The Multimodal Music Stand

Bo Bell‡ Jim Kleban† Dan Overholt‡ Lance Putnam‡ John Thompson‡ JoAnn Kuchera-Morin‡

†Electrical and Computer Engineering
‡Media Arts and Technology
hci@mat.ucsb.edu

California NanoSystems Institute, Room 2001
University of California – Santa Barbara
Santa Barbara, CA 93106

ABSTRACT

We present the Multimodal Music Stand (MMMS) for the untethered sensing of performance gestures and the interactive control of music. Using e-field sensing, audio analysis, and computer vision, the MMMS captures a performer's continuous expressive gestures and robustly identifies discrete cues in a musical performance. Continuous and discrete gestures are sent to an interactive music system featuring custom designed software that performs real-time spectral transformation of audio.

Keywords

Multimodal, interactivity, computer vision, e-field sensing, untethered control.

1. INTRODUCTION

New musical instruments need to be accessible, offer expert control, and develop a repertoire in order to survive. While custom interfaces based on a single instrument are useful in a particular context, the Multimodal Music Stand (MMMS) is designed for use in many different contexts. Because of this approach, the MMMS allows any traditional performer to extend their technique without physically extending their instrument. In this way, it has the potential to become a breakthrough in interactive electroacoustic music performance.

Many traditional performers are unwilling to introduce any prosthetics requiring physical manipulation, citing interference with their normal manner of playing [1]. The MMMS provides untethered interaction allowing expressive control of signal processing while maintaining performers' traditional instrument expertise. It augments the performance space, rather than the instrument itself, allowing touch-free sensing and the ability to capture the expressive bodily movements of the performer.

Our approach is to use non-contact sensors, specifically microphones, cameras, and e-field sensors, embedded in the MMMS. Using this array of sensors, the music stand provides data to a multimodal analysis system that identifies a set of predetermined gestures that are then mapped to real-time audio synthesis and transformation processes.

Our primary goals are to:

- Enable untethered interactive electroacoustic performance
- Take a generalized approach towards instrument augmentation (i.e. allow extended performance techniques without instrument modifications)
- Capture performance gestures and map them to audio synthesis and transformation parameters
- Use multimodal analysis to reinforce cue detection

2. BACKGROUND

2.1 Similar Musical Interfaces

In 1919, Léon Theremin invented the world's first non-tactile electronic instrument, the Theremin, that sensed the distance to a performer's hands using changes in capacitance. The MMMS adopts this technique of sensing and expands it to be used in conjunction with audio analysis and computer vision techniques.

The music stand is a good candidate for an unobtrusive alternative controller in a traditional music setting. A related interface is Hewitt's Extended Mic-stand Interface Controller (e-Mic) [2]. The e-Mic, while providing an alternative interface, uses a microphone stand to provide interactive control to a vocalist. While similar to the MMMS in its use of a familiar musical stage item, the e-Mic requires the performer to physically manipulate controls attached to the microphone stand.

Other approaches to creating augmented music stands have been taken in the realm of traditional music as well as in new media arts. Two primary commercial augmented music stands are available for traditional musicians: the muse [3] and eStand [4]. These music stands feature such attributes as graphical display of scores, the ability to annotate digitally and save those annotations, automatic page turns, built-in metronomes and tuners, and networkability. These stands serve a practical purpose, but are not designed as expressive gestural interfaces.

Leaning toward new media arts, MICON [5] is an installation based on a music stand for interactive conducting of pre-recorded orchestral audio and video streams. Through gesture recognition

software, amateur participants are able to control the tempo of playback as well as the dynamic balance between instruments using a Buchla Lightning II infrared baton. MICON features graphical display of the musical score and automatic page turning. While MICON is intended for musical conducting within the context of an interactive installation, the MMMS is intended for professional musicians in the context of an interactive performance.

2.2 Gesture In Instrumental Music

The MMMS is primarily concerned with capturing ancillary performance gestures, i.e. gestures that are not directly related to the actual production of sound. Wanderley and Cadoz [6] have shown that certain ancillary gestures are repeatable and consistent to a particular piece. Interestingly, Wanderley [7] points out that these gestures, while not altering the sound production of the instrument itself, affect the way the sound is heard by an audience. Thus, the gestures of performers carry important details that can be used to inform the interactive music system.

Computer vision tools such as EyesWeb [8] and EyeCon [9] are capable of identifying such ancillary gestures on basic (syntactic) and advanced (semantic) levels. Thus far, these tools have mainly been used to analyze dancers' movements and control musical parameters [10]. Similar work was attempted on the syntactic level by Qian et al. [11], and on the semantic level by Modler and Myatt [12]. We build upon these concepts with the MMMS, both by defining a lexicon for control and by mapping these to continuous gestures.

3. SYSTEM OVERVIEW

The MMMS system consists of three parts. The *Multimodal Input Analysis* segment involves electronic field sensing, visual, and audio analysis. These methods gather information about the musician's performance. The *Multimodal Detection Layer* analyzes this input data, sending trigger messages according to user defined conditions. The *Audio Synthesis and Transformation* engine listens for triggers and continuous controls that affect how it creates and/or alters the musical accompaniment.

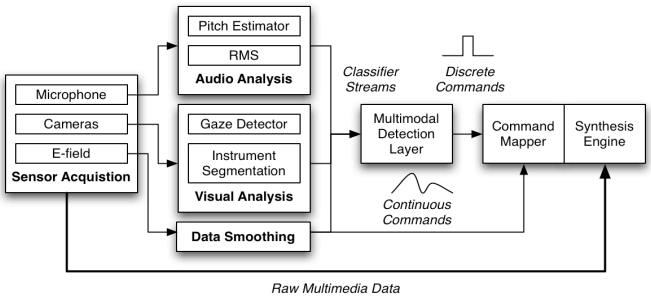

Figure 1: Multimodal Music Stand system model.

3.1 Multimodal Input Analysis

The Multimodal Music Stand incorporates four electric field sensors [13]. These are designed to capture the physical bodily (or instrumental) gestures of the performer via sensor antennas. The synthesis engine (detailed below), uses these as input sources for the control of continuously variable musical parameters. This gives traditional performers an added dimension of musical control: they can directly influence the interactive computer music aspects of a performance without modifying their instrument or tethering themselves with wires.

Figure 2: The prototype MMMS with e-field sensor antennas mounted at the corners of the stand.

The electric field sensing technique is based on Theremin circuit topology [14], but done entirely in the digital domain [15]. The original circuit design for the Theremin used an analog heterodyning technique. Only the front end of this circuit is used in the MMMS. We combine this with the measurement of pulse timing information using custom-written firmware on the CREATE USB Interface [16].

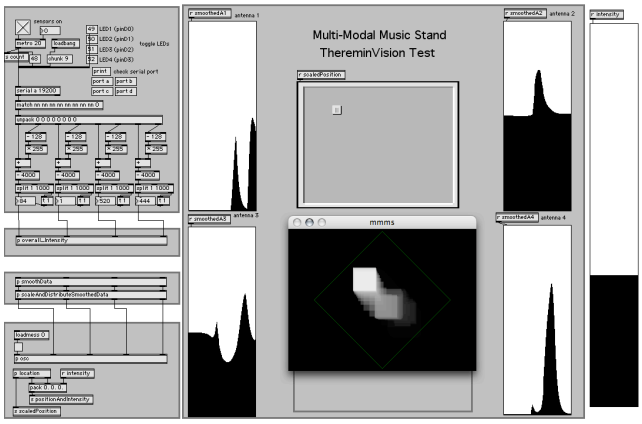

Figure 3: The software in Max/MSP/Jitter to receive, filter, and process e-field sensor data / send OSC.

The data from each of the independent e-field sensors is received in Max/MSP/Jitter, analyzed, and sent out via Open Sound Control (OSC). Using four channels of sensing makes it possible for the MMMS to track the performer in three dimensions. The overall intensity of all four antennas determines the z-axis gestural input. The incoming sensor data is smoothed with a simple median filter in Max and visualized in Jitter.

3.2 Multimodal Detection Layer

The MMMS allows musicians to transcend traditional performance by providing a sensing space in which to sculpt their work. Multimodality, specifically combining the use of cameras, microphones and e-field sensing, adds robustness in detecting discrete cues for control and provides additional degrees of freedom in mapping continuous gestures for accompaniment.

Multimodal analysis has been applied successfully in areas such as audio-visual speech recognition, multimedia content analysis, and tracking for interactive installations. The MMMS fuses gesture detection in the e-field, visual, and audio domains, along with combinatorial dependencies and predefined timing knowledge, to mitigate control errors. In addition, feedback from the system to the performer during rehearsal and performance periods helps to remove problems of missed and false triggers.

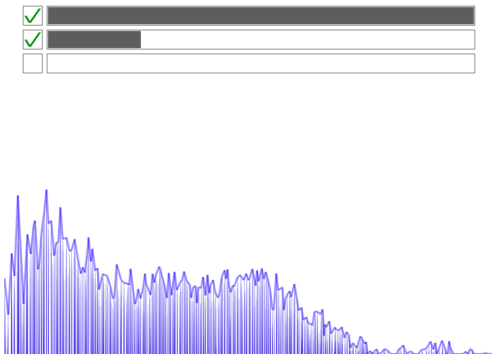

Figure 4. Graphical feedback of performance cues.

The first real test bed application for our system is an interactive flute piece *timeandagain* by JoAnn Kuchera-Morin. By observing flutists playing in duet, we identified a lexicon of gestures in the visual domain. We then developed computer vision algorithms to detect the common cueing mechanism of looking in a particular direction, and also to gather the expressive dipping and raising movements of the flute. While the general-purpose nature of the MMMS allows any common computer vision techniques to be used, many compositions with use techniques tailored to the piece.

The MMMS uses two cameras to detect the aforementioned gestures. One camera mounted atop the stand, segments and tracks the flute. Another camera placed to the side of the stand detects a performer's gaze to the side.

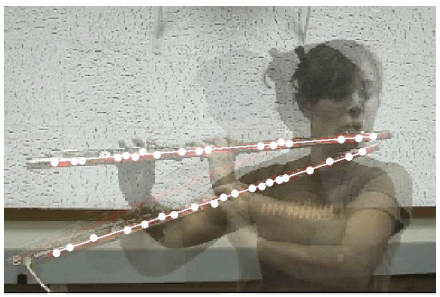

Figure 5. Flute segmentation using RANSAC.

For *timeandagain*, the gestures of flute dipping and raising are defined, respectively, as the angle of the flute below or above a certain threshold. The first step in determining the flute angle is to isolate the flute in the scene. We approach the problem of flute segmentation by noting that the flute itself is often the only specular object in the scene. Thresholding the brightness domain in HSB color space removes most of the performer and background. Because the flute is not purely reflective, and because there are occlusions due to the performer's hands, the thresholding operation returns specular "blobs" along the line of the flute, along with noise. We reduce these blobs to their centroids, and then perform linear regression using the random sample consensus (RANSAC) algorithm [17]. The result of these calculations, processed by our custom Max/MSP/Jitter object (jit.findflute), is a robust calculation of the angle of the flute.

For the second gesture, head turning, we detect cueing motions that involve gaze and eye contact. Sensing eye contact using a retinal tracking system would prove unwieldy in a performance environment. Instead, the head turning gesture is recognized using a face detection algorithm from a laterally placed camera. An OpenCV implementation of the Viola-Jones [18] method detects the frontal pose of the performer's face as she turns slightly to make a cue.

While the visual analysis is accurate, by itself it is not reliable enough for the stringent demands of cueing in the context of a musical performance. Our solution is to define gestures as multimodal aggregates consisting of visual, audio, and e-field features. To derive audio features, a condenser microphone mounted on the stand sends audio from the flute into a computer running Tristan Jehan's "analyzer~" Max/MSP object [19]. We use OSC to send two audio features (pitch and RMS amplitude), along with the visual features, to the multimodal detection layer where higher-level gestures are derived.

The multimodal detection layer is software that integrates the audio and visual classification results for gesture recognition. Users define the types of gestures occurring in the piece. The gestures can be restricted to certain time windows or allowed during any time in the piece. Furthermore, gestures can be defined to occur in multiple modalities together, such as, a gaze to the side along with a particular loudness of playing. Upon detection of the predefined gesture, OSC messages are sent to the synthesis and transformation machine.

3.3 Audio Synthesis and Transformation

The synthesis engine uses a client-server control model, similar to SuperCollider Server [20]. The server receives discrete commands from the multimodal detection layer and continuous commands from the multimodal analysis layer. Discrete commands can reset state, begin playback of sample buffers, or start/stop transformation effects. This allows the performer to "set and forget" specific processes, giving him/her more time to focus on other musical tasks and thus avoid problems with control simultaneity. Continuous commands can position sounds or be mapped to parameters of a variety of spectral processing techniques, such as band-thinning or spectral blurring. The nature of the software allows a multitude of generic array transformations applicable to both time- and frequency-domain data.

The synthesis server was built using *synz*, a C++ library of frequently encountered signal generation and transformation patterns [21]. *synz* shares many design ideologies of existing signal-processing frameworks such as platform-independence, ease-of-use, and flexibility. In addition, *synz* stresses computational efficiency together with modularity and generic programming to make it a solid foundation for higher-level dataflow models, such as 4MS [22], and plug-in architectures. The core functions of *synz* are divided into scalar, array, memory, table filling and random-based operations. Whenever possible *synz* use generics through C++ templates to enhance algorithm/object reuse. Signal objects use single-element evaluation and return methods called 'next*()', similar to STK's 'tick()' [23], and thus do not enforce any buffering or multi-channel processing scheme. This maximizes flexibility in building arbitrary signal flow graphs. Furthermore, the methods are inlined so as not to incur additional penalties for function calls.

4. CONCLUSIONS AND FUTURE WORK

The Multimodal Music Stand is both a musical device and a research platform. It was developed with the goal of realizing interactive musical works in ways that increased expressivity without hindering performers. We hope to expand its repertoire and collaborate with other expert performers. Through these collaborations, we want to bridge the gap between traditional performers and the electroacoustic community.

In the future we hope to develop optimal multimodal fusion schemes enabling richer and more natural musical human-computer interaction. This would move us closer to a more generalized approach to instrument augmentation.

5. ACKNOWLEDGEMENT

The authors would like to thank JoAnn Kuchera-Morin and B.S. Manjunath for their oversight in this project. Support was provided by IGERT NSF Grant# DGE-0221713.

6. REFERENCES

[1] McNutt, E. "Performing electroacoustic music: A wider view of interactivity." *Organised Sound*, vol. 8(3), pp 297-304, 2003.

[2] Hewitt, D. and I. Stevenson. "E-mic: Extended mic-stand interface controller." *Proc. of the 2003 Conference on New Interfaces for Musical Expression*, Montreal, pp. 122-128, 2003.

[3] Graefe, C., D. Wahila, J. Maguire. "muse: A digital music stand for symphony musicians." *Interactions*, vol. 3:3, pp. 26-35, 1996.

[4] Sitrick, D. "System and methodology for musical communication and display." U.S. patent 7,157,638 (January 27, 2000).

[5] Borchers, J., A. Hadjakos, and M. Muhlhauser. "MICON: A Music Stand for Interactive Conducting." *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression,* Paris, pp. 254-9, 2006.

[6] Cadoz, C. and M. Wanderley. "Gesture-music." M. Wanderley and M. Battier, eds. *Trends in Gestural Control of Music*, Paris: IRCAM - Centre Pompidou, 2000.

[7] Wanderley, M. "Interaction musicien-instrument: Application au contrôle gestuel de la synthèse sonore." Ph.D. thesis, University Paris 6, Paris, France, 2001.

[8] Camurri, A., G. De Poli, M. Leman, and G. Volpe. "Communicating expressiveness and affect in multimodal interactive systems." *IEEE Multimedia Magazine*, 12:1, pp. 43-53, 2005.

[9] Wechlser, R., Weiss, F., and Dowling, P. "EyeCon -- A motion sensing tool for creating interactive dance, music, and video projections." *Proc. of the SSAISB Convention*, Leeds England, 2004.

[10] Camurri, A., B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. "Multimodal analysis of expressive gesture in music and dance performances." A. Camurri and G. Volpe, eds. *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop*, Berlin: Springer-Verlag, 2004.

[11] Qian, G., F. Guo, T. Ingalls, L.Olsen, J. James, and T. Rikakis. "A gesture-driven multimodal interactive dance system." Paper presented at ICME 2004.

[12] Modler, P. and T. Myatt. "A video system for recognizing gestures by artificial neural networks for expressive musical control." A. Camurri and G. Volpe, eds. *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop*, Berlin: Springer-Verlag, 2004.

[13] Paradiso, J., and N. Gershenfeld. "Musical applications of electric field sensing." *Computer Music Journal*, vol. 26:2, pp. 69-89, 1997.

[14] Smirnov, A. "Music and gesture: Sensor technologies in interactive music and the THEREMIN based space control systems." *Proc. of the 2000 International Computer Music Conference*, Berlin, pp. 511-4, 2000.

[15] Fritz, T. "ThereminVision II Instruction manual" http://thereminvision.com/version-2/TV-II-index.html.

[16] Overholt, D. "Musical interaction design with the CREATE USB Interface: Teaching HCI with CUIs instead of GUIs." *Proc. of the 2006 International Computer Music Conference*, New Orleans, 2006.

[17] Fischler, M. A., R. C. Bolles. "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography." *Comm. of the ACM*, Vol 24, pp. 381-395, 1981.

[18] Viola, P. and M. Jones. "Robust real-time object detection." *Technical Report* 2001/01, Compaq CRL, February 2001.

[19] Jehan, T. "Tristan Jehan's Max/MSP Stuff." http://web.media.mit.edu/~tristan/maxmsp.html.

[20] McCartney, J. "Rethinking the computer music language: SuperCollider." *Computer Music Journal*, vol. 26:4, pp. 61-68, 2002.

[21] Putnam, L. "synz." http://www.uweb.ucsb.edu/~ljputnam/synz.html (accessed January 2007).

[22] Amatriain, X. and S. Pope. "An object-oriented metamodel for multimedia processing." *ACM Transactions on Multimedia Computing, Communications and Applications*, in press, 2006.

[23] Cook, P. R. "The Synthesis Toolkit (STK)." *Proc. of the 1999 International Computer Music Conference*, Beijing, China, 1999.