The Expressionist: A Gestural Mapping Instrument for Voice and Multimedia Enrichment

J. Cecilia Wu, University of California, Santa Barbara, USA

Abstract: This paper presents the Expressionist, a three-dimensional human interface designed to enhance human vocal expression through interactive electronic voice filtering. The system is composed of a two-handed magnetic motion sensing controller and a computer program that processes the recorded voice of the operator based on the performer's body motion. This system is designed for electroacoustic vocal performers who are interested in multimedia live performances and wish to explore the possibilities of using body movement to enrich musical expression.

Keywords: Gesture Mapping, Musical Body Motion, Human-Computer Interaction, Virtual-Reality, Real-Time Voice Filtering, ChucK, CHAI3D

Introduction

The desire for musical expression runs deeply through human cultures; although styles vary considerably, music is often thought of as a universal language. As new technologies have appeared, inventors and musicians have been driven to apply new concepts and ideas to improve musical instruments or create entirely new means of controlling and generating musical sounds. Recently, with the explosion of digital technology, the computer has become an essential tool for creating, manipulating, and analyzing sound. Its precision, high computational performances and capacity to generate almost any form of audio signal make it a compelling platform for musical expression and experimentation. Moreover, the availability of many new forms of human interfaces ranging from camera trackers to tactile displays offers an infinite spectrum of possibilities for developing new techniques that map human gestures on sophisticated audio events. Although many approaches have been proposed to use gesture and movement perception to drive electronic instruments (Winkler 1995, 1998; Hunt et al. 2000; Knapp and Cook 2005; Machover and Chung 1989; Overholt 2001), little attention has been dedicated to exploring the use of digitized body motion for enhancing the musical expression of the human voice in real-time.

Related Work

Previous work in gesture tracking can be categorized in two groups: movement sensor-based and position tracking based. Movement sensor-based recognition is most suitable for mobile applications where the user may traverse large workspaces. These systems typically rely on small accelerometers or joint sensors that capture the motions of the human body. Such systems have been mounted on instruments and used for musical performance conducting systems (Rekimoto 2001). Furthermore, wearable devices such as glove-based devices designed for capturing sign language (Hideyuki and Hashimoto 1998; Sawada 2001) have demonstrated the capacity to even sense small finger motions. Tsukuda et al. (2002) introduced the Ubi-finger, a wearable device that uses acceleration, touch and bend sensors to capture a fixed set of hand gestures.

In contrast, position-tracking based devices are used for stationary applications and generally require a tracker or camera to be setup at a fixed location and sometimes calibrated to the environment. Oikonomidis et al. (2011) present different strategies for recovering and tracking the 3D position, orientation and full articulation of a human hand from marker-less visual observations obtained by a 3D Kinect sensor. More advanced multi camera systems such as the Vicon 8 have also been used to capture human motion during large stage performances where optical occlusions may occur (Dobrian and Bevilacqua 2003). In the area of music expression,

The International Journal of New Media, Technology and the Arts
Volume 10, Issue 1, 2015, www.artsinsociety.com, ISSN 2326-9987

these categories of devices can be used in various ways to create new instruments and new ways to enhance the expression of audio or visual performances. Butch et al. (1997) present different gesture mapping strategies to control MIDI instruments or synthesizers. Taylor et al. (2006) further explore the use of motion tracking techniques to extract musical feature data to generate responsive imagery.

Design and Motivation

Human beings are expressive. When interacting and connecting with other people, we use facial expressions, body language, and speech to communicate our thoughts and feelings to other individuals. Expressions occur outside the speaker's control. During musical performances, audiences often connect to the performer and mentally model thoughts and feelings in the performer's mind (Gabrielsson and Juslin 1996), in what is called emotional contagion (Hatfield et al. 1993). In addition, humans read faces, bodies, and voices; this ability begins at infancy and is refined to an art by adulthood (Darwin 1872). In this paper we introduce the Expressionist, a novel 3D interface that captures human body motions to enrich vocal expression during live performances. The aspiration of this project is to capture the natural forms of human expression and embody them musically in real-time. By capturing body gestures, we also explore new methods for directing other instruments or musicians participating in the performance.

System Architecture

The proposed system is composed of a 3D human motion tracker and microphone that captures the performer's voice and hand gestures. Hand gestures are relayed to a computer to extract absolute position, orientation, velocity and acceleration values. These measurements are then converted into audio events that activate in real-time vocal filters programmed according to the desired musical style. Finally, the resulting audio signal is transmitted to an amplifier for stage performance. In Figure 1 we illustrate the overall architecture of the system. The different parts of the systems are developed in the following paragraphs.

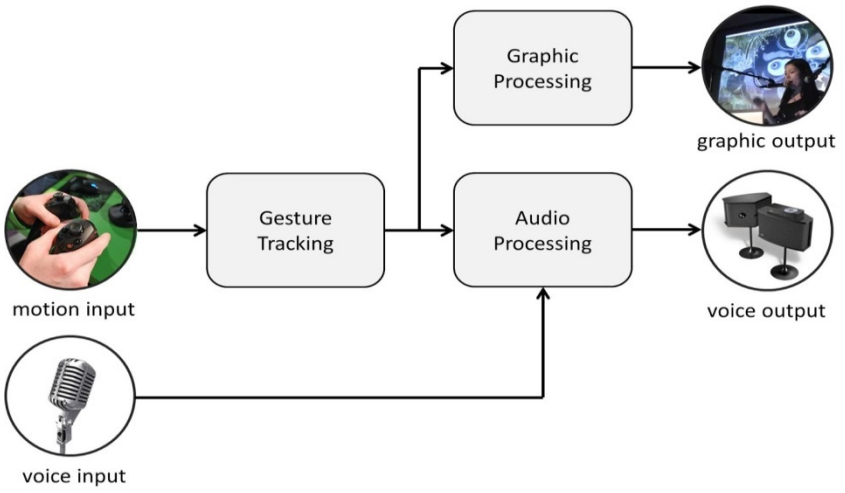

Figure 1: The Expressionist: a general overview of the system.

3D Human Interface

The vast majority of motion-controllers available today rely on small accelerometers that detect hand motions generated by the operator. Although such inertial sensors are used in almost any game controller or smart phone sold on the market today, their precision remains limited, and they lack the ability to track absolute positions in 3D space. For this project we selected a new type of interface called the Razer Hydra. As compared to accelerometer based, ultrasonic or camera driven motion sensors, the Razer Hydra uses magnetic motion sensing. Magnetic motion sensing enables this new interface to track in true 3D space for absolute position and orientation with its magnetic field. This technology also allows for full six degree-of-freedom movement tracking with sub millimeter accuracy, while eliminating the need for a line of sight to operate.

The overall user interface of the Expressionist adopts a paired system, with a base station emitting a low energy magnetic field that both hand controllers record to estimate their precise position and orientation. The controllers and the base station each contain three magnetic coils. These coils work in tandem with the amplification circuitry, digital signal processor and positioning algorithm to translate field data into position and orientation data. Because of the real-time nature of the technology, the interface delivers near instant response between the hand motions generated by the operator and the actions commanded by the computer. An illustration of the device is presented in figure 2.

Figure 2: The six degree-of-freedom Razer Hydra interface. (Courtesy of Sixense TrueMotion)

Gesture Tracking and Audio Processing

The Expressionist is composed of two applications that operate on a single laptop computer and handle both the gesture tracking and audio processing events. The gesture tracking application connects to the Razer Hydra interface and acquires at 100 times per second the position, orientation, velocity and acceleration values of each input device. Additionally, the states of the input switches and configuration of the miniature joysticks mounted at the extremities of both end-effectors are also retrieved. The raw signals are then processed using a first order filter and converted to a MIDI protocol before being broadcast to the audio stage. The audio processing application, designed using the ChucK language (Wang and Cook 2003, 2008), interprets MIDI signals from the gesture-tracking module and activates a selection of voice filters according to the desired vocal performance. The application communicates all input and output of audio signals using a Roland UA-101 interface, and supports multiple channels that include both voice and (optionally) electronic instruments.

Interactive Visual Design

Video design or projection design is a creative field of stagecraft. It is concerned with the creation and integration of film and motion graphics into the fields of theatre, opera, dance, fashion shows, concerts and other live events (Taylor 2006). Video design has only recently gained recognition as a separate creative field. The creation of visuals for live music performances bears close resemblance to music videos, but is typically meant to be displayed as

back plate imagery that adds a visual component to the music performed onstage. Graphic images, light shows, and/or interactive animations can be used in combination with audio tracks to convey setting and place, action, and atmosphere of a composition.

As an extension to our system, we developed an interface to easily animate two- and three-dimensional interactive and dynamic environments that respond to hand motions captured by the 3D tracker. Our display system builds upon the CHAI3D framework, an open source set of C++ libraries designed to create sophisticated applications that combine computer haptics, visualization and dynamic simulation (Conti 2003). The software libraries support many commercially available multi degree-of-freedom input devices, including the Razor Hydra. The framework provides a convenient interface for loading image and 3D model files. Furthermore the libraries also support tools to easily create complex physical systems that include particle, rigid body and deformable-based models. An XML file format is used to describe scene graphs and animation sequences that are loaded by the graphic display module. These files describe the objects composing the virtual environment, their physical properties as well as the interactions, events, and behaviors that can be triggered by the performer.

Gesture Mapping Strategies

To intuitively operate the different voice filters of the audio processing stage, we identified a set of body postures that express different emotional states of the performer. A selection of these postures is illustrated in figure 3. First reading the absolute position and orientation of the magnetic hand controllers and then computing their relative position and angular offsets, postures are distinguished. Once a posture has been correctly identified, a state machine is used to smoothly transition from one filter selection to another. The interpolation reduces the occurrence of acoustic clips that can typically occur during sudden discrete changes. According to the desired effects and hand movements of the performer, different filter functions are used to modify the audio signal. Each filter may alter different parameters of the voice signal. These parameters may include for instance the amplitude, frequency and phases, reverberation, or the spread of the sound in an auditorium. It is clear that the relationship between the gestural variables and the sound effects vary greatly from the style of performance.

Figure 3: Mapping body movement to vocal expressions.

Experimental Results

In the following paragraphs we illustrate a series of vocal and instrumental performances that were presented at Stanford University. For each piece, the body motion of the vocalist was captured in real-time and processed to enhance sound expression.

"High & Low" – A Duo Performance

To evaluate some of the possibilities offered by the system, we conducted an improvisation piece between a vocalist and musician playing the Celleto. We selected a divergent gesture mapping strategy, which allowed the vocalist to control parameters of her voice. During the improvisation, the vocalist triggered sound effects and controlled the output of the Celleto. A pitch-shift filter was applied to create a second layer of voice to emulate Tibetan throat singing. Sound panning and echoes were controlled in real-time by the hand gestures captured from the singer. A picture of the musical performance is illustrated in figure 4.

Figure 4: Onstage performance. Hand gestures from the vocalist are captured using two Razor Hydra 6-DOF interfaces.

"Mandala" – A Live Performance with Graphic Expression

The word "Mandala" comes from Sanskrit, meaning "circle." It represents wholeness, and can be seen as a model for the organizational structure of life itself. In Tibet, as part of a spiritual practice, monks create mandalas with colored sand (Bryant 2003). The creation of a sand mandala may require days or weeks to complete. When finished, the monks gather in a colorful ceremony, chanting in deep tones (Tibetan throat singing) as they sweep their Mandala sand spirally back into nature. This symbolizes the impermanence of life and the world. As part of a musical performance created at Stanford University, an interactive display of a sand Mandala was choreographed with the vocal performance. Near the end of the musical piece, hand motions from the singer were directed to initiate the destruction of the virtual Mandala. Our graphic and dynamic modeling framework was used to simulate in real-time the physical interaction between the captured hand motions of the performer and the small mass particles composing the virtual Mandala.

Figure 5: Construction and destruction of a sand Mandala.

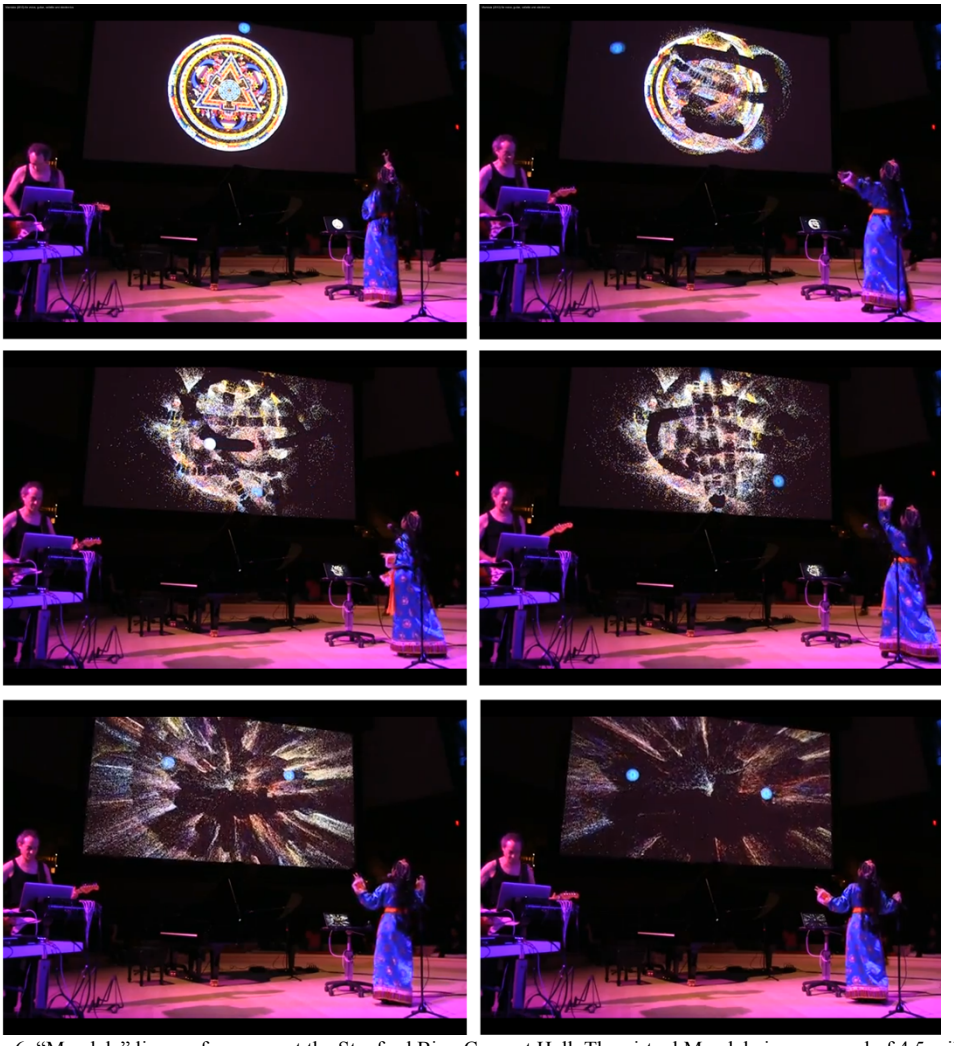

Figure 6: "Mandala" live performance at the Stanford Bing Concert Hall. The virtual Mandala is composed of 4.5 million sand particles that are dynamically updated 100 times a second. Force interactions are modeled in real-time by capturing the position of the performer's hands and by computing collision events with and between the individual sand particles.

"Synergy" – Bringing Multiple Performers Together

This piece explores an arrangement composed of three performers where multi body movement tracking is used to enhance human voice and produce environmental sounds. The performance combines the acts of a vocalist and two musicians who move around the stage as the music evolves. The vocalist stands in place and sends a combination of whispers and extended vocal input. Hand gestures are tracked in real-time and vocal signals processed. The two mobile performers accompany the vocalist by creating various computer-generated sounds using their own body gestures. The wireless controllers carried by the musicians' track their own gestures and process the output sounds in real-time. The ChucK and Pure Data frameworks were used to develop and implement the different acoustic filters and environmental sound generators. The main composition elements in this piece include a human voice, a low-pitched heartbeat, a fuzzy static, a low-pitched "monk" drone, and an electric spark.

Figure 7: "Synergy" live performance at Stanford University.

Conclusion

In this paper we presented the design for a new interface to explore the imaginative relationships between singing and natural emotional expression by using gesture control to modulate the texture and shaping of the voice. For its portability, lack of sensitivity to occlusions, and ease of use, we selected a magnetic motion tracker with absolute position sensing capabilities and sub millimeter accuracy. We developed a software application to convert gesture motions into audio events that activate and modulate sound filters in real-time. Different interaction metaphors were created and evaluated during a series of live performances that are available for viewing online. These preliminary results validated the intuitive design of the Expressionist and its capacity to respond with real-time performances to vocal and audio signals. Furthermore, these results also confirmed our original idea to explore natural human gesture as an effective way to instinctively enrich voice expression.

REFERENCES

Bryant, Barry. The wheel of time sand mandala: Visual scripture of Tibetan Buddhism. Snow Lion Publications, 2003.

Conti, François. The CHAI libraries. No. LSRO2-CONF-2003-003. 2003.

Darwin, Charles. The expression of the emotions in man and animals. Oxford University Press, 1998.

Dobrian, Christopher, and Frédéric Bevilacqua. "Gestural control of music: using the vicon 8 motion capture system." In Proceedings of the 2003 conference on New interfaces for musical expression, pp. 161-163. National University of Singapore, 2003.

Gabrielsson, Alf, and Patrik N. Juslin. "Emotional expression in music performance: Between the performer's intention and the listener's experience."Psychology of music 24, no. 1 (1996): 68-91.

Hatfield, Elaine, and John T. Cacioppo. Emotional contagion. Cambridge university press, 1994.

Sawada, Hideyuki, and Shuji Hashimoto. "Gesture recognition using an acceleration sensor and its application to musical performance control."Electronics and Communications in Japan (Part III: Fundamental Electronic Science) 80, no. 5 (1997): 9-17.

Hunt, Andy, Marcelo Wanderley, and Ross Kirk. "Towards a model for instrumental mapping in expert musical interaction." In Proceedings of the 2000 International Computer Music Conference, pp. 209-212. 2000.

Knapp, R. Benjamin, and Perry R. Cook. "The integral music controller: introducing a direct emotional interface to gestural control of sound synthesis." In Proceedings of the International Computer Music Conference (ICMC), pp. 4-9. 2005.

Machover, T., and J. Chung. "HyperInstruments: Musically Intelligent and Interactive performance and Creativity Systems. Intl." In Computer Music Conference (Columbus, Ohio, 1989).

Oikonomidis, I. Kyriazis, N. Argyros, and A. Efficient Model-based. "3D Tracking of Hand Articulations using Kinect." In Proceedings of the British Machine Vision Conference 2011Dundee UK, pp. 101-1.

Overholt, Dan. "The MATRIX: a novel controller for musical expression." In Proceedings of the 2001 conference on New interfaces for musical expression, pp. 1-4. National University of Singapore, 2001.

Rekimoto, Jun. "Gesturewrist and gesturepad: Unobtrusive wearable interaction devices." In Wearable Computers, 2001. Proceedings. Fifth International Symposium on, pp. 21-27. IEEE, 2001.

Rovan, Joseph Butch, Marcelo M. Wanderley, Shlomo Dubnov, and Philippe Depalle. "Instrumental gestural mapping strategies as expressivity determinants in computer music performance." In Proceedings of Kansei-The Technology of Emotion Workshop, pp. 3-4. 1997.

Taylor, Robyn, Pierre Boulanger, and Daniel Torres. "Real-Time Music Visualization Using Responsive Imagery." In 8th International Conference on Virtual Reality, pp. 26-30. 2006.

Tsukadaa, Koji, and Michiaki Yasumurab. "Ubi-finger: Gesture input device for mobile use." In Ubicomp 2001 Informal Companion Proceedings, p. 11. 2001.

Wanderley, Marcelo M. "Gestural control of music." In International Workshop Human Supervision and Control in Engineering and Music, pp. 632-644. 2001.

Wang, Ge, and Perry R. Adviser-Cook. The chuck audio programming language. a strongly-timed and on-the-fly environ/mentality. Princeton University, 2008.

Wang, Ge, and Perry R. Cook. "ChucK: A concurrent, on-the-fly audio programming language." In Proceedings of the International Computer Music Conference, pp. 219-226. Singapore: International Computer Music Association (ICMA), 2003.

Winkler, Todd. "Making motion musical: Gesture mapping strategies for interactive computer music." In ICMC Proceedings, pp. 261-264. 1995.

Winkler, Todd. "Motion-sensing music: Artistic and technical challenges in two works for dance." In Proceedings of the International Computer Music Conference. 1998.

ABOUT THE AUTHOR

J.Cecilia Wu: PhD Student, Department of Media Arts and Technology, University of California, Santa Barbara, California, USA