

Tongue 'n' Groove

Florian Vogt, Graeme McCaig, Mir Adnan Ali, Sidney Fels

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 through speaking, and has the potential for sensitive and fine control. Other systems for music control using the lips [6] have shown such properties.

Looking at the role of the tongue in speech modelling [2], [5], [4], 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 -I think the mouth harp uses more than the "tip" of the tongue. Also, instruments such as Mouthesizer and Talkbox use various elements of the human vocal tract to control or modulate sound. "Mouthesizer" uses the lips as the sole means of input. "Talkbox" utilizes a speaker placed

in the performer's mouth, which records the filtering effect of the mouth using an external microphone. Another related music controller, the Vocoder, 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. 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 with an image processing algorithm a tongue model in real-time. The mapping from the tongue model to the music controller is performed by computer software that analyzes the input video and generates the tongue model, which drives sound synthesis parameters. 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.

TONGUE 'N' GROOVE SYSTEM

Figure 1 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 reflective property and therefor 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 [3]. It calculates the amount of motion within each of 10 vertical bands of the tongue image. The other algorithm, described in the next section, 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. 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. 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 (type). 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.

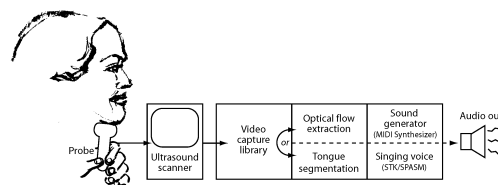


Figure 1: System diagram of the Tongue & Groove controller.



Fig

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 PC. As can be seen from figure 2, the algorithm is subject to a significant amount of noise and error, which causes unintended fluctuations in the musical output. We are currently testing simple techniques, such as outlier removal, to increase the accuracy of the algorithm. This will allow more precise control and improve the expressive potential of the Tongue'n'Groove. Further we are investigating the application of more computational expensive methods, such as discrete snakes and Kalman filtering, in the spacial 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 the tongues position as previously, this algorithm analyzes the tongues motion in ten horizontal segments. The flow extraction algorithm computes the motion intensity by taking the difference between consecutive image frames.

MAPPING OF SYNTHESIS PARAMETERS

The Tongue'n'Groove is a flexible musical controller, which will be tested with several different virtual instruments as output. Thinking about the mapping from a vector of tongue positions into a vector of musical control parameters, we have identified three broad categories, moving from a direct mapping to a more abstract mapping:

1. The vector of vocal-tract segment heights can be used to control the equivalent vector of heights in a physically modelled human singing voice synthesizer.
2. The vector of vocal-tract segment heights can be used to control filter shaping parameters in a non-vocal virtual instrument.

3. The vector of vocal-tract segment heights can be used to control a set of non-filter-related parameters in a non-vocal virtual instrument.

In our application we implemented, two instruments: Activity sound trigger and singing voice, which match the prior indicated first and third category.

SINGING VOICE

Using the Synthesis Toolkit (STK) code as a base, at least one virtual instrument from each category will be implemented and tested with the Tongue'n'Groove. The mapping depends on the synthesizer, but some mapping principles apply across all forms.

When the tongue controller is used to control a realistic vocal tract model, it is important to have accurate readings of tongue/hard palate positions. However, when tongue movements are used to control another instrument, the precise measurement of tongue position becomes less important. As long as a consistent mapping is applied, users will learn the relationship between tongue motion and sonic effect. For this reason, a fairly simple image-processing algorithm may be adequate for many types of musical control. This allows for real-time implementation using a standard Unix personal computer. In this case I the tongue position can be mapped in vocal tract diameter.

Bases on the noisy ultrasound image and minimized processing a fuzzy mapping into the combines temporal and spacial might be another way of approaching the problem.

ACTIVITY SOUND TRIGGER

This instruments intended for novice performers triggers send note information to a MIDI software synthesizer notes get triggered, when exceeding a threshold, with a velocity of the input parameter. Each of the parameters maps to a different note in an octave and after a number to triggers the current activated octave changes to a difference active octave.

CONCLUSION

We build a unique ultrasound-video-based tongue music controller, which can be adopted for a variety of instruments by choosing a Image processing to extract tongue features and map them to synthesis parameters. One of the two image processing algorithms extracts temporal, the other positional changes. We incorporated two instruments, the singing synthesizer SPASM and the MIDI activity Note trigger, to be controlled by the tongue. Musical performer tried our system and reported its was fun and compelling to perform. A ultrasound video display was found to improve tongue controllability for performance novices and enhanced the performance experience for spectators too.

REFERENCES

1. P. Cook. Spasm, a real-time vocal tract physical model controller and singer, the companion software synthesis system, 1993.
2. G. Fant. *Acoustic Theory of Speech Production*. S'Grovenhage, Mouton, 1960.
3. S. S. Fels and K. Mase. Iamascope: A graphical musical instrument. *Computers and Graphics*, 2(23):277–286, 1999. (original is pdf).
4. J. N. Holmes. Formant synthesizers: Cascade or parallel? *Speech Communication* 2, 1983.
5. D. H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustic Society of America*, 67:971–995, 1980.
6. Michael J. Lyons and Nobuji Tetsutani. Facing the Music: A Facial Action Controlled Musical Interface. In *Proceedings ACM CHI*, 2001.