method or model seam parameters

(d) Specify timelines for parameter values for animation of vocal tract

(e) Compare speech output using different synthesis methods

4. Compare different vocal tract models

(a) Monitor vocal tract geometry and identify differences

(b) Monitor speech output and identify differences in time domain and frequency domain both analytically and perceptually

(c) Specify timelines for animation of different models

5. Compare different data for driving vocal tract models

(a) Import data from MRI, Ultrasound, EMA and other data sources

(b) Link data sources to model parameters

(c) Specify intervals for driving model parameters from data while specifying model parameters to be driven by simulation

(d) Compare both vocal tract shapes and acoustic output using both perceptual and objective measures

6. Synthesize speech

(a) Specify time intervals for synthesizing speech from either data-driven or simulated articulatory parameters

(b) Concatenate and interpolate articulatory parameters using different methods for text-to-speech synthesis

(c) Alternate between simple and complex synthesis models including 2-D tube models, simplified aeroacoustic model and complete aeroacoustic model

(d) Alternate excitation methods ranging from simple glottal waveforms to complex vocal fold models for pitch control

7. Integrate vocal tract model with face models

(a) Explore integration with models ranging from geometric face models to complex dynamic face models

(b) Perform audio and/or visual perception tests

3.1 Probes: Articulatory Parameter Abstraction

With the ArtSS project, we are creating a tool set and environment that will allow the seven tasks outlined above to be performed with relative ease. Notice that we are not planning to integrate data extraction and analysis tools that many researchers have developed and that are part of the normal process of speech research. Instead, we intend to provide an abstraction of all data sources as probes. An input probe will be either an input data source such as a sequence of video images from MRI, a sequence of articulatory parameters derived from rules, or any other set of extracted data provided by the researcher. Likewise, an output probe is any data extracted from the articulatory speech synthesizer specified by the researcher through the vocal tract model. These output probes include such things as the acoustic output, a specific articulatory parameter as it changes over time or a virtual MRI image of the synthetic vocal tract.

The intent for implementation of the probe abstraction is to provide hooks into the vocal tract modelling representation that link a probe with particular parts of the vocal tract model. Thus, an engine link can be used to use a data-derived articulatory parameter to drive a specific set of model points over time. Likewise, a virtual microphone at the virtual mouth provides a probe of the acoustic waveform. Other probe types can be defined and included in the model. The use of the probe abstraction provides an interface for the wide variety of data sources and sinks that are important for speech research.

4 SCENARIOS

At this stage of the research, we are attempting to specify the needs of researchers in the field by looking at our own requirements and deriving the requirements of others from the literature and through personal communications. The task analysis in Section 2 gives a high-order decomposition of some of the activities that need to be supported in the 3-D articulatory speech synthesizer. In this section we present a scenario that illustrates how a typical researcher might use the synthesizer. This approach provides a means of looking at the work flow and types of interfaces that would be required. We plan to further explore different scenarios to validate the elucidation of the task space.

In our first scenario, we consider Bob, a post-doc working on a new tongue tip model. He has created a new 3-D FEM-based tongue tip model in Matlab. He has also extracted various parameters for his tongue tip model from ultrasound data for various speech sounds. He wants to try his tongue tip model in the 3-D articulatory synthesizer so that he can hear the effects of different movement parameters on sound output. Bob plans to use the ArtSS system, since he knows that he would be able to test his model in a complete speech synthesis framework without having to write his own speech synthesis algorithms. He starts up the ArtSS

system.

Upon starting ArtSS, Bob sees the initial user interface appear as illustrated in Figure 1. The upper left window, called the SG window, shows a scene graph representation of the default spring-mass model of the vocal tract. The upper right window, called the graphics window, shows a 3-D rendering of the current vocal tract model. The window below, called the timeline window, shows a timeline of probes that are used to drive the synthesizer. At this point, in this scenario, there would be no input probes selected for driving the simulation so the timeline window would normally be empty rather than having the probes as indicated in the figure.

The first thing Bob wants to do is to replace the default tongue tip model with his own tongue tip model while leaving the remainder of the vocal tract model unchanged. In the graphics window, he zooms in and selects the region of the tongue in the vocal tract model that he wants to replace. The selected parts appear in the SG window as well. He then cuts the section out. At this point, a new window appears with the names and positions of the end points where the cut was made. This list is used to interface to Bob's model that he wants to insert.

Next, Bob imports his model. He loads his Matlab code into the simulator. His code provides calls that are made as the simulation proceeds through the vocal tract model and reaches references to his model. The critical points are the places where his model attaches to the default model. A close up of how the connections are made is displayed in a new window. Bob assigns links between his model's nodes and the nodes of the default vocal tract model at the points that were cut. He notes that he successfully made the scale of his model the same as the rest of the components so that the connections are easily made.

For each group of links he specifies parameters for how each node at the boundary affects attached nodes. Essentially, Bob specifies the flow of forces and the boundary conditions for the seam between his model and the default model. As Bob is not so concerned about the interaction between his model and the default model, he selects that the links are treated as anchors. As such, they will only have geometric constraints, effectively providing a solid anchor for his model.

Next, Bob imports his probe values that he has extracted from his ultrasound data so that he can drive his tongue tip model. At this point in his research, he has only extracted the movement of a single location of the tongue tip. He notices that when he imported his model in the step before, the timeline window indicated that additional probes were available. This is due to the fact that as part of the interface, Bob indicated which parameters of his model were available as probes. These probes have associated functions that are called at each time step to update if there is data specified for them in the timeline window. As he only has data for one point on the tongue tip, he leaves other previously defined probes unspecified so that the simulation will automatically calculate the values of these probes over time. When he imports his probe data, the probe appears in a probe clip window, much like a non-linear video editing suite uses video clip windows. Bob drags the probe clip into the timeline so that it will be used to drive his model.

At this point, Bob can press play and watch the probe data drive the FEM model h e has created. However, his main goal is to produce speech using the whole vocal tract. His ultrasound data is from a subject saying /la/. Thus, he needs some data for the rest of the vocal tract along with a sound source for the excitation. He opens the data import window and looks in the default data directory. He finds a probe file for an /l/ and an /a/ for the default model and loads them in. These clips appear in the clip window. He drags them into the timeline window and adjusts their durations dynamically to fit with his probe data. He also adjusts the linear interpolation between the probe data to get a smooth transition for the default probes. He selects a simple acoustic glottal pulse for his excitation and an area-function sound synthesis method. He places a virtual microphone at the mouth of the vocal tract model. The microphone probe appears on the timeline. He presses play and the synthesis engine steps through the timeline and synthesizes speech by updating the model parameters and estimating the vocal tract shape at each time step. The acoustic wave appears in the timeline, and is synchronized with the other input probes. Bob listens to the sound and is happy with it. He saves the sound file for further processing. He also then saves the entire state of the simulator so that he can show this to his colleagues later.

5 CURRENT DESIGN APPROACH

The above scenario illustrates some of the complex operations that would be useful for researchers developing and using a 3-D articulatory speech synthesizer. In our current vision of the design of the system, we have grouped the various parts of the system into five main components, and a graphics rendering module, as shown in Figure 2. The architecture is composed of: 1. a simulator engine, 2. a 3-D geometry module, 3. a graphical user interface (GUI) module, 4. a synthesis engine, 5. a numerics engine, and 6. a graphics rendering engine. In this design, 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. Details of

the overall infrastructure design are described in [20].

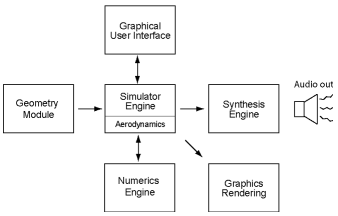

Figure 2: Block diagram of Art proposed design.

The geometry module provides nodes for constructing hierarchical 3-D models including: 3-D coordinates, transformations, allowances for user control, and material properties. Material properties may be as simple as colour or as complex as a dynamic node that specifies how a node can change its behaviour over time. These dynamic nodes provide data sources for changing the model as the simulation progresses. Data sources may include: physical models such as spring-mass models,

FEMs, image data such as MRI and ultrasound, and acoustic data. The GUI module provides a convenient mechanism for interaction with the model. The numerics engine provides general purpose numerical methods that can be switched on demand to provide a simple mechanism for using different simulation methods.

The 3-D vocal tract and face model consists of a hierarchical object-oriented structure which represents multiple levels of detail. Anatomical structures are represented at the top level by nodes such as tongue, lips and larynx. These high-level structures consist of medium-level structures such as muscle groups, bones and tissue. The medium-level structures are composed of lower-level structures like geometry, muscle fibres and ligaments. A library of modules allows the user to define deformable models as well as muscular activation of a model in three-dimensions.

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 the combining of models of different structures, each potentially having different levels of detail, from different research groups, thus providing a testbed for articulatory based speech research and production. Our ultimate aim is to have a fully functioning 3-D vocal tract model that uses aeroacoustic models to produce speech. We intend to use Matlab as our main platform as many researchers already use it for developing models as well as for data extraction and data analysis. However, we will need to augment Matlab to provide enhanced user interface capabilities and a general purpose application interface for researchers who code models in different computer languages such as C. The simulator will look like a Matlab toolkit with additional elements that are provided outside of the Matlab interpreter.

Our design goals are structured around modularity, flexibility and ease of use of the articulatory speech synthesizer. Although we will likely compromise performance (i.e. processing speed) to achieve these goals, our belief is that a common platform/toolkit that provides a multifunction articulatory speech synthesizer and that can be modified piece-by-piece will allow researchers to try their ideas out easily in the full vocal tract context.

6 SUMMARY

At this point in the process we have identified many of the main tasks required of a general purpose research synthesizer. The main tasks are: importing different vocal and excitation models, analyzing and comparing new vocal tract models, driving vocal models from medical image data, synthesizing speech and combining vocal models with face models. We are making progress on providing an easy-to-use interface for researchers to combine their models with the results of others in the articulatory speech synthesis community. We are employing a user-centred design process and therefore are encouraging other members of the research community to participate in the development of this speech synthesizer. Based on our analysis thus far, we are building a modular infrastructure that integrates measured articulatory data, vocal models, excitation models, sound synthesis models, visualization tools, and numerical methods. Ultimately, we hope to create a general purpose 3-D articulatory speech synthesizer that can be driven from both medical image data as well as synthetic articulatory parameters.

7 ArtSS Contact

For more information or to participate, please contact the authors and refer to: http://hct.ece.ubc.ca/research/speech

REFERENCES

[1] 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.

[2] S. Basu, N. Oliver, and A. Pentland. 3d lip shapes from video: A combined physical-statistical model. *Speech Communication*, 25(12):131–148, 1998.

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

[4] M. Jackson, C. Y. Espy-Wilson, and S. E. Boyce. Verifying a vocal tract model with a closed side-branch. *JASA*, 109(6), 2001.

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

[6] Y. Lee, D. Terzopoulos, and K. Waters. Constructing physics-based facial models of individuals. In *GI*, pages 1–8, 1993.

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

[8] S. Maeda. Improved articulatory model. *JASA*, 84(S1):146pp, 1988.

[9] H. D. Maxey. Smithsonian speeech synthesis history project (ssshp), 2002.

[10] J. S. Perkell. Properties of the tongue help to define vowel categories : hypotheses based on physiologically-oriented modeling. *Phonetics*, 24:3.

[11] L. Reveret, G. Bailly, and P. Badin. Mother : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *ICSLP*, volume 2, pages 755–758, 2000.

[12] L. Reveret and C. Benoit. A new 3d lip model for analysis and synthesis of lip motion in speech production. In *Proc. of the Second ESCA Workshop on Audio-Visual Speech Processing*, pages 207–212, 1998.

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

[14] P. Rubin and E. Vatikiotis-Bateson. Webpage: Talking heads, 1998.

[15] 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.

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

[17] M. Stone, E. Davis, A. Douglas, M. NessAiver, R. Gullapalli, W. S. Levine, and A. Lundberg. Modeling the motion of the internal tongue from tagged cine-MRI images. *JASA*, 109(6):2974–2982, 2001.

[18] I. R. Titze. *Principles in Voice production.* Allyn and Bacon, 1994.

[19] E. Vatikiotis-Bateson and D. J. Ostry. An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics*, 23:101–117, 1995.

[20] F. Vogt, S. S. Fels, B. Gick, C. Jaeger, and I. Wilson. Extensible infrastructure for a 3d face and vocal-tract model. In *ICPhS*, 2003.

[21] W. R. von Kempelen. *Mechanismus der menschlichen Sprache nebst Beschreibungeiner sprechenden Maschine. Mit einer Einleitung von-Herbert E. Brekle und Wolfgang Wild.* Stuttgart-Bad Cannstatt F. Frommann, Stuttgart, 1970.

[22] D. Whalen. Icphs symposium on atriculatory synthesis, 2003.

[23] R. Wilhelms-Tricarico. Physiological modeling of speech production: methods for modeling soft-tissue articulators. *JASA*, 97(5):3085–98, 1995.

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