A FRAMEWORK FOR SONIFICATION OF VICON MOTION CAPTURE DATA

Ajay Kapur[1,2,3,4] *George Tzanetakis*[2,3] *Naznin Virji-Babul*[4] *Ge Wang*[5] *Perry R. Cook*[5,6]

Departments of Electrical Engineering[1], Computer Science[2], Music[3] and Psychology[4]
University of Victoria
Victoria, British Columbia, Canada

Department of Computer Science[5] (and Music[6])
Princeton University
Princeton, New Jersey, USA

ABSTRACT

This paper describes experiments on sonifying data obtained using the *VICON* motion capture system. The main goal is to build the necessary infrastructure in order to be able to map motion parameters of the human body to sound. For sonification the following three software frameworks were used: *Marsyas*, traditionally used for music information retrieval with audio analysis and synthesis, *CHUCK*, an on-the-fly real-time synthesis language, and *Synthesis Toolkit (STK)*, a toolkit for sound synthesis that includes many physical models of instruments and sounds. An interesting possibility is the use of motion capture data to control parameters of digital audio effects. In order to experiment with the system, different types of motion data were collected. These include traditional performance on musical instruments, acting out emotions as well as data from individuals having impairments in sensor motor coordination. Rhythmic motion (i.e. walking) although complex, can be highly periodic and maps quite naturally to sound. We hope that this work will eventually assist patients in identifying and correcting problems related to motor coordination through sound.

1. INTRODUCTION

The motion of the human body is a rich source of information, containing intricacies that needs be analyzed. The goal of this work is to build the necessary infrastructure to study the use of sonification for understanding human motion. In order to achieve this, VICON1, a commercial vision based motion capturing system was interfaced with various sound producing languages and frameworks.

Sonification of human motion can yield results that are not observable by vision alone. Perception of periodicity, regularity and speed of motion are few of the attributes that are easier to observe with the aid of sound.

Rather than forcing a particular sonification method, our goal is to provide maximum flexibility in the mapping of motion parameters to sound. This would enable individual users interacting with the system to choose their own personalized sonification approach.

Data from a variety of human motions including, musicians performing on traditional instruments, dancers acting out emotions, as well as data from individuals having impairments in sensor motor function, was collected and preliminary experiments in sonifying the data were conducted.

Although the proposed infrastructure can be applied to many areas of research, we are interested in studying how musician's posture and gestural movement during performance affect the sound produced as well as the emotional content of the performer. We also intend to use sound to learn how to build machine-based emotion detection, also known as affective computing. Finally, it is our hope that this work will eventually enable individuals with motor disorders to understand and modify their patterns of movement. For example, sonification can be used as a biological feedback mechanism to help in modulating specific parameters of human body motion such as movement speed.

The paper is structured as follows: related work will be presented in section 2. A brief background on the technology and software that is being used in the research will be discussed in section 3. Section 4 explains the infrastructure and framework of our system. Finally section 5 gives a variety of case studies and reports to initial experiments using the system built.

2. RELATED WORK

There have been experiments with using motion capture systems to study a musical performers interaction with their instrument. Sources [1,2] describe comprehensive studies of three main factors that influence performance: (1) The instruments constraints on the body, (2) the characteristics of the performance (e.g. rhythm, articulation, tempo, etc.) and, (3) the interpretive momentary choices of the performer. The same team did further research to analyze the production and reproducibility of the performer's ancillary body movements [3]. They also presented results for controlling digital audio effects using ancillary gesture capture data [4].

[1] http://www.vicon.com (January 2005)

Training a machine to recognize human emotion is an active field of study, which is changing human-computer interactive interfaces and applications. For more details on the evolution and future of affective computing as well as more complete lists of references readers are pointed to papers [5,6]. There are researchers sonifying machine-based emotion recognition. EmotionFace [7] is a software interface for visually displaying machine-based emotion expressed by music. There is also work in automatic emotion recognition derived from the body gestures of dancers using the EyesWeb framework presented in [8].

A review of the literature also shows that researchers are experimenting with the sonification of human body gestures for medical applications. Experiments to determine whether auditory signals can provide proprioceptive information normally obtained through muscle and joint receptors proved that sonification of joint motion though strong time/synchronization cues were a successful approach.[9] Other medical applications include researchers who are experimenting with auditory displays of hyperspectral colon tissue images using vocal synthesis models[10]. Rhytmic sonification of temporal information extracted from the scalp EEG helped in analysis of complex multivariate data with subtle correlation differences [11].

3. TECHNOLOGY BACKGROUND

This section will describe the technologies being used for this research: The Vicon Motion Capture System, Marsyas, Synthesis Toolkit (STK), and ChucK.

3.1. VICON Motion Capture System

VICON[12] is a system for motion capturing. Spheres covered with reflective tape, known as markers, are placed on visual reference points on different parts of the human body. The system enables custom models to be built to focus on specific aspects and details of motion (e.g. more markers can be placed on lower body to examine walking, while more makers can be placed on hand for examining drumming). The VICON system consists of 6 cameras and is designed to track and reconstruct these markers in 3-dimensional space. When a marker is seen by one of the cameras, it will appear in the camera's view as a series of highly illuminated pixels in comparison to the background. During capture the coordinates of all the markers in each camera's view are stored in a data-station. The VICON system then links the correct positions of each marker together to form continuous trajectories, which represent the paths that each marker has taken throughout the capture and thus how the subject has moved over time. At least three of the cameras must view a marker for the point to be captured. Therefore to obtain continuous signals interpolation is used to fill in the gaps.

3.2. Marsyas

Marsyas [13] is a software framework for rapid prototyping and experimentation with audio analysis and synthesis with specific emphasis to music signal and music information retrieval. It is based on a data-flow architecture that allows networks of processing objects to be created at run-time. A variety of feature extraction algorithms both for audio and general signals are provided. Examples include Short Time Fourier Transform, Discrete Wavelet Transform, Spectral and Temporal Centroid, Mel-Frequency Cepstral Coefficients and many others. In addition Marsyas provides integrated support for machine learning and classification using algorithms such a k-nearest neighbor, gaussian mixture models and artificial neural networks. Using Marsyas, the motion capture signals from the Vicon system can either be sonified at various playback speeds or they can be used directly to train machine learning models for classification. Marsyas is freely available from http://marsyas.sourceforge.net and contains code for reading and processing files in the Vicon file format.

3.3. Synthesis Toolkit (STK)

The Synthesis ToolKit (STK) [14] is a set of open source audio signal processing and algorithmic synthesis classes written in the C++ programming language. STK was designed to facilitate development of music synthesis and audio processing software, with an emphasis on cross-platform functionality, realtime control, ease of use, and educational example code. The Synthesis ToolKit is particularly novel because of its collection of physical models of instruments. Physical models model the time domain physics of the instrument and take advantage of the one-dimensional paths in many systems (ex. strings, narrow pipes) and replace them with delay lines (waveguides) [15].

STK provides a wealth of synthesis possibilities. However, due to its low-level nature, it can be difficult to rapidly develop and experiment with sonification algorithms. To make a modification, programmers must stop, write code (C++), compile, and restart to listen to result. Furthermore, there are no facilities for concurrent programming. In the next section, we discuss how a higher-level language might be used to flexibly address these issues.

3.4. CHUCK

ChucK [16,17] is a concurrent, strongly-timed audio programming language that can be programming on-the-fly. It is said to be *strongly-timed* because of its precise control over time. Concurrent code modules that independently process/sonify different parts of the data can be added during runtime, and precisely synchronized. These features make it straightforward to rapidly implement and experiment with sonification algorithms, and also to fine tune them on-the-fly.

Using the timing framework alone (with a single process), it is possible to "step through" the captured data with respect to time, sonifying any single marker or a group of markers - using STK (via ChucK) or other synthesis modules. The sonification rate can be dynamically increased or decreased by modifying the syntax of time advancement. Concurrency builds on this mechanism by allowing interchangeable processes to sonify different parts of the program in parallel. We provide an example of this in 4.3.

4. FRAME WORK

The framework to map motion parameters of the human body to sound is described in this section. It is a three-part process: (1) collect data using the VICON system, (2) import data into desired synthesis language, (3) choose sonification algorithm.

4.1. Data Collection

After motion capture trials are run using the VICON system, all data is labeled and interpolation algorithms are run to obtain continuous streams of marker positions. Next, each trial is exported to a text file. The first line of the text file contains the label names of the markers, delineated by a comma. Each line is time stamped and represents the x-, y-, z-, coordinates of all the markers for that particular time instance. All our experiments were captured at a 120 Hz sampling rate.

For this research, data collected falls into three different categories. The first category is data collected on performers playing traditional instruments, namely the tabla and the violin. Tabla performance recordings were of traditional *Tin Taal Theka* excerpts of 16-beat cycles. As shown in Figure 3, a full model of the right hand was captured using a custom built VICON plugin to capture 28 marker points. The violin performances were of simple songs in moderate tempo in a major scale. As shown in Figure 4, markers were placed to capture upper body movements including the head, arms, and upper clavicle.

The second category is data collected on dancers acting out different emotions. The subjects were asked to enact four basic emotions using their body movements. No specific instructions for how these emotions should be enacted were given resulting in a variety of different interpretations. The basic emotions used were sadness, joy, anger, and fear. Markers were placed at 14 reference points on the subjects bodies as seen in Figure 5, in order to get a full representation of the skeletal movement.

The third category is data collected from individuals having impairments in sensor motor coordination. Specifically, we obtained data from children with Down syndrome and children with cerebral palsy. A detailed skeleton of the body was obtained using 34 markers as seen in Figure 6. Subjects were asked to perform tasks, such as walking at variable speeds across a room. We also collected data for subjects who do not have disorders for comparison.

4.2. Importing Data

Once the data is collected, the files are imported into the desired synthesis language. For both Marsyas and ChucK, custom ViconFileSource classes were designed to read the marker sets. The motion capture data and derived features can be used to control parameters of different synthesis algorithms and digital audio effects for sonification purposes. Both languages were modular enough to allow for two streams of data, (in this case, Vicon and audio) to run at two different sampling rates.

4.3. Sonification Algorithms

Using Marsyas for audio analysis, feature extraction and classification, and STK for synthesis, and finally ChucK, a high-level language for rapid experimentation, we are able implement a breadth of sonification algorithms.

Using Marsyas, we were able to design a simple gesture based additive synthesis module, which took *n* different makers and used them to control the frequency of *n* sinusoids. The code sketch in Figure 1 shows how the 3 markers of the x,y,z wrist position can be used to control 3 sinusoidal oscillators in Marsyas. Another method was to use the gesture data to control the gain values of the *n* different wavetables. In order for this to work, each marker's data stream had to be normalized.

Another technique that was easy to implement in Marsyas was gesture-based FM synthesis. FM synthesis is a method of creating musically interesting sounds by repetitively changing the basic frequency of a source. We set up a system to have the modulation index and source frequency change with data from the marker streams.

```
01# while (viconNet->getctrl("bool/notEmpty"))
02#     {
03#          // read marker data from file
04#          viconNet->process(in,out);
05#
06#          // control frequencies of sine oscillators
07#          pnet->updctrl("real/frequency1", out(1,0));
08#          pnet->updctrl("real/frequency2", out(2,0));
09#          pnet->updctrl("real/frequency3", out(3,0));
10#          // play the sound
11#          pnet->tick()
12#     }
```

Figure 1. *The following code sketch shows has the 3 markers of the x,y,z wrist Position can be used to control 3 sinusoidal oscillators in Marsyas.*

An example of motion controlled digital audio effect implemented in Marsyas is a real-time Phase Vocoder[18]. A Phase Vocoder is an algorithm for independent control of time streatching and pitch shifting. Thus the marker data streams can control the speed of the audio playback and the pitch independently of each other.

Using STK, we were able to control physical models of instruments. This way we could use marker streams to control different parameters (such as tremlo rate, hardness, direction, vibrato, reed apeture, etc) on instruments including flute, clarinet, mandolin, shakers, sitar, etc.

```
01#     // read single column of data from input
02#     ColumnReader r( input, column );
03#     float v;
04#
05#     // time-loop
06#     while( r.more() )
07#     {
08#         // read the next value
09#         r.nextValue() => v;
10#         // do stuff with v
11#         ...
12#         // advance time as desired
13#         T::ms => now;
14#     }
```

Figure 2. *Example template ChucK code for sonification of body motion data.*

ChucK provides the ability to (1) precisely control the timing of a sonification algorithm and (2) easily factor many complex sonification algorithms and digital audio effects into concurrent modules that are clearer (and easier) to implement and reason about.

In this example (Figure 2), we show a simple template used for sonifying multi-valued streams of marker data by factoring into concurrent processes - one process for each value stream (column). The template first creates a reader for a specific column (line 2). In the loop (lines 6-14), the next value is read (line 9) and used for sonification (to control synthesis, etc., line 11). Finally, time is advanced by any user-definable amount (line 13).

This template can be instantiated one or more times as concurrent processes, each with a potentially different column number, time advancement pattern, and synthesis algorithm. Together, they sonify the entire dataset, or any subset thereof. For example, one process uses a granular model to sonify column 2, and another uses a plucked string model to sonify column 5. One of the properties of ChucK is that all such processes, while independent, are guaranteed to be synchronized with sample-precision. Furthermore, it is possible to add/remove/replace a process on-the-fly, without restarting the system.

5. CASE STUDIES

The following case studies are presented to portray how the framework presented in section 4 can be useful in different types of research.

5.1. Experiments with Musical Instruments

The goal in this area of study is to sonify events of the gestures of performers playing different instruments. There are numerous areas of interest that we hope to explore using sound.

First, we are interested in finding which markers contain musical information. This is being tested using STK's physical models of the instrument, in order to try and reproduce a performance, using the marker's data to control parameters of appropriate physical models. Our goal is to find how few markers can be used in order to reconstruct a musical phrase. Another interesting question is to observe interchanging traditional mappings (e.g. map plucking hand to bowing, and bowing hand to plucking) to obtain new types of sound.

Another area of interest is to observe ancillary gestures during performance (e.g. how the head moves during a violin performance). Specifically, we ask the following questions: When a perform plays the same composition, does the ancillary body gestures move in the same way? What is the minimum number of markers that need to be the same in order for the same performance to be played? What type of information do the ancillary markers obtain? Answering these questions using the framework we have designed allows us to observe subtle differences in movements that are difficult to see using only visual feedback.

As seen in Figure 3 and 4, we are performing initial experiments on a tabla performance and violin performance. The challenge with the tabla is the precise timing of fingers. Thus we use a detailed model of the hand, as described above, in order to preserve the performance. Challenges with the violin include timing like the tabla, but also another dimension of pitch.

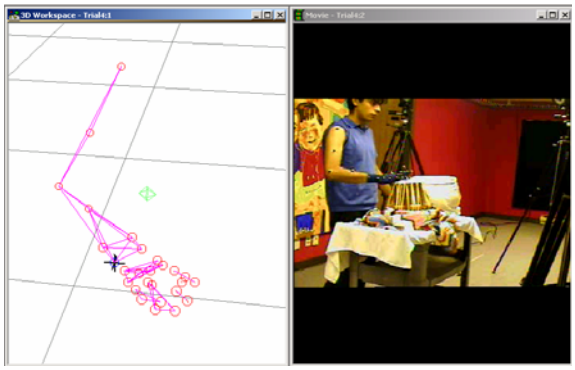

Figure 3. *Screenshot of data capturing process for tabla performance.*

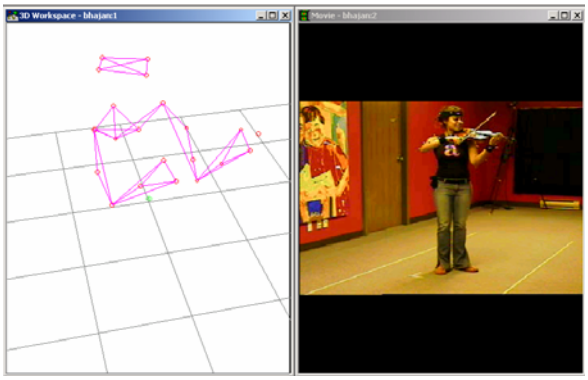

Figure 4. *Screenshot of data capturing process for violin performance.*

5.2. Experiments with Affective Computing

The goal in this area of study is to use our sonification infrastructure to help aid in analyzing human emotion. Using data collected (described in section 4.1) our first experiments have been to simply use the position data to control frequencies of the sinusoids in an additive synthesis representation of emotion. Using all the markers at the same time ended up distorting sound and no information can be found. We mapped data from the wrist movements of the subject to control a small set of sinusoids. With this method, it was easy to discern happy from sad, but angry and scared did not display in changes in sounds.

Using feature extraction algorithms in Marsyas, we began to derive data such as velocity and acceleration of the markers and use that information for sonification. These initial results are more promising for discerning the four emotions from sound. We are currently exploring other features such as centroid, flux and root-mean-square for sonification. Our goal is to find what is the critical feature that distinguishes the different emotions. What is the maximum number of markers that hold the data, and where are they located? One really interesting aspect of these experiments is the fact that different emotions can still be distinguished in the abstract mapping of motion capture data to sound.

Another approach described in detail in [19], is to train a classifier using the marker data and machine learning algorithms such as k-nearest neighbor, artificial neural network, and Gaussian classifiers. Then give the system marker data (not used for training), have it calculate a prediction class, and sonify the result. In this case, each emotion would have a different melody, or soundfile and be triggered by the output of the classifier. This is a useful approach implemented in Marsyas that is easily adaptable to other areas of research.

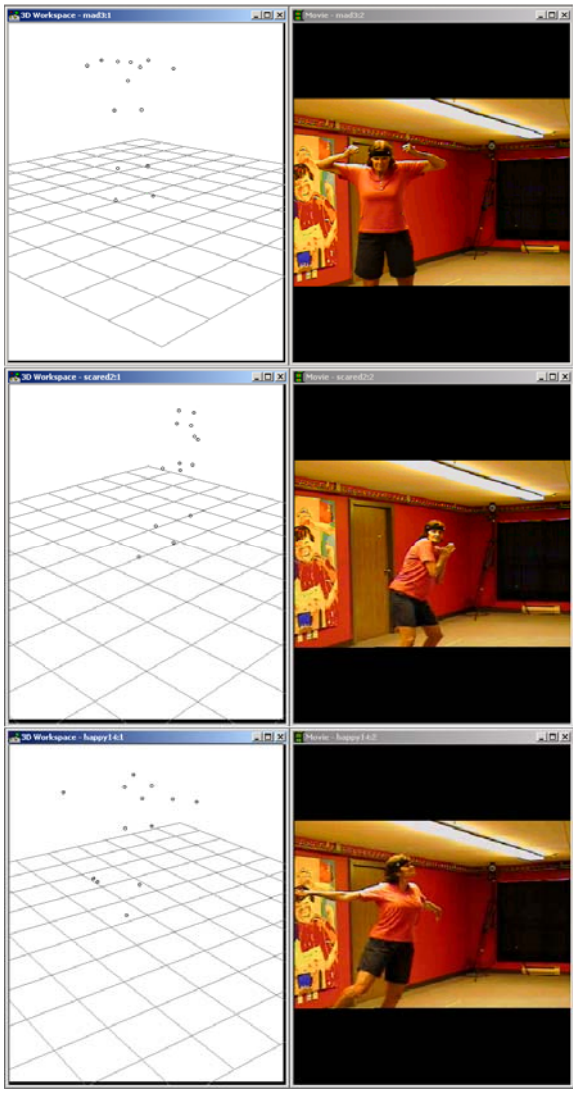

Figure 5. *Screenshots of data capturing process of actress acting angry (top), scared (middle), and happy (bottom).*

5.3. Auditory Input and Motor Control

It is well documented that individuals with Down syndrome have specific impairments in sensory-motor processing and coordination. These impairments are thought to result from a combination of changes in the functional organization of the brain as well as delays in sensory conduction. Recently it has been shown that adults with DS display better motor coordination in response to auditory information in comparison with visual information [20]. It has been suggested that auditory information may be more effective in facilitating continuous motor performance tasks [21].

Previous work in this area has focused primarily on providing external auditory feedback (using tones from a metronome) to control the timing of the movement. [22] We are proposing to sonify dynamic information about the moving limb(s) to determine if patients can use this information to improve movement speed, accuracy and motor coordination. Since auditory information is processed much faster than other sensory information (e.g. vision) we predict that following training, patients will be able to learn to use auditory information related to their current joint position and velocity to make changes to their subsequent movements. Auditory information in conjunction with visual information may enhance both feedback and feedforward processes for conveying rhythm and periodicity of limb movement.

Preliminary experiments with sonifying the data revealed that one of the most salient features is the representation of movement speed. Therefore, we mapped the motion of specific joints to frequencies of sine waves. Since individuals with Down Syndrome generally have longer movement times than neurologically normal individuals, a key question is whether these individuals can use targeted auditory information to modulate movement speed. As a first step, we sonify the movement of key joints in neurologically normal (NN) individuals during rhythmic movements (i.e. walking, jumping, running) and "replay" the motion of their joints at different speeds. Following a period of learning, we will determine whether NN individuals are able to correctly match movement speed with the sonified data. We hope that these experiments will be critical for determining the auditory-motor transformations required to modulate movement speed.

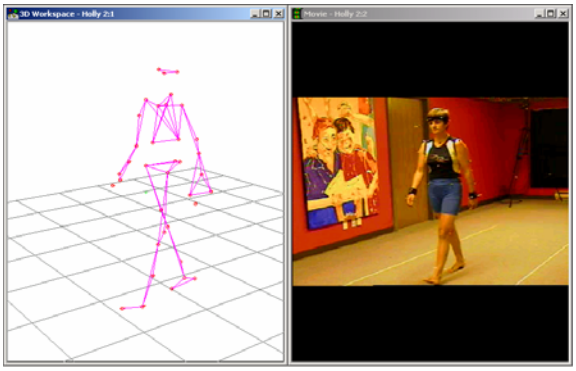

Figure 6. *Screenshots of data capturing process of a subject walking across a room.*

6. CONCLUSIONS

We have presented a framework for sonifying human body motion using a variety of synthesis languages and algorithms. Body gesture data is captured using the commercially available VICON motion capture system and exported to Marsyas, STK, and ChucK to aid in realizing the data. We have presented initial experiments in different areas of research including data captured from musicians performing traditional instruments, dancers acting out emotions, and individuals with motor control disorders. Our initial experiments show that sound aids in realization of the data captured.

There are many directions for future work. One of the currently limitation of the system is that the data capture and analysis from VICON is not obtained in real-time. We have plans to acquire a real-time VICON motion capture system. We hope to use our infrastructure to further learn how a musician's posture and ancillary gestural movement during performance affect the sound produced as well as the emotional content of the performer. We also plan use sound to aid in building machine-based emotion detection interface. Our hope is also to eventually aid individuals with motor disorders to understand and modify their movement patterns.

7. ACKNOWLEDGEMENTS

We would like to thanks Asha Kapur for helping collect the emotion data, and for performing the violin. We would also like to thank all the dancers/subjects who participated in emotion capture, including Jane Henderson, Alyson Ryder, and Lilah Montague. Thanks to the Down Syndrome Research Foundation and Queen Alexandra Center for Children's Health for use of their space and for providing support. Special thanks to patients who participated in our tests. Also thanks to Peter Driessen and Andrew Schloss for their support.

8. REFERENCES

[1] M. M. Wanderley, "Non-Obvious Performer Gestures in Instrumental Music." In A. Braffort, R. Gherbi, S. Gibet, J. Richardson, and D. Teil (eds.) Gesture-Based Communication in Human-Computer Interaction. Springer-Verlag, pp. 37 – 48, 1999.

[2] M. M. Wanderley, "Quantitative Analysis of Non-obvious Performer Gestures." In I. Wachsmuth and T. Sowa (eds.) Gesture and Sign Language in Human-Computer Interaction. Springer Verlag, pp. 241-253, 2002.

[3] M. Peinado, B. Herbelin, M. Wanderley, B.L. Callennec, R. B., D. Thalmann, and D. Méziat1, "Towards Configuarable Motion Capture with Polorized Inverse Kinematics" Sensor, 2004.

[4] M.M. Wanderley, and P. Depalle, "Gesturally-Controlled Digital Audio Effects," *In Proceedings of Conference on Digital Audio Effects.* Limerick, Ireland, December 2001.

[5] R. W. Picard, "Towards Computers that Recognize and Respond to User Emotions." *IBM System Journal*, vol 39, pp.705-719, 2001.

[6] M. Pantic, "Toward an Affect-Sensitve Multimodal Human-Computer Interaction," *Proc of the IEEE.* vol. 91, no. 9, September 2003.

[7] E. Schubert, "EmotionFace: Prototype Facial Expression Display of Emotion in Music," in *Proc. Int. Conf. On Auditory Displays (ICAD)*, Sydney, Australia, July 2004.

[8] A. Camurri, P. Coletta, M. Ricchetti, R. Trocca, K Suzuki, and G. Volpe, "EyesWeb – Toward Gesture and Affect Recognition in Interactive Dance and Music Systems," *Computer Music Journal,* 24:1, pp 57-69, MIT Press, Spring 2000.

[9] C. Ghez, T. Rikakis, R.L. DuBois and P.R. Cook, "An Auditory Display System for Aiding Interjoint Coordination," in *Proc. Int. Conf. On Auditory Displays (ICAD)*, Atlanta, Georgia, April 2000.

[10] R. J. Cassidy, J. Berger, K. Lee, M. Maggioni, and R.R. Coifman, "Auditory Display of Hyperspectral Colon Tissue Images using Vocal Synthesis Models," in *Proc. Int. Conf. On Auditory Displays (ICAD)*, Sydney, Australia, July 2004.

[11] G. Baier and T. Hermann, "The Sonification of Rhythms in Human Electroencephalogram." in *Proc. Int. Conf. On Auditory Displays (ICAD)*, Sydney, Australia, July 2004.

[12] A. Woolard, *Vicon 512 User Manual*, Vicon Motion Systems, Tustin CA, January 1999.

[13] G. Tzanetakis, and P. R. Cook, "MARSYAS: A Framework for Audio Analysis," Organized Sound, Cambridge University Press, vol. 4, no. 3, 2000.

[14] P.R. Cook, and G. Scavone. "The Synthesis ToolKit (STK)," *In Proceedings of the International Computer Music Conference (ICMC)*. Beijing, China, 1999.

[15] P. R. Cook, *Real Sound Synthesis for Interactive Applications*, A.K Peters Press, 2002.

[16] G. Wang, and P. R. Cook. "ChucK: A Concurrent, On-the-fly Audio Programming Language," *In Proceedings of the International Computer Music Conference (ICMC)*, Singapore, September 2003.

[17] G. Wang and P. R. Cook. "On-the-fly Programming: Using Code as an Expressive Musical Instrument." *In Proceedings of the International Conference on New Interfaces for Musical Expression,* Hamamatsu, Japan, 2004.

[18] J. L. Flanagan, R. M. Golden, "Phase Vocoder," Bell System Technical Journal, November 1966, 1493-1509.

[19] A. Kapur, A. Kapur, N.Virji-Babul, G. Tzanetakis, and P.F. Driessen, "Gesture-Based Affective Computing on Motion Capture Data," submitted for publication, 2005.

[20] R. Chua, D.J. Weeks, & D. Elliott. (1996). A functional systems approach to understanding verbal-motor integration in individuals with Down syndrome.Downs Syndrome:Research and Practice,4 ,25 –36.

[21] S. D. Robertson, V. Gemmert, W. A. Arend, and K. V. Maraj, Brian; "Auditory Information is Beneficial for Adults with Down Syndrome in a Continuous Bimanual Task," Acta Psychologica, Vol 110(2-3), Jun 2002.

[22] M.H. Thaut, G.P. Kenyon, M.L. Schauer, and G.C. McIntosh, "The Connection between Rhytmicity and Brain Function," IEEE Engineering in Medicine and Biology, pp 101-108. April 1999.