Tongue 'n' Groove: An Ultrasound based Music Controller

Florian Vogt, Graeme McCaig, Mir Adnan Ali, Sidney Fels
Human Communications Technology Laboratory
Department of Electrical and Computer Engineering
University of British Columbia
2356 Main Mall, Vancouver BC, Canada V6T 1Z4
{fvogt, rgmccaig}@ece.ubc.ca, maali@pnr.ca, ssfels@ece.ubc.ca

ABSTRACT

Here we propose a novel musical controller which acquires imaging data of the tongue with a two-dimensional medical ultrasound scanner. A computer vision algorithm extracts from the image a discrete tongue shape to control, in real-time, a musical synthesizer and musical effects. We evaluate the mapping space between tongue shape and controller parameters and its expressive characteristics.

Keywords

Tongue model, ultrasound, real-time, music synthesis, speech interface

INTRODUCTION

Musical controllers may be activated by different parts of the body. Each combination of musical controller and body part results in a different quality of control, expression, and richness of interaction. The human vocal tract is the body part most commonly used for sound generation. Examples are speech, singing, and other non-speech sounds. The tonal shaping of the human voice is to a large extent controlled by the tongue. Using the tongue as an input modality leverages the skills human have acquired though speaking, and has the potential for sensitive and fine control.

Looking at the role of the tongue in speech modeling [3], [6], [7], the vocal tract shape is primarily controlled by the tongue. In voice production modeling the airspace of the vocal tract, from the glottis to the lips, can be considered as a linear filter. This filter acts on input generated by the glottis, also known as the excitation function. This implies strong potential for using the control mechanism of the vocal tract, starting with the tongue shape, to control an external sound synthesis device.

Existing physical instruments which make use of the tongue as a control mechanism include reed instruments, the harmonica, and the mouth harp. Also, instruments such as Mouthesizer and TalkBox use various elements of the human vocal tract to control or modulate sound. The Mouthesizer [8], created by Michael Lyons et al., uses the lips as

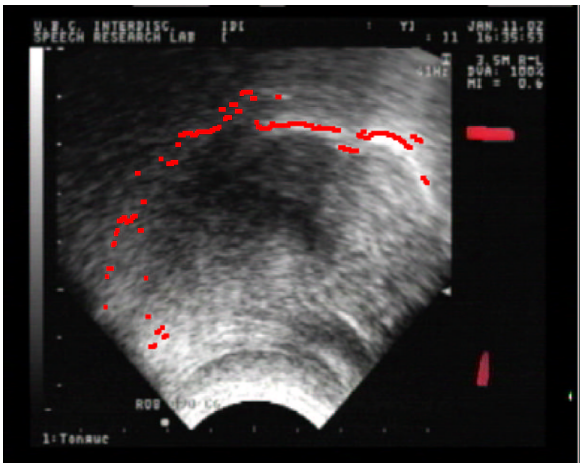

Figure 1: Tongue contour reconstruction algorithm.

the sole means of input. The TalkBox [9], utilizes a speaker placed in the performer's mouth, which records the filtering effect of the mouth using an external microphone. The TalkBox got very popular in the 70's, and is played by many performers such as Peter Frampton [5].

Another related music controller, the Vocoder (Voice Operated reCorDER) [2], extracts from acoustic voice signals the formant frequencies. With the assumption of a single linear filter model, the formant frequencies would be the equivalent of the filter coefficients.

The proposed system is different and novel, in that instead of acoustic measurement, we use an articulatory model based on measurement of the physical configuration of the vocal tract in real time. The principle of using the tongue as a music controller was proposed by David Wessel in [10]. These measurements are used in an active sense to control a digital instrument, rather than the more passive embodiment found in TalkBox where the interior of the mouth is used as a physical acoustic chamber. In the present project,

the mapping of the vocal tract to the sound output is reconfigurable. The goal of this study is not to directly model the vocal tract as used in everyday speech, but rather to explore how to leverage the fine motor control skills developed by the tongue for expressive music control.

Our system is composed of an ultrasound device, positioned under the chin to provide continuous imaging of the performer's tongue. The tongue video is acquired into a computer with a video capture card, which extracts a basic tongue model in real-time with an image processing algorithm. The translation mapping from the tongue model into sound synthesis parameters makes our system a music controller that analyzes the input video as shown in Figure1 and generates the tongue model, which drives . One advantage of this approach is the relatively non-intrusive nature of the ultrasound device, as compared with a system such as TalkBox where mechanical hardware must be inserted into the performer's mouth.

DESIGN CONSIDERATIONS

By building and testing the Tongue 'n' Groove we hope to evaluate the potential of the tongue as an expressive musical controller. We have identified many factors that will determine the effectiveness of this controller. Our design and test plan attempts to explore the following issues:

Physical constraints on motion: The tongue moves within a spatially limited region, and each portion of the tongue is elastically connected to neighboring regions. This is one of the most unique aspects of tongue control.

Accuracy and Speed: The spatial accuracy of a tongue controller is limited on the human side by the accuracy of tongue motor control. Algorithms on the computer side should be designed to support this maximum spatial accuracy. Furthermore, the tehmporal resolutino of the video stream and software processing should be adequate so that time lag is not an obstacle to good control.

Learned abilities with the tongue: People have a preexisting set of skills from using their tongues for singing, speaking, eating and caressing. They also have certain expectations regarding the role of the tongue in sound production.

Intimacy/Emotional Connection: Because of its situation in the body, and its involvement with intimate and communicative activities, a tongue controller may heighten the performer's emotional connection with the produced sound.

Sensor dimensionality: In our system we use a 2D ultrasound device. Our investigation starts with imaging of the mid-sagittal tongue profile. It may be discovered that better control results from the use of 3D or an alternate 2D plane.

SYSTEM DESIGN

Figure2 shows the components of the Tongue 'n' Groove system. An Aloka SSD-900 ultrasound scanner is used with a small probe, similar in shape to a microphone. The performer presses the probe against the underside of the jaw. Sound-conductive gel may be used to lubricate the skin for better probe contact. The probe can be held in hand, or used with a microphone stand.

The SSD-900 produces 2-dimensional images of the tongue profile in analog NTSC video format. Thirty frames per second are obtained at 768x525 resolution. The SSD-900 calibrates the ultrasound image so that image distances correspond to scaled real-world distances. The intensity in different parts of the image depends on the ultrasonic reflectivity of body parts. The tongue-air boundary layer on the upper surface of the tongue, has high reflectivity and therefore creates the most intense region of the image.

The ultrasound image is digitized using a Linux workstation with a video capture card. A video capture library written in C makes image data available to the Tongue 'n' Groove image processing algorithm. Two different algorithms have been tested with the Tongue 'n' Groove so far. One algorithm uses optical flow, and is based on Sidney Fels' Iamascope system [4]. It calculates the amount of motion within each of 10 vertical bands of the tongue image. The other algorithm calculates a vector of vertical positions along the tongue surface.

The output of the image processing algorithm is used to provide constantly updated control parameters to a synthesis engine at the video frame rate. The optical flow algorithm sends MIDI signals to the PC internal sound card, causing notes to be triggered when motion occurs. The 10 different bands control 10 pitches within a predefined chord. The tongue-height algorithm sends its output vector to the Singing Physical Articulatory Synthesis Model (SPASM)[1], where the tongue heights are mapped to radii of cylindrical segments in the virtual resonant chamber.

IMAGE PROCESSING

Tongue Contour Reconstruction

If the Tongue 'n' Groove were intended to control realistic human voice sounds, it would be important to have accurate readings of tongue/hard palate positions in order to drive a vocal tract model. However, our goal is broader: to use the tongue to control expressive musical sounds, including abstract vocal-type sounds and non-vocal sounds. As long as a consistent mapping is applied, users can learn the relationship between tongue motion and sonic effect. Therefore we have implemented a fairly simple image-processing algorithm that outputs a vector of relative heights along the top surface of the tongue.

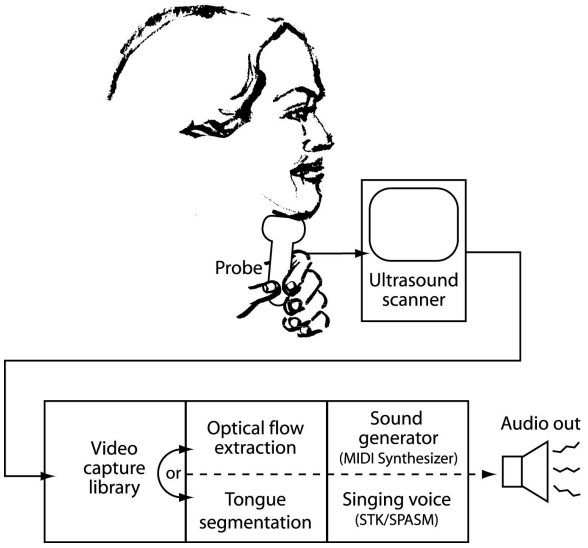

Figure 2: System diagram of the Tongue 'n' Groove

It does not attempt to measure absolute position within the throat or the shape of the hard palate.

In measuring the configuration of the tongue, we acquire an NTSC image from the ultrasound scanner. The intensity levels of the image are then normalized under the assumption that the ambient intensity is inversely proportional to a small power of the distance from the center of the probe. The actual exponent, computed by averaging over several data-sets acquired by the ultrasound, has a value of about 2.2.

Since the probe is held under the chin, the tongue is approximately in the same position relative to the probe regardless of the user. The region of interest in the ultrasound scan is therefore fixed. This region is scanned for maximum pixel intensities across the field of the image. These values correspond to the distance from the probe to the lower contour of the tongue. Since the hard palate is fixed, these values give all the required information to estimate the configuration of this portion of the vocal tract.

The currently implemented algorithm does not compensate for shifting and rotation of the entire ultrasound image. This allows the user to change the vector of outputs by changing position and pressure of the ultrasound probe against the throat. We view this as a desirable feature, since the user can learn to employ the probe as a second controller, modifying the sound output in unique ways.

Due to its simplicity, the algorithm is capable of 30 frames per second output on a 800MHz Pentium-II Workstation. As can be seen from Figure 1, the algorithm is subject to a significant amount of noise and error, which causes unintended fluctuations in the musical output.

We are improving the current algorithm with methods such as outlier removal and simple spline and polynomial curve fitting, to increase the accuracy within our real-time constraint. With the algorithm improvements we predict more precise control and improved expressive potential for the Tongue 'n' Groove. We plan to investigate the application of more computationally expensive methods, such as discrete snakes and Kalman filtering, in the spatial and temporal domain for this real-time constraint problem.

Optical Flow Extraction

The second image processing algorithm for this application extracts optical flow. Instead of analyzing the tongue's position, this algorithm analyzes the tongue's motion in ten horizontally-spaced segments. The flow extraction algorithm computes the motion intensity by taking the difference between consecutive image frames.

SOUND SYNTHESIS

We plan to implement music synthesis algorithms as output for the Tongue 'n' Groove. We believe that an important factor in determining the success of a tongue-to-music mapping is the similarity between the artificial control mapping and the natural frequency-shaping function of the tongue. On this basis, we have identified three broad categories of instrument to be implemented and compared:

1. The tongue image data can be used to control the resonant tube shape in a physically modeled human singing voice synthesizer.
2. The tongue image data can be used to control filter-shaping parameters in a non-vocal instrument.
3. The tongue image data can be used to control a set of non-filter-related parameters in a non-vocal instrument.

Currently, the two instruments, Tongue-SPASM and the Tongue-Scope, we have implemented fit Categories 1 and 3 respectively.

Tongue-SPASM

The Tongue-SPASM instrument is based on Perry Cook's Singing Physical Articulatory Synthesis Model (SPASM). SPASM simulates human voice sounds, by modeling a vocal excitation function and filtering it through a virtual vocal tube with varying cross-section. The Tongue-SPASM maps tongue heights to radii of cylindrical segments in the virtual resonant tube.

We adopted the Linux version of SPASM to allow for real-time control. The original code is designed to read vocal

tract radii values from a script file, changing the radii values at certain intervals as defined in the script. We modified SPASM to read radii vectors from an Unix file descriptor to update the simulated vocal tract on input changes.

The Tongue-SPASM algorithm is capable of reading in new control vectors and changing the sonic output at a rate of at least 30 signals/second, corresponding to the video frame rate to the ultrasound signal.

Tongue-Scope

The Tongue-Scope instrument is based on the musical output code of Sidney Fels' Iamascope system. The algorithm proceeds at a constant rate through a predefined, looped, chord sequence. Each chord contains 10 possible pitches that are triggered asynchronously by detected motion in the corresponding tongue-image segment. Notes are play-ed through the internal MIDI synthesizer of a PC sound card. The instrument sound changes periodically according to a predefined cycle.

Future Instruments

Here are our future design concepts for alternate Tongue 'n' Groove instruments:

Squeezable Sax A saxophone instrument from Synthesis Tool Kit will be modified so that the resonant tube of the saxophone takes on different radii values along different portions of its length. These radii values will be controlled by the incoming vector of tongue heights.

Mouth Cathedral The tongue will be used to control the parameters of an echo-chamber effect, giving the user a feeling of controlling a large resonant room with the space of the mouth. This mapping is appealing since it extends the intuitively correct concept of the mouth as resonant chamber: by extending the time scale of the resonance, an echo operation is achieved instead of a filter operation, but perhaps an intuitive sense of control will remain.

Marble-Mouth The tongue will be used to control a number of virtual, bouncing spheres within the mouth. Each sphere occupies a fixed horizontal position and bounces up and down between the tongue surface and an arbitrary upper surface. As the tongue positions increase in height, the period of bouncing decreases. Each sphere creates a distinctive pitch, while timbre depends on the velocity of impact. The resulting gestalt of repeated, tonal/percussive sounds should have an effect similar to a gamelan ensemble. This mapping is interesting because it leverages people's experience of using the tongue to manipulate objects in the mouth, as in the act of eating.

OBSERVATIONS

After completing the current prototype of Tongue 'n' Groove, we performed some preliminary testing by playing each instrument, Tongue-Scope and Tongue-SPASM ourselves. Here are our observations:

Tongue-Scope Given the current mapping, playing the Tongue-Scope only remained interesting for a few minutes. Tongue motion resulted in sound output, but the control did not seem intuitive or precise. This could perhaps be useful for a basic walk-up demo.

Tongue-SPASM Tongue-SPASM produced a richer experience than the Tongue-Scope. Despite the strong noise originating in the image extraction algorithm, it seemed that patterns of tongue motion could be repeated to form musical phrases. We found it intuitively satisfying to map the tongue as a spectral controller. Due to the presence of noise, there was significant variation in the tongue position vector from one reading to the next, which caused a grainy, rhythmic segmentation in the audio. However, this led to the unplanned discovery that adding an automatic rhythmic component to the output increases the musical interest of playing the Tongue 'n' Groove (an effect we may purposely employ in the next generation of instruments). Overall the tongue and sound feeling was appealing.

PROPOSED EVALUATION

Our first method of testing the Tongue 'n' Groove will be based on user surveys and qualitative observation. Participants will try each instrument, and answer questions similar to the following:

- Was it easy to understand the relationship between tongue movement and musical result?
- Were you able to shape the sound according to your intentions?
- Does the musical output sound aesthetically pleasing, and/ or musically expressive?

Non-performing listeners will also be surveyed regarding the musicality of the output. Comparing the three instruments will allow us to make inferences about which types of musical mapping are most appropriate with a tongue-profile based controller.

We are also developing tests of a quantitative nature, to be performed as time permits:

- Simple measurements from the captured image stream to determine maximum tongue velocity and range of motion.
- "Produce the same musical phrase repeatedly" tests to measure variation of tongue motion.
- "Mimic a supplied musical phrase" tests to measure subject's understanding of the control space.
- Tests to compare the precision and expression of tongue control with the control afforded by other body parts. We envision a simple implementation of a controller based on video capture of a hand viewed from the side. Since the

hand shares the same physical connectivity constraint as the tongue, the two controllers could be compared with a variety of tasks.

- The "sing-along test". If at some point an image-processing algorithm is implemented that estimates a vector of tongue-to-hard palate distances with low error, we will compare a fully scripted, artificially controlled song passage (as produced by the original SPASM) with the same passage controlled by the Tongue 'n' Groove. Users will sing along in real time with a scripted time-varying sequence of source excitations and consonant sounds, controlling only the filter parameters. A comparison will be made of which output sounds most natural and expressive.

ACKNOWLEDGMENTS

We thank Perry Cook for his contribution of SPASM and Bryan Gick for help with the ultrasound scanner. We also thank Paula Wirth for creating the illustration in Figure 2.

REFERENCES

1. P. Cook. SPASM, a real-time vocal tract physical model controller and singer, the companion software synthesis system, 1993.

2. H. Dudley. Remaking speech. *Journal of the Acoustic Society of America*, 11:169–177, 1939.

3. G. Fant. *Acoustic Theory of Speech Production*. S'Grovenhage, Mouton, 1960.

4. S. S. Fels and K. Mase. Iamascope: A graphical musical instrument. *Computers and Graphics*, 2(23):277–286, 1999.

5. P. Frampton. The TalkBox. http://www.frampton.com, accessed on Apr. 3 2002.

6. J. N. Holmes. Formant synthesizers: Cascade or parallel? *Speech Communication 2*, 1983.

7. D. H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustic Society of America*, 67:971–995, 1980.

8. M. Lyons and N. Tetsutani. Facing the Music: A Facial Action Controlled Musical Interface. In *Proceedings ACM CHI*, 2001.

9. NewMusicBox. Effects and signal processors, no 6 1999. http://www.newmusicbox.org/third-person/oct99/effects.html, 1999.

10. D. Wessel. Instruments that learn, refined controllers, and source model loudspeakers. *Computer Music Journal*, 14(4):82–84, 1991.