Investigation of Gesture Controlled Articulatory Vocal Synthesizer using a Bio-Mechanical Mapping Layer

Johnty Wang
Media and Graphics Interdisciplinary Centre (MAGIC)
University of British Columbia
Vancouver BC, Canada
johnty@ece.ubc.ca

Nicolas d'Alessandro
Media and Graphics Interdisciplinary Centre
University of British Columbia
Vancouver BC, Canada
nda@magic.ubc.ca

Sidney Fels
Electrical and Computer Engineering & MAGIC
University of British Columbia
Vancouver BC, Canada
ssfels@ece.ubc.ca

Robert Pritchard
Media and Graphics Interdisciplinary Centre
University of British Columbia
Vancouver BC, Canada
bob@mail.ubc.ca

ABSTRACT

We have added a dynamic bio-mechanical mapping layer that contains a model of the human vocal tract with tongue muscle activations as input and tract geometry as output to a real time gesture controlled voice synthesizer system used for musical performance and speech research. Using this mapping layer, we conducted user studies comparing controlling the model muscle activations using a 2D set of force sensors with a position controlled kinematic input space that maps directly to the sound. Preliminary user evaluation suggests that it was more difficult to using force input but the resultant output sound was more intelligible and natural compared to the kinematic controller. This result shows that force input is a potentially feasible for browsing through a vowel space for an articulatory voice synthesis system, although further evaluation is required.

Keywords

Gesture, Mapping, Articulatory, Speech, Singing, Synthesis

1. INTRODUCTION

The human voice is one of the most intimate musical instruments known. This intimacy, based on the fact that it resides within the human body, coupled with its usage for communicative purposes, makes the voice a difficult instrument to analyze, and is therefore the subject of a significant amount of scientific research. Within the NIME context, because all hearing people are essentially "expert listeners", the voice becomes an interesting platform for the development and evaluation of new instruments.

The Digital Ventriloquist Actor (DiVA) system is a real time gesture controlled speech and singing synthesizer [5] The DiVA, as shown in Figure 1, is a solo voice instrument that has been used in various performances [10, 11] as well as speech research.

NIME'12, May 21 – 23, 2012, University of Michigan, Ann Arbor.

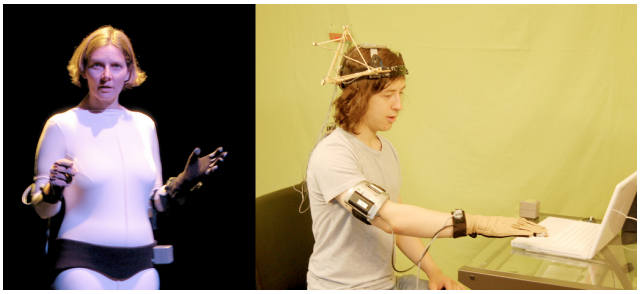

Figure 1: The DiVA system used during performance and speech research

The original DiVA system maps gesture input directly into sound space to control a formant synthesizer [7]. The synthesizer is capable of producing high quality speech as evident by the sample phrases, but requires a large number of input parameters. Mapping the input gestures to the synthesis parameters has been the central focus of the DiVA system. The gesture to sound mapping scheme requires significant user input bandwidth and the result is a system with a gradual learning curve requiring long training time and unnatural sounding output. At the other extreme, concatenative based text to speech synthesis systems can offer natural sounding speech, but at the price of controllability which makes it unsuitable for real time performance. Figure 2 shows the relative positions of the formant-based DiVA system, a concatenative synthesis system and the target goal for the new system.

The proposed method to improve the sound quality while maintaining the ability to control the system as a performance instrument is through the implementation of an articulatory synthesis system with an underlying bio-mechanical model. The motivation of this method lies in the fact that the constraints imposed by the model is based on physiology of the human vocal apparatus and as such, a meaningful mapping system from the input hand gestures, that are also muscle based, will provide a more natural vocal tract configuration and the output sound is produced through articulatory synthesis. This method should not only create a more expressive vocal instrument, but also provide new avenues for speech research since the remapping of articulators (to the hands) allows exploration on the cognitive process of

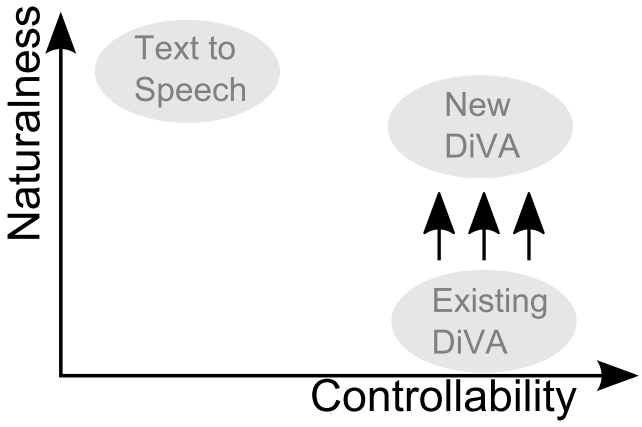

Figure 2: Controllability vs Naturalness

vocal production. With the vocal tract as the underlying input representation for control for the user, it remains unclear whether force control (i.e., isometric, where gestures force relates directly to muscle activation) or position control (i.e., isotonic, where gesture position maps to the shape of the vocal tract) provide effective control techniques. As a first step evaluation, a force input controller mapped to articulators was implemented using FSRs, and compared with a position controlled articulator shape mapping using a touch pad.

This paper first describes related work, then provides an overview of the implemented system and then describes preliminary evaluation that compares the new and existing systems focusing on the gesture input and synthesized results. A discussion on the limitations of the evaluation is presented followed by suggestions for future development.

2. RELATED WORK

The DiVA system is based on GloveTalkII [6] where hand gestures are captured using an instrumented glove and sent to a software mapping system implemented using neural networks. The mapping system converts the input gestures to formant synthesis parameters in real time to produce sound. Further developments of the system as a musical instrument [9, 5] led to various modifications and additions to the system for artistic applications, with considerable effort spent on developing the aesthetics, modularity and robustness as required for use by musicians during rehearsal and performance.

The HandSketch controller and RAMCESS [3] synthesizer is another example of an existing gesture controlled speech and singing synthesis system.

In terms of articulatory synthesis, VocaltractLab [2] utilizes a parametric model of the vocal tract and is capable of high quality speech and provides an off-line control system used to generate input parameters.

The bio-mechanical model of the tongue used by the new system is based on [13] where a computationally expensive finite element method model was sped up with a small loss in accuracy using stiffness warping. The synthesizer used in the new system [12] provides real time articulatory synthesis based on a tube geometry and parameters driving a self oscillating glottal source model.

In terms of developing new musical interfaces, [8] provides a framework for selecting and evaluating input devices, and [14] provides discussion on sensor choice and their suitability for various musical tasks.

3. SYSTEM OVERVIEW

The new synthesis system consists of a number of modules, as shown in Figure 3. Communication between the modules was implemented using Open Sound Control [1] which allows flexible routing and transmission of the data and the possibility of running the modules on different machines. The following sections explain each module in more detail.

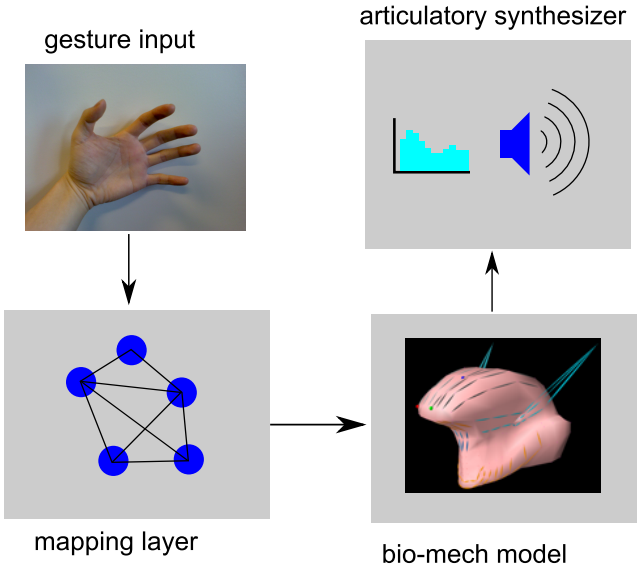

Figure 3: System Diagram

3.1 Force Input and Mapping

The force input and mapping system consists of a series of Force Sensitive Resistors (FSR's) attached to an Arduino microcontroller connected via a serial port to a Max/MSP patch, as shown in Figure 4. The forces are mapped to muscle activations and grouped according to their effect on the tongue body: front, back, up, and down. This input system allows opposing muscles to be activated at the same time.

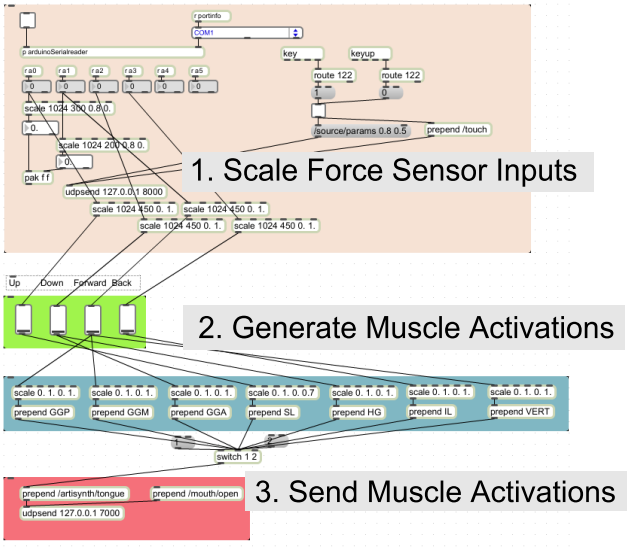

Figure 4: Force Input and Mapping

3.2 Bio-Mechanical Model

The model was implemented in the ArtiSynth modelling environment [4]. A series of beams was constructed around the

tongue model [13] to represent sections of the vocal tract, and 22 marker points were placed at set intervals along the tract. The distances between these marker points and the tongue surface are computed in real time which allows an effective cross sectional area function to be calculated. The muscles in the tongue model are controlled by an OSC listener that listens for messages with address tags corresponding to each muscle.

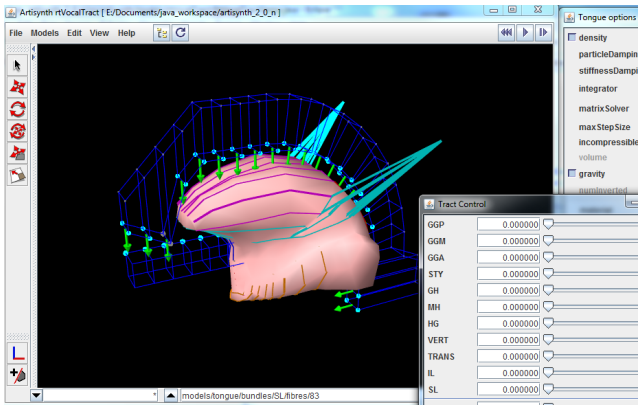

Figure 5: Artisynth Vocal Tract Model

3.3 Synthesizer

The synthesizer is an extended version of [12] that implements an OSC listener and updates the tube shape in real time while synthesizing audio at 44100 Hz. The number of tube sections is set to the same as the output of the bio-mechanical model, although a linear interpolation function is also available for a different number of sections if required.

3.4 System Integration and Tuning

When integrated, the system runs on a single laptop (Intel C2D 2.4GHz Macbook). The static positions in the tract model was manually matched to provide a certain output vowel space when the tongue was activated through the force input system. The vowel space was tuned to include **i**, ɛ, ɑ and **u**, which represent relative extremes of the tongue body along the top/bottom and forward/back positions.

4. COMPARISON WITH EXISTING SYSTEM

To evaluate our sensor choice the new input system is compared with the existing system. A modified version of the existing DiVA synthesis system was set up with the same vowel space (based on formant frequencies) as the new system and a 2D browsing input was implemented on a iPad. The X-Y finger position on the touch screen is fed into the existing mapping system to control formant frequencies, and the finger down/up is used to trigger the sound on/off. For the force input system a separate USB foot-switch was used to control the switching of the sound. Figure 6 shows an image of the two controllers side by side.

4.1 Experiment

A pilot experiment was set up with 3 performers and 4 listeners. The producers were introduced to the interfaces and then asked to perform certain "words" composed of 1 to 3 different vowel sounds. The gesture trajectories were recorded to allow later playback for producing audio samples during listener evaluation. Due to the limited time the performers have practised on the interface for the pilot, only relevant qualitative feedback is currently available.

Table 1: Identification Accuracy

User	Kinematic	Force
1	18 %	35 %
2	36 %	31 %
3	34 %	52 %
4	33 %	62 %
average	30 %	45 %

For the listener evaluation, a word identification task was set to compare the intelligibility of the system outputs. In addition, a series of descriptors such as *sharp*, *exciting*, *natural*, *speechlike* and *intelligible* was provided and the listener had to rank two (unknown) sound samples based on each term.

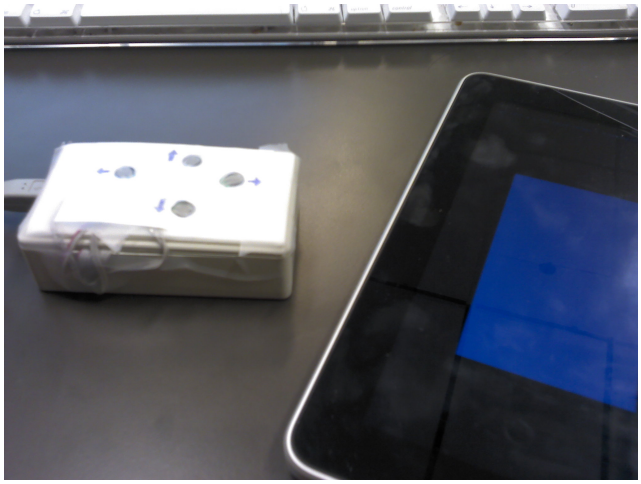

Figure 6: Force and Kinematic Controllers

5. RESULTS

5.1 Performer Evaluation

While input trajectories were captured for all the user input and deviation from target values could be calculated, it should be acknowledged that because the force input system has targets at the saturation point (maximum input activation for the bio-mechanical model), it is not too relevant to make any quantitative comparison between the two interfaces in this respect. All performers, given the limited amount of practice, seem to prefer the touch interface due to its relative ease of use. The fact that the kinematic controller is implemented on a polished commercial product may also influence the performers' preference.

5.2 Listener Evaluation

126 audio samples were used from the performer input data (61 from the existing system and 65 from the new one). They were placed in a randomized list, and played back to each listener for identification. Table 1 shows the accuracy rates for each user. Overall the accuracy for the force controller output is considerably higher than the kinematic system.

For the qualitative descriptors (Table 2), the output from the kinematic system was rated by most listeners as *sharp*. This may be due to the different synthesizer used in the existing system, and suggest that for a more comparable analysis the two input and mapping systems should use the same synthesizer if possible.

Table 2: Qualitative descriptors for system outputs

	Kinematic	Force
sharp	98%	2 %
exciting	77%	23 %
natural	20%	80 %
speech-like	27%	73 %
intelligible	56%	44 %

5.3 Discussion

Through the pilot experiment various issues were discovered that motivates modification and refinement of the evaluation procedure. First, it is clear that for a difficult to use interface, a sufficient period of learning should be expected for proper evaluation [8]. A significant difference in output sound quality was noticed by listeners, and for better comparison the same synthesizer should be used to isolate differences between the input mappings. Since the initial submission of this paper, kinematic input was implemented to drive the articulatory synth between the static tube shapes of the boundary force input cases so the kinematic and force input spaces can be better compared using the same sound output.

An interesting outcome of the preliminary evaluation is that despite the sound from the existing kinematic controller and formant synthesis system is considered "more intelligible" by listeners, the actual identification rate was considerably higher in the new system for 3 out of the 4 listeners. At this point however, it is not certain if the force input is the most feasible approach, and further investigation with a larger number of samples is required.

6. FUTURE WORK

This work represents the preliminary evaluation of using force input for gesture controlled articulatory synthesis. There are limitations in the various components of the current system and it is far from comprehensive as a vocal synthesizer (lacking consonants, stops, etc). However, one of the most crucial aspects of a gesture controlled synthesizer is input system and the immediate focus of future work is to explore and evaluate appropriate input and mapping methods. Based on the findings of the current evaluation, the same synthesis engine should be used to provide a more balanced comparison by eliminating the effect of the difference in sound quality. If certain aspects of kinematic control is found to be appropriate alongside with force control, a hybrid controller such as [15] may be employed to make use of relevant features of both concepts (and indeed may be necessary for generating a wider number of sounds). In addition, as a musical interface, creative tasks such as "composing a short piece" or "performing a musical phrase" are potential avenues for future exploration.

7. CONCLUSION

A force input system was implemented to control muscle activations to browse through a vowel space in an articulatory vocal synthesis system with a bio-mechanical mapping layer. The new input system and corresponding audio output is evaluated by a comparison with the existing system from the performer and listener perspectives. Preliminary user evaluation suggest that the force-based system was more difficult by novice users but the resultant output was no less intelligible and rated to be more natural by listeners. Force input appears to be a feasible controller for browsing through a vowel space in an articulatory voice synthesis system, but further evaluation is required.

8. REFERENCES

[1] Open sound control. `http://opensoundcontrol.org`. Accessed: 30/09/2011.

[2] P. Birkholz, D. Jackel, and K. Kroger. Construction and control of a three-dimensional vocal tract model. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[3] N. d'Alessandro and T. Dutoit. RAMCESS / HandSketch : A Multi-Representation Framework for Realtime and Expressive Singing Synthesis. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[4] S. Fels, J. Lloyd, K. Van Den Doel, F. Vogt, I. Stavness, and E. Vatikiotis-Bateson. Developing Physically-Based, Dynamic Vocal Tract Models using ArtiSynth. In *International Seminar on Speech Production*, volume 6, pages 419–426, Ubatuba, Brazil, 2006. Citeseer.

[5] S. Fels, R. Pritchard, and A. Lenters. Fortouch: A wearable digital ventriloquized actor. In *New Interfaces for Musical Expression (NIME2009)*, pages 274–275, 2009.

[6] S. S. Fels and G. E. Hinton. Glove-TalkII-a neural-network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 9(1):205–12, 1998.

[7] J. Holmes, I. Mattingly, and J. Shearme. speech synthesis by rule. *Language and Speech*, 7(3):127, 1964.

[8] N. Orio, P. I. Stravinsky, N. Schnell, and M. M. Wanderley. Input Devices for Musical Expression : Borrowing Tools from HCI. *Real-Time Systems*.

[9] B. Pritchard and S. Fels. GRASSP : Gesturally-Realized Audio , Speech and Song Performance. In *Proceedings of the 2006 conference on New Interfaces for Musical Expression (NIME2006)*, pages 272–276, Paris, France, 2006.

[10] B. Pritchard, S. Fels, N. d'Alessandro, M. Witvoet, J. Wang, C. Hassall, H. Day-Fraser, and M. Cadell. Performance: what does a body know? In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 2403–2407, New York, NY, USA, 2011. ACM.

[11] B. Pritchard and M. Witvoet. Performance: what does a body know? In *Proceedings of New Interfaces for Musical Expression (NIME2010)*, Sydney, Australia, 2010.

[12] K. van den Doel and U. Ascher. Real-Time Numerical Solution of Webster's Equation on A Nonuniform Grid. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1163–1172, Aug. 2008.

[13] F. Vogt, J. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S. Fels. Efficient 3d finite element modeling of a muscle-activated tongue. *Biomedical Simulation*, pages 19–28, 2006.

[14] M. Wanderley, J. Viollet, F. Isart, and X. Rodet. On the choice of transducer technologies for specific musical functions. In *Proceedings of the 2000 International Computer Music Conference (NIME2000)*, pages 244–247, 2000.

[15] J. Wang, N. d'Alessandro, and S. Fels. Squeezy: Extending a multi-touch screen with force sensing objects for controlling articulatory synthesis. In *Proceedings of New Interfaces for Musical Expression (NIME)*, 2011.