Tongue 'n' Groove: An Ultrasound based Music Controller

Florian Vogt, Graeme McCaig, Mir Adnan Ali
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

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.

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

1. 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 modelling 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 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].

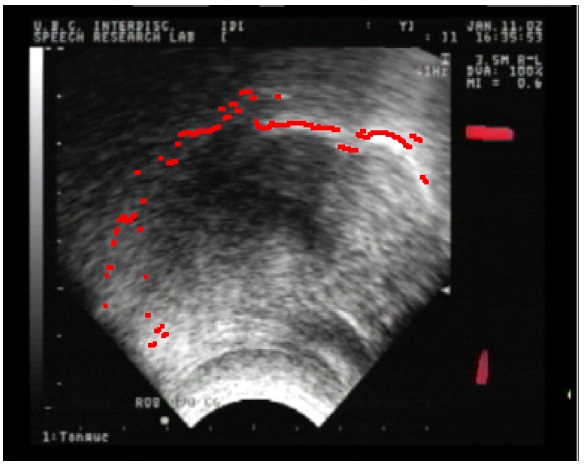

Figure 1: Tonguecontour reconstruction algorithm.

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 reconfig-

urable. 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 utilizes 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. 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.

2. 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 testing explores the following issues:

2.0.0.1 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.

2.0.0.2 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 temporal resolution of the video stream and software processing should be adequate so that time lag is not an obstacle to good control.

2.0.0.3 Learned/Evolved abilities of the tongue:

People have a pre-existing 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. People are especially skilled at stringing together spatio-temporal patterns of tongue motion, with a high degree of regularity (as in speech).

2.0.0.4 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.

2.0.0.5 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.

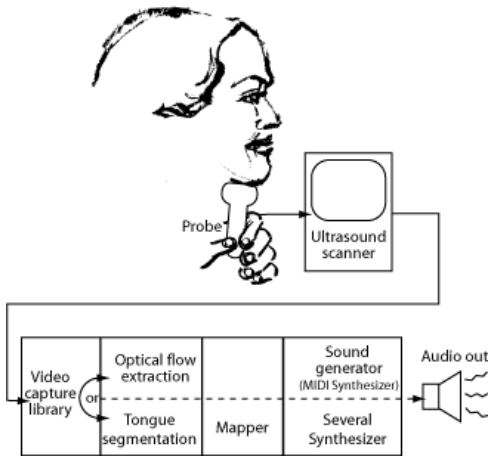

Figure 2: System diagram of the Tongue 'n' Groove

3. 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 which we send though a control parameter mapping program to a synthesis engine at the video frame rate. The mapping program allows automatic and manual adjustment of gain, bias and noise threshold separately for each image parameter. Output from the mapping program is used to control one of the various music synthesis algorithms we have investigated.

4. IMAGE PROCESSING

4.1 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. 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. By using the median of adjacent intensity maxima we reduce the number of outputs, but improve the noise robustness. 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.

We compensate the exponential spacial falloff of reflection sensitivity with a exponential function in our image extraction. We also minimize the noise problems by adding static background noise subtraction of the the mean of a series of calibration frames. This calibration feature can be triggered with a button press as an automatic procedure from the GUI.

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.

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

4.2 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.

5. CONTROL PARAMETER MAPPER

In our system architecture we chose to build a number of simple independent components. The control parameter mapper is one component, which allows direct control of the mapping of image to synthesis parameters. Each of the parameter channel can be manually controlled with sliders for gain, bias, threshold and a button to disable the parameter mapping to the output. In addition to the manual controls,we implemented a calibration function using a frame sequence. As a first step, with the tongue in rest position, we determine for each channel the thresholds based on the noise. (Tongue displacements which do not exceed the threshold produce no change in output.) Then in a second step with a frame sequence of extreme tongue motions we set the gains and biases. The separate mapper is very useful in investigation of the physical control abilities.

6. SOUND SYNTHESIS

We are experimenting with different music synthesis algorithms as output for the Tongue 'n' Groove. As discussed above, a good tongue controller will leverage the learned/evolved ability of the tongue to assist in speaking and eating. Therefore we see three categories of instrument as especially attractive for tongue control:

1. Instruments where the tongue plays a frequency-shaping role similar to its role in speech.
2. Instruments where the tongue is used to articulate repeated spatio-temporal phrases, as in talking.
3. Instruments which simulate the manipulation of objects in the mouth, as in eating.

So far we have developed five Tongue'n'Groove instruments, with the most successful fitting at least one of the above categories. By playing each instrument ourselves, we have made informal observations.

6.1 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 played through the internal MIDI synthesizer of a PC sound card. The instrument sound changes periodically according to a predefined cycle.

This instrument was a preliminary attempt to achieve basic sound output. It exhibited a low degree of control due to the noise present throughout the ultrasound image, causing unwanted triggers. The new Tongue-Motion instrument presents a better implementation of the motion-detection concept. However, the use of a predefined loop as non-controlled parameter variation is a potentially useful technique that we may incorporate into future instruments.

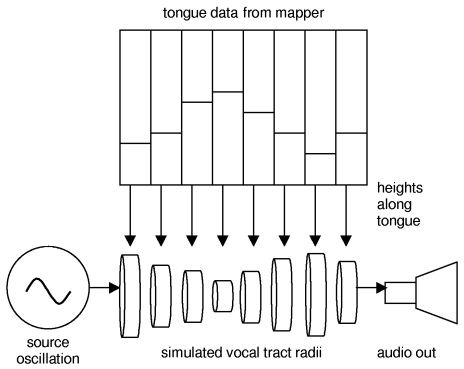

Figure 3: Sound Synthesis: Tongue-SPASM.

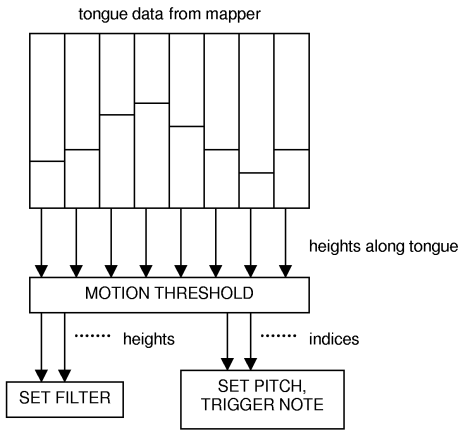

Figure 4: Sound Synthesis: Tongue-Motion.

6.1.0.6 Marble-Mouth

In this instrument, the tongue controls 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.

This mapping was our first attempt to simulate objects within the mouth. We found that influencing the bouncing period did not give a satisfying feeling of control, since this mapping does not provide continuous feedback of tongue behaviour (i.e. you must wait until the next impact to hear the effect of a change). The lack of pitch or timbre control makes for a sound that quickly grows boring.

6.2 Tongue-SPASM

The Tongue-SPASM instrument is based on Perry Cook's Singing Physical Articulatory Synthesis Model (SPASM)[1]. SPASM simulates human voice sounds, by modelling 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 (see Figure 3).

We adapted 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. An earlier version of Tongue-SPASM contained a bug which introduced uncontrolled, rhythmic pulsing to the output. We observed that this actually made the instrument more fun to play. This suggests that, as in the body, the tongue might be best used as a secondary controller that modifies a primary stream of musical information.

Tongue-SPASM leverages the familiarity of filter-shaping with the tongue. For this reason, the instrument has a more 'natural' feeling than the others. However, with our rough mapping into vocal tract space, the familiarity actually causes the problem of unexpected behaviour- a given tongue configuration does not make the sound a user would expect. This frustration can be removed with further refinement of the mapping from tongue measurements to vocal tract space.

6.3 Tongue-Motion

Tongue-Motion follows the motion-detection idea of Tongue-Scope, but with motion detection performed only on the detected tongue vector. This reduces the control noise. Different horizontal tongue points trigger notes of different pitch, with the height of the detected activity controlling a lowpass filter cutoff frequency (see Figure 4).

Tongue-Motion is difficult to control, because horizontal portions of the tongue cannot be easily wiggled independently. Thus the output has a limited musical range, giving a similar effect to someone randomly strumming the open strings on a guitar.

6.4 Tongue-Max

Tongue-Max employs a similar mapping to Tongue-Motion, making for interesting comparison. In Tongue-Max, only the tongue point with greatest (or optionally, least) height triggers output. As in Tongue-Motion, the index of that point determines note pitch, and the height of the point determines note timbre (see Figure 5). Thus, Tongue-Max shows the merits of relative-position-based control, rather than absolute-velocity-based control.

Tongue-Max seems to be the most satisfying and expressive instrument we have implemented. A variety of musical phrases can be produced, and a given phrase can be repeated with regularity. Tongue-Max follows the 'object manipulation' concept to some degree; one can imaging lifting an object to the roof of the mouth, or perhaps stroking over a set of virtual strings in the mouth. The success of Tongue-Max points to a strategy of dimension reduction,

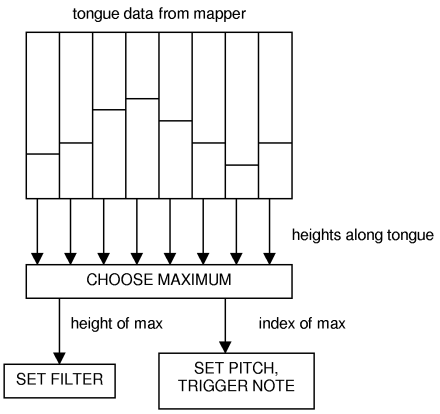

Figure 5: Sound Synthesis: Tongue-Max.

extracting one 2-dimensional point from the behaviour of eight 1-dimensional tongue points.

Both Tongue-Motion and Tongue-Max were also tested with the filter-control disabled. In each case this resulted in a decrease in musical interest and expressiveness. This could be attributed simply to the reduction of output dimensions, but it seemed to us that using the tongue for filter control felt intuitively 'right' as hypothesized earlier.

Overall, the Tongue'n'Groove instruments were rewarding to play, perhaps more fun than would be expected from the limited degree of control. This agrees with our hypothesis about the intimacy of using the tongue as a controller.

7. PROPOSED EVALUATION

We are planning to conduct a user study to determine the success of the Tongue 'n' Groove instruments. 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 various 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:

- 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.
- A single-point control test is a easy quantifiable benchmark to compare the system performance and its setting for controllability. Applying a measure, like the "fitts test"[**?**] for pointing devices, to the tongue only investigates one control aspect of the tongue and does not represent its full abilities.
- 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.

8. DISCUSSION

Our results are based on informal testing of our working prototype by a small number of users. We achieved tongue tracking performance for multiple control points of the tongue contour at video frame rate. By improving noise robustness we are able to track control points within 5 pixel accuracy at NTSC resolution. We identified the "single point control test" as a simple controllability benchmark for tongue controller. Performers are able to control a single point with little effort. A quantified study needs to be still conducted in a similar way as the Fitts test for pointing devices.

Further we found that performers had difficulties in controlling multiple tongue points independently, which suggests that the tongue is not suitable to control independent parameters by using a one-to-one mapping of tongue control points to independent sliders. A better way to think about the tongue is a "spatio-temporal contour controller"- i.e. many gestures of the tongue can be controlled accurately and reliably as we observe in speech production. Gesture modelling and mapping seems to present a promising avenue for further investigation of the tongue as a intimate music controller.

9. 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.

10. 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.