Problems and prospects for intimate and satisfying sensor-based control of computer sound

Matthew Wright
Center for New Music and Audio Technologies (CNMAT), UC Berkeley
1750 Arch Street, Berkeley, CA 94709 USA
(1) 510 643-9990
matt@cnmat.berkeley.edu

ABSTRACT
Electronic musical instruments consist of some kind of sensor/input system, software to interpret the sensed data, and sound synthesis. Electronic instruments should provide intimate and satisfying control over sound comparable to the control that an expert musician has over a fine acoustic instrument. These elements are necessary for such electronic instruments: low latency and virtually no jitter, precise and accurate sensing, reproducibility and reliability. Sensor-based electronic musical instruments should also embody parsimony and transparency.

General Terms
Measurement, Design, Experimentation, Human Factors.

Keywords
Accuracy, Jitter, Latency, Transparency, Parsimony, Reproducibility

1. INTRODUCTION:
The field of computer music has been a fertile ground for research and development of systems incorporating a wide variety of sensing and input devices.

Sound synthesis requires control, both for traditional musical parameters such as pitch, loudness, and duration, and for parameters unique to particular synthesis structures, e.g., carrier/modulator ratio for FM, virtual breath pressure for physical models of wind instruments, and position in a timbre space [10] for timbre-interpolating synthesizers. One approach to generating control information is to do so out of real time in a procedure analogous to a composer sitting down with staff paper and a pencil. The success of the orchestra/score model of the Music-N languages [7] attests to the effectiveness of this approach.

My interest is in systems along the lines of Figure 1. In this model, the sound synthesis control parameters are derived in real time based on the data sensed by one or more sensors. The mapping system may be as simple as "map the X position measurement to logarithmically-scaled frequency and scale it so that the 12 inches of input dynamic range correspond to the 8 octaves from 40 Hertz to 10240 Hertz." The mapping system can also be very complex, potentially involving, e.g., statefulness, switching among different modes, stochastic processes, or high-level control of algorithmic processes to generate musical material. The mapping system may also involve the triggering and transformation of stored musical material. (Since the stored musical material is often prepared out of real time before the performance, the supposed dichotomy between nonrealtime "composing" and realtime "performing" is not really so clear.)

Joel Chadabe argues eloquently that the concept of "mapping" may be too limited to describe complex algorithmic and/or nondeterministic systems between the sensor measurements and the synthesis control [4]. Nevertheless, in my mind if the input to a subsystem is sensor measurements and the output is parameters to control sound synthesis, then that subsystem, no matter how complex or indirect, is doing the job of mapping.

1.1 Why Sound?
Since this conference is aimed at sensing and input for multimedia systems in general, not just musical instruments, I will give some reasons that sensor-controlled audio systems should be of interest to anyone working with sensors and input systems.

One reason is that human hearing has much higher temporal precision than vision, touch, smell, or taste [8], so the challenges (and rewards) of building satisfyingly responsive real-time systems are greatest when the output is sound.

Another reason is that systems along the lines of Figure 1

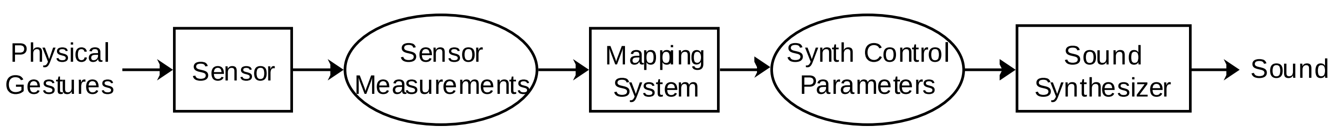

Figure 1. Structure of a sensor-based system for control of sound synthesis

Sensors and Input for Multimedia Systems (SIMS) '02, June 20-21, 2002, Santa Barbara, CA. Copyright 2002.

naturally invite comparison with traditional acoustic instruments. Of course the point of the comparison is not to attempt to decide whether acoustic or electronic instruments are "better"; each have their strengths and weaknesses. My point is that the finest acoustic instruments set standards for nuance, intimate fine control, dynamic range, temporal responsiveness, and the possibility for a performer to develop virtuosity, and that electronic instruments have definitely not surpassed acoustic instruments in all of these areas. The challenge to meet or come near these standards with our electronic instruments is a powerful force driving development, and this development will naturally include some solutions that can be generalized to non-musical uses of sensors.

Traditional acoustic instruments also provide tried and true models for the relationship of physical gestures to the resulting sounds. This results in a large supply of research projects of the form "simulate behavior X of instrument Y with an electronic instrument." I believe that the emulation of acoustic instruments is an interesting goal only to the extent that it increases our general knowledge of sensor-based instruments or results in hybrid instruments that combine the strengths of electronic sound production with the features of acoustic instruments.

1.2 Who I am

Since this is a position paper, I will say a few words about my background as a musician and as a technologist to make it more clear where my biases and aesthetic tendencies come from.

From age 8 I was programming in BASIC on early personal computers, typically very limited interactive programs based on text or later on Apple-II graphics. At about age 14 I lost interest because I had run out of things I knew how to do and I had no tools for structuring more ambitious projects.

I was trained in western classical music from a fairly early age, but was always too lazy to practice enough to attain more than a basic proficiency.

In high school in the 1980s I became interested in synthesizers, drum machines, electric guitars (and electric guitar effects processing), sequencers, and 4-track cassette recorders. In those days I spent a lot of my free time making music with these technologies along the lines of the popular music I was listening to. This included both playing with friends in bands and working alone or with friends in our various little home studios. During this time I got a lot of experience using electronic music technology to make up for deficient musicianship.

I also played piano in my high school's "jazz workshop," which gave me experience with a certain kind of improvisation as well as exposure to a lot of (recorded) music played very intimately by master musicians who had developed very personal relationships with their instruments.

As a university student I was finally introduced to the ideas of computer science [1], and my interest in writing computer programs as a tool to express ideas was rekindled. I took David Wessel's computer class and was infected with his love of interactive live performance, his way off looking at music-making from the point of view of the scientific study of music perception and cognition, and his openness to music from around the world.

Since about 1995 I have been extremely interested in Indian, Afghan, Persian, Kurdish, Arabic, Turkish, Armenian, and Andaluse music, and today I spend almost all of my non-computer-music time studying, practicing, observing, or performing these kinds of music. Northern California is blessed with a wealth of master musicians from these parts of the world. To make a few generalizations: all of this music is based heavily on melody and rhythm; even the strictest classical traditions have retained elements of improvisation (always within certain frameworks); and for the most part the ideal is a fairly small ensemble of virtuosi.

Where does all this leave me? I love rhythm, and it takes a lot for slow timbral evolution or other kinds of non-rhythmic music to sustain my attention. On the other hand I'm tired of drum-machine-perfect mechanical rhythms; my interest is in the "feel" and "groove" of expert performers. (For example, one of my current obsessions is the difference in microtiming between Indian and Afghan 7-beat rhythms.)

I'm also drawn towards the model of the improvising virtuoso instrumentalist. Many of my profoundest musical moments have been watching a master musician create music in real time.

2. TERMINOLOGY

Referring again to Figure 1, I will define some of the terminology I use to discuss these kinds of systems.

A "sensor" is any device that measures properties of the physical world and translates them to data that can be interpreted by the rest of the system.

I will use the word "instrument" to refer to the entire system from the sensors to the sound output. In my mind, for example, changing the mapping from sensor measurements to synthesis control parameters results in a new instrument.

The "interface" that an instrument presents to the performer is a combination of the physical interface, i.e., the sensors, with what might be called the "semantic interface," i.e., everything a performer knows about what sound will result from which gestures.

3. ELEMENTS OF SATISFYING AND INTIMATE CONTROL WITH SENSORS

In this section I will list the factors that I believe determine whether a sensor-controlled music system will be intimate, satisfying, and expressive, and whether it will be possible for a performer to develop virtuosity with it.

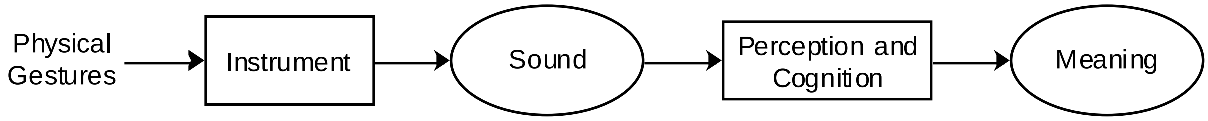

Figure 2. The role of music perception and cognition

3.1 Listening to the Instrument

Ultimately, the purpose of an instrument is to make a sound that people will hear, e.g., the performer(s) and the audience. So the question is not "what sound does the instrument make?" but "how do we perceive the sound that the instrument makes?" Figure 2 shows how we might think of perception and cognition as giving meaning to the sounds we hear.

With instruments not matched to human perception, much of the range of possible output sounds may be "wasted" on possibilities that make no difference to listeners. An extreme example would be a frequency control that went from 30Khz to 40KHz; in this case all of the possible outputs would be out of range of human hearing and this control would have no perceivable effect whatsoever. A more realistic example would be a volume fader that controlled linear rather than logarithmic amplitude; near the top of the range small changes in the fader position would have very little effect, while near the bottom of the range small changes in the fader would have such a large perceptual effect that it would be difficult to exercise fine control.

The effect of playing the instrument must be perceivable. Part of the art of designing instruments' mapping systems is in mapping the range of input gestures (according to the limits of the sensors) to a perceptually meaningful range of sound output.

3.2 Time

In my opinion, timing is the most important aspect of music, and a musician's mastery of timing is the most important aspect of her skill. Skilled instrumentalists time their gestures exactly so as to produce exactly-timed musical results. We can perceive very slight deviations in timing, especially against a somewhat steady pulse. In rhythmic music, these small variations convey musical information and can make the difference between an expert performance and a mediocre performance.

Therefore we must consider the latency and jitter inherent in our sensor-based instruments. In this context, an instrument's "latency" is the total elapsed time between a physical gesture and that gesture's effect on the sound output. "Jitter" is simply the variation of latency.

3.2.1 Latency

Low latency is critical to the feeling of control intimacy. The rule of thumb is that sound travels approximately one foot per millisecond in air, so a 10 millisecond (ms) latency is equivalent to playing through a loudspeaker 10 feet away. This figure of 10ms seems to be a good goal for latency; for example, electric guitarists often place their loudspeakers 10 feet away from their ears, but rarely 20 feet away.

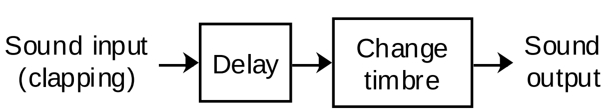

Figure 3. Test system to determine audibility of latency

Consider an experiment using the system shown in Figure 3. The subject makes a percussive sound, for example, a clap, and listens to the output of the system on headphones. The "change sound" element ensures that the output will be timbrally distinct from the input. (I used a bank of resonant filters.) The delay time is set randomly on each iteration of the experiment. How big can the delay time be before the subject can notice the delay? With myself as a subject the figure is about 10ms. I encourage electronic instrument builders to try this experiment on themselves to get a feel for the effects of various system latencies. This system also allows one to compare the relative audibility of the delay depending on the type of sound; in general we are most sensitive to the timing of percussive transients.

3.2.1.1 The sounds of the sensors themselves

Although the functional purpose of the sensors is only to measure physical gestures, many sensors make a sound as they are used. Examples include the click when a button is pressed, the thud of a drumstick hitting a rubber trigger pad, and the clicking of a pianist's fingernails on the keys. Obviously, from a mechanical engineering standpoint, we should attempt to dampen these kinds of sounds as much as possible; that's why concert pianists trim their fingernails.

It's tempting to say "the synthetic sound coming out of my loudspeaker is loud enough to mask the sound my sensor makes, so I won't worry about it." Unfortunately, our perception in these situations is dominated by the precedence effect, also known as the "law of the first wavefront" [2]. This effect tells us that when we are hearing a sound from two locations, and the wavefronts arrive at our ears more than about a millisecond apart, the one that we hear first will determine our perception of where the sound is coming from, even if the later sound is drastically louder.

So it's not enough to simply drown out the sound of the sensors; the synthetic sound must reach the audience before the sound of the sensor. Since our electronic instruments have latency, the only way to defeat the precedence effect is to take advantage of the slowness of sound by placing the loudspeaker between the sensors and the audience.

3.2.1.2 Gestures that come before a note sounds

On many instruments the performer must prepare in advance before actually making each sound. Percussionists must begin moving their hand or stick in advance so that the strike comes at the right time. String players must stop the strings at correct length before plucking or bowing. Singers must be ready with enough breath and with certain vocal tract musculature in the proper state. In more extreme examples such as the pipe organ and alpenhorn, there is a substantial delay between the gesture that plays a note and the actual sounding of that note. Each of these examples can be viewed as performer timing her gestures to account for latencies inherent in the instrument.

I believe that the sensing of these kinds of "pre-note" preparatory gestures will result in more intimately controllable sensor-based instruments. I know of one example: Don Buchla's *Marimba Lumina* [3]. This is a commercially available instrument modeled somewhat after the acoustic marimba; the performer uses special mallets to strike the virtual bars on the surface of the instrument. Unlike most percussion sensors, the Marimba Lumina continuously senses the distance of each mallet to the instrument surface. Computing the velocity allows the Marimba Lumina to predict when the mallet will strike; the resulting "advance notice" compensates for latencies in the rest of the system, resulting in an intimate instrument usable for precise rhythmic material.

3.2.2 Jitter

The above examples of performers compensating for instruments' latencies apply to fixed, deterministic latencies. For example, if there is a 100ms delay between depressing a pipe organ key and the resulting sound, that delay will always be 100ms; this consistency is what allows the performer to learn to compensate. Unpredictable latency, however, guarantees that the performer will not be in control of the rhythm of the output sounds.

Figure 4 shows a test system similar to that shown in Figure 3, but for jitter instead of latency. It is difficult to produce a variable delay for an audio signal without pitch shifting or complex granular techniques, so this system is based on an event representation such as MIDI output by a keyboard. Synthesized sounds with sharp attacks are most illustrative. The question is: how large of a range of latency can the system impose without adversely affecting the rhythm that the performer plays? Again I encourage builders of sensor-based instruments to experiment for themselves with the tolerability of different amounts of jitter.

The musically acceptable amount of jitter depends greatly on the kinds of sounds being produced. At one extreme are examples such as dense textures, timbral changes in sustaining sounds, and so forth, where latency and jitter up to even 200ms may have very little impact in the musical result.

In the middle is the common case of each gesture causing an individual note to be played. For a skilled performer, jitter of around 10 ms makes the difference between feeling that she does or does not have complete control of rhythms played on the instrument.

The other extreme is illustrated by the case of "flams," i.e., pairs of notes played in very rapid succession on the same instrument. Skilled drummers can vary the timbre produced by a flam by adjusting the time between the two notes, for example, by controlling the relative height of two drumsticks before setting them in motion towards a drum head. Timing differences between the two notes result in differences in the timbre of the flam, and even a single milliseconds' difference can be audible [5, 9]. Therefore a sensor-based electronic instrument would have to have jitter no more than the 1ms to allow for this playing style.

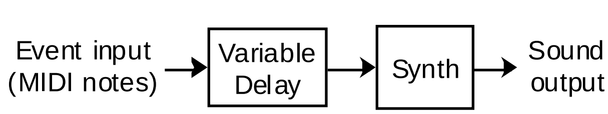

Figure 4. Test system to determine audibility of jitter

3.2.3 A System's Cumulative Latency and Jitter

Generally, a system's overall latency and jitter is at least as bad as the sum of the latencies and jitters of each component.

For isochronously sampled input sensors, the sampling rate determines the jitter. Suppose a sensor outputs its current measurements every 5 milliseconds; in this case, gestures that happen to come just before the sampling time will have near-zero latency, while gestures that come just after the sampling time will have 5ms latency. Thus, in general, 200Hz sampling results in 1 / 200Hz = 5ms of jitter.

The communications protocols used to get the sensed data into the mapping system have their own latencies and jitters. All communications networks have latency and jitter. Serial protocols such as MIDI add additional jitter in the case where a new message needs to be sent but a previous message is currently being transmitted.

If the system includes a computer then there will be more latency and jitter having to do with interrupt processing for I/O, e.g.., USB and Ethernet interrupts. Audio I/O for computers is always in "blocks" or "vectors" of multiple samples, sometimes 64 samples or lower but often 512 or 1024; this introduces latency but no jitter.

3.2.4 Trading Latency for Jitter

I know of two methods that remove jitter by adding latency.

To remove jitter introduced by network transmission, each message can include a time tag saying when it is supposed to take effect. The receiving device then implements a delay equal to the longest expected time for network transmission. Messages that happen to arrive early can be held until the correct time, while messages that happen to arrive late can take effect as soon as they are received. In effect this delays all messages by the expected worst-case delay time. The OpenSound Control protocol [11] includes this mechanism.

The other method is a form of temporal quantization. A process with very little jitter, e.g., CNMAT's Rhythm Engine [6], continuously schedules musical events. We assume that the input from the performer will arrive at this process with some amount of jitter, so instead of using the performer's gestures to schedule events directly, we use it to control the process that is scheduling the events, e.g., to bring different streams of material into or out of the texture, or to change a parameter such as tempo.

3.2.5 Temporal Limits of Internet Musical Collaboration

There has been a lot of interest in musical collaboration over the Internet and attempts to build systems that allow musicians to play together over long distances.

Latency and jitter are the main problems for collaborative Internet music-making. Even ideal network communications are limited by the speed of light, which is 299.792458 kilometers per millisecond. Light travels the 446 kilometers from Santa Barbara to Berkeley in a mere 1.5ms, a short enough latency to allow intimate music-making, but to get the 8782 kilometers from Berkeley to Amsterdam takes 29ms. In practice, network packets generally take more like 150 ms to get from Berkeley to Europe.

Successful Internet music making over long distances always involves a compositional solution to the latency and jitter problem. One kind of solution is to make music that doesn't depend on precise timing, for example, gradual timbral evolution of dense ambient textures. Another class of solution synchronizes incoming (and therefore delayed) musical material with some kind of precisely-timed musical process, e.g., the repetition of a loop. For example, each remote collaborator could be listening to a 10-second loop and occasionally adding or removing musical material. Even though the performer in Amsterdam receives new material after it has been played in Berkeley, the system can wait until the next time the loop repeats and place the new sound from Berkeley at exactly the right time in the loop.

3.3 Accuracy and Precision

More accurate and precise sensors give rise to better instruments. It should be possible for small gestures to have subtle effects on the output sound, while still allowing large gestures. Continuous control of pitch requires a resolution of about 2 cents over a range of at least 2 octaves.

Consider baseball pitchers as an extreme example of the capabilities of the human body. Today's fastest pitchers achieve speeds of 105-115 miles per hour; no doubt that is about the fastest that humans can possibly move their hands.

3.4 Parsimony

I use "parsimony" not in the sense of "unwilling to spend money," but in the Ockham's Razor sense of "prefer the simplest tool that does the job." In the case of sensor-based music systems, the "job" is allowing the performer to achieve the desired sonic result by playing the instrument. We must always ask ourselves "Is this the simplest interface that could do the job?"

As a baseline reality check, I like to compare the sensors used in an instrument with a bank of faders and a MIDI keyboard. What can a performer do with instrument X that she couldn't do with some faders and switches?

As an example, I once played with a set of four light-interruption sensors with four parallel thin beams of light stretching in front of my face. The instrument lit up my hand in a beautiful way when I played it, and I could make all kinds of theatrical motions with my arms and hands as I played it. However, there was not much I could play on the instrument that generated sensor measurements that I couldn't have made with a 1/3 octave non-velocity-sensitive keyboard. The one exception I found was a rapid sort of tremolo effect with the fingers each interrupting each light in a staggered fashion. This allows the playing of certain musical figures much faster than would be possible with 4 discrete keys, because it is not so easy to make each finger strike and release each key in turn.

By the principle of parsimony, if I were going to perform with these sensors (for a (paying) audience), I would feel obliged to do something with them that I couldn't do without them. As something of a musical purist I would probably attempt to take advantage of the tremolo technique. Other options would include crafting an overall visual spectacle around the aesthetics of the lights or combining the lights with acoustic instruments and/or other sensors to produce a more complex instrument.

3.4.1 Economy of Motion

Economy of motion is a universal attribute of skilled acoustic instrument performance. The concept of economy of motion is relatively well-defined for acoustic instruments because the physics of sound production demand certain kinds of gestures, so any other motion beyond the bare minimum needed to get the air vibrating can be easily classified as "extra".

Economy of motion is not the only possible aesthetic, of course, and many of the most entertaining musicians perform gestures whose effect is mainly visual, e.g., the guitarist's "windmill," rotating the right shoulder and arm through a wide circle on the way to striking the strings. More subtle body motions such as rocking back and forth can be an effective device for controlling timing.

For sensor-based instruments, our ability to choose a mapping function can always give us, in a certain sense, unlimited economy of motion, because the tiniest, most barely-detectable gesture can cause an entire musical work to come out the loudspeakers. But pressing "play" on a tape recorder is a pretty weak form of performance.[1] For me a true performance must include constant decision-making (conscious or not) by the performer and the constant possibility for the performer to control or at least influence the resulting sound in some way. So I build instruments on the assumption that a performance will consist of one or many gestures over time. The point is that we designers of sensor-based instrument get to decide what range of motions the instrument will require.

So the first question is "what musical elements will the player of the instrument affect?" This determines the purpose of the instrument, and from this "requirements description" there is often a fairly straightforward engineering tradeoff between accuracy and economy of motion. For example, mixing boards have faders with a "throw" (i.e., usable range) in the 80mm to 120mm range. Purely on the basis of economy of motion one might prefer 10mm faders, but given the anatomical limits of the human hand and the desired dynamic range of gain control, these would not have enough accuracy. So "economy" doesn't mean to avoid as much motion as possible, but to avoid *unnecessary* motion.

Given the musical purpose of the instrument and the necessary accuracy, designing for economy of motion involves questioning the ergonomics of the sensors and the user interface design of the mapping system.

Another view of economy of motion has to do with the congruence of the "size" of a gesture with "size" of gesture's sonic result. "Big" gestures should make "big" changes on the sound output.

3.5 Transparency

The popularity of laptop performers has given rise to a certain performance practice and aesthetic of inscrutability. The performer sits staring at a laptop, typing and whatnot, with perhaps a head nodding to the beat. The audience sees this but has no idea what relationship these gestures have to the sound they are hearing. This phenomenon is not tied to the laptop interface; it is far too easy to devise mapping strategies that obscure what the performer is actually doing to affect the sound. Often instruments based on "experimental" or "alternative" sensors are played with a certain amount of drama, sometimes incorporating gestures whose effect is purely theatrical and actually has no effect on the sound.

[1] However, if the performer is not actually affecting the musical output, it would be more parsimonious to admit that the piece is for tape and dancer!

I have no desire to judge these ways of performing. My point is simply that in the space of possible gesture/sound mapping systems, the "transparent" ones[2] are few and far between. The tendency towards non-transparency is just a fact of life in our field. I invite all performers to embrace this and make their performances as mysterious as they desire.

Nevertheless, my personal aesthetic tends towards the virtuoso performer, and I like to be able to watch a musician play an instrument. A further challenge to designing transparent mappings is that many of the easy-to-conceive ones give rise to instruments that are too limited or unsubtle. In my work I look for mappings that support the performer's complexity and nuance while at the same time allowing an audience to understand the skills being displayed in the performance. I encourage other designers of sensor-based instruments to do the same.

3.6 Reliability and Reproducibility

Learning to play an instrument requires that the instrument continue to respond in the same way to the same gestures. Therefore our instruments must reliably give the same behavior over time.

Complex real time sensor-based electronic music systems have a staggering range of possible ways to fail, and it is depressing how much time the performers of these instruments spend troubleshooting them.

4. ACKNOWLEDGMENTS

First I would like to acknowledge and thank Stephen Pope for inviting and encouraging this paper.

David Wessel has been my mentor and later collaborator for my entire history of making music with computers, and this paper is full of ideas that I either got from him or refined in the course of discussions with him. Likewise Adrian Freed has been a teacher and coworker during my entire time in this field. Most of my recent work in designing instruments has been in the context of collaborations with composer / guitarist / improviser John Schott, whose musical ideas and technological dreams have guided me greatly. All three of these people also gave me specific feedback on this paper.

Richard Andrews, John Cooper, Juliet Lee, Keith McMillen, and Ali Momeni all gave me valuable feedback on early drafts.

[2] By "transparent mapping" I mean a mapping in which it is obvious to audience members how the performer's gestures affect the resulting sound.

5. REFERENCES

[1] Abelson, H., Sussman, G.J. and Sussman, J. *Structure and Interpretation of Computer Programs*. MIT Press, Cambridge, Massachusetts, 1985.

[2] Blauert, J. *Spatial hearing : the psychophysics of human sound localization*. MIT Press, Cambridge, 1997.

[3] Buchla, D. Marimba Lumina, 2001.

[4] Chadabe, J., The Limitations of Mapping as a Structural Descriptive in Electronic Instruments. in *International Conference on New Interfaces for Musical Expression (NIME)*, (Dublin, Ireland, 2002), Keynote-2.

[5] Henning, G.B. and Gaskell, H. Monaural phase sensitivity measured with Ronken's paradigm. *J. Acoust. Soc. Am.*, *70*. 1669-1673.

[6] Iyer, V., Bilmes, J., Wessel, D. and Wright, M., A Novel Representation for Rhythmic Structure. in *Proceedings of the 23rd International Computer Music Conference*, (Thessaloniki, Hellas, 1997), International Computer Music Association, 97-100.

[7] Mathews, M. *The Technology of Computer Music.* MIT Press, Cambridge, Massachusetts, 1969.

[8] Moore, B.C.J. *An Introduction to the Psychology of Hearing.* Academic Press, London, 1997.

[9] Ronken, D. Monaural detection of a phase difference between clicks. *J. Acoust. Soc. Am.*, *47*. 1091-1099.

[10] Wessel, D.L. Timbre space as a musical control structure. *Computer Music Journal*, *3* (2). 45-52.

[11] Wright, M. and Freed, A., Open Sound Control: A New Protocol for Communicating with Sound Synthesizers. in *International Computer Music Conference*, (Thessaloniki, Hellas, 1997), International Computer Music Association, 101-104.