bterry@burnettgroup.com

Brett Terry, Burnett Group (.us)

SOUND/ACTION PARADIGMS IN MULTIMEDIA

Summary
A basic set of terms and models are developed to describe a range of possible sound-action models. Mention is made of how existing technologies could serve to implement these models.

This paper present some ideas for thinking about paradigms of sound in multimedia. I will begin by defining a few of the terms that I am going to use so their meanings won't be equivocal...

I like to differentiate between local and networked multimedia. For example, CD-ROMs and kiosks are local multimedia, meaning they run on a single computer, in contrast to the world wide web, which comprises distributed information, and is subject to a more stringent set of constraints in how it can look and sound Networked multimedia is typically characterized by a client/server model - the viewer is usually in the role of a client making requests for information from servers connected to the network. Although streaming data technologies are beginning to emulate what is possible with local multimedia, the predominant experience is still based on a transactional model.

I will also be using the term agent, a term borrowed from artificial intelligence, in order to represent some kind of software program that is trying to act within the confines of a some limited form of intelligence, by which I mean it, at a minimum, has some basic mnemonic and decision-making abilities.

I will treat the agent as a black box, the innards of which will be left deliberately vague, although I will make some suggestions about how implementations might be made using existing technology.

I will also be differentiating between primary/secondary, and simple/complex, in the context of controllable aural parameters. I take a primary sonic or aural parameter to mean a parameter that controls something about a sound directly--its pitch, amplitude, etc...--and a secondary parameter to mean a parameter that affects an abstraction or representation of a sound such as style or density. By a simple sound-action, I mean changing one parameter at a time versus a complex sound-action that changes more than one parameter simultaneously.

This paper is really about interactive sound/action paradigms, and concentrates mostly on networked multimedia, where the challenges are more significant, at present, than in local multimedia.

There will always be many uses for non-interactive sound paradigms in multimedia. Using a background soundtrack, delivering information by voice are all effective ways to set a *mood or convey essential information. What is less clear* is what the possibilities for interactive sound/action paradigms might include, and what perceptual ramifications might ensue.

I begin by extending the most obvious and naive interactive sound paradigm - what might be called "response feedback", simple a direct mapping between some kind of action (whether triggered by a user or an internal process within the computer) and a sound:

This might be represented, more generally as a set-based model. By set-based, I mean that it regards sounds as atomic objects, rather than as a continuous stream, although this distinction is not as mutually exclusive as it might appear.

So, more generally we have what really amounts to a mapping of two sets:

Notice that each action is mapped to exactly one sound, and a *sound can be mapped to any number of actions. Sounds in* this model might be extremely simple and functional such as sounds that get the user's attention, or to indicate success or failure.

Some things to note about this model are that the amount of information conveyed by an action is very small - usually just a message such as "I'm a button and I was clicked" or "I'm a process who is done".

This model also makes use of a single perceptual phenomena - reinforcement. For simple purposes, reinforcement is useful in extending the dissemination of information to the auditory sense. For complex purposes, in which there are repeated *actions or excessively many actions or sounds, the brain* quickly finds a repeated association banal and repetitious In the case of too many actions and sounds, one has difficulty remembering which sound meant what.

Certainly, more complex and more musical extensions of this model are not always warranted to improve the quality of a presentation. However, there are cases in which a presentation might benefit from a more musical treatment of sounds or might make use of highly complex information, aspects of which might be apprehended by an aural representation of data. This aural representation might convey patterns or structures in a different way than purely visual information.

Sonfication is the name given to rendering data aurally (often in conjunction with a visual representation). NCSA--the National Center for Supercomputing Applications, for example, has employed sonification in weather representations 3-d sound and other complex parametrization of sound is increasing employed by researchers in large-scale virtual reality environments such as the Cave, and in situations such as piloting or medicine where presenting rapid and useful feedback aurally is useful because the visual sense is presumably preoccupied.

What I propose to do at this point is to add one little piece, an agent, to this basic model (and its stream-based equivalent) and develop a basic set of oppositions that can characterize these models, albeit in a non-hierarchical way. For each of the models I will present, I will comment briefly on how they might be implemented, using the technologies which seems to be changing substantially month to month.

The insertion of an agent into the picture amounts to making the more relationship between a single action and a single sound less overt. A repeated action might trigger different sounds, a sequence of actions in a certain context might lead to a unique result. Of course, this scheme might easily result in aural confusion, but what we are looking at is what happens when a successful case emerges.

That is, assuming it is possible to construct a suitable agent, what might this model have to offer above and beyond the naive model?

The data structure or message format of an "action" forms the basis for what the agent has to work with. At present, most

real-time actions in multimedia contain very little information. Hypertext links, for example, which form the basis of world wide web navigation, only offer one piece of information, a destination page. In another paper, entitled Beyond Hypertext, I explore this subject in more detail. For the purposes of this paper, I will mention a few examples of some information that actions might include.

One example is categorical information hierarchies - if my basic message was "Fred" I might include the categories Fred belongs too - race, gender, income. If actions communicate information about categories, the agent can derive more general inferences about what someone is interested in, or try to develop connections and patterns.

Another example is temporal and historical information - if my action is a mouseUp message on a word in the midst of a tract on 18th century French economics, it might also be useful to know how long the page was being viewed, how fast, what the sequence of pages the user followed to get to this point, or how many times the viewer has chosen to look at this page.

In addition to discrete information, actions can also send a continuous stream of information - the most common example is the position of the cursor or other input devices.

It seems inevitable that actions will grow more complex as multimedia technology evolves - if only driven by corporate and marketing interests to glean as much information as possible. The interactive paradigms of music right now are primitive and although they certainly might remain so, it seems likely that people will explore and attempt to develop more complex paradigms. Should the right outstanding examples and channels of dissemination combine, some kind of Kuhnsian paradigm shift in this area might feasibly occur.

It is worth noting that certain technologies can often impose a stream-based or set-based way of thinking.

For example, Quicktime 2.5 can play MIDI information (which is essentially stream-based) but it imposes the set-based idea of a movie - hence each sound must be loaded as a set. Shockwave's audio capabilities are similar. Most streaming audio currently available (such as RealAudio and LiveWire) relies on the server first loading a complete sound file and then sending it, little provision is made for modifying what is being sent. Thus, the server side of most networked client/server models imposes a set-based methodology even when the client is stream-based.

A purely stream-based server/client model would most likely utilize MIDI. Part of the difficulty, of course, is that clients are usually freely available whereas server technologies are most often proprietary and expensive. When stream-based servers for MIDI or other sound abstractions are developed, their widespread acceptance is always predicated, to an extent, on their marketing niche

Having a MIDI stream-based server (in which data is created rather than read from a file), makes it easy to change many music parameters about the music being transmitted. Depending on the client, the sound itself (that is, the timbre) can be changed Many computer music research centers such as CNMAT in Berkeley and the Audio Development Group of NCSA have working client/server models in which the client directly synthesizes sound using FM synthesis or some other simple and fast synthesis method. In these cases, the message format uses some proprietary format. Ideally, one might conceive of a synthesis engine that runs on a computer that can produce a musically diverse and comprehensive set of timbres and controllable parameters while, at the same time, accommodating some degree of standardization

The trade-off in the above dilemma is that, while is not complicated to extend the computer's ability to generate a wide range of electronic sound within the context of networked multimedia, the ability to play real-world sounds remains difficult. It remains an obvious milestone to shoot for, driven by the urge to have networked multimedia match the quality of local multimedia This does contribute to thinking about how close we could be to having reasonable synthesized sound Although there are certainly moments in which vocal and sampled sound are desirable, the tools of synthesized sound are not even available to a widespread public. The release of QuickTime 2.5 with its built-in synthesized instruments is one of the first such tools, another is the Crescendo Netscape plug-in. Both of these use MIDI, probably because it is public, free, and standardized.

A major issue that conditions the whole notion of the server loading a sound file and sending it is that it does not seem possible for a server to send sampled data in such a way that the server is actively modifying, resynthesizing, mixing the sounds (here local multimedia has made some significant improvements in the last few years- just look at the sound manipulations taking place in the You Don't Know Jack CD-ROM).

Sending raw high-quality sound would require about 170 k/sec, or compressed with a 8.1 ratio 22k/sec, whereas the 28.8 modem or the 56-128 fractional-T1 users that comprise a large portion of the audience for the next few years average between 3 and 7k/sec. It remains to be seen just how tolerable compression ratios above 16:1 will be, and if some hardware sound decoding device, such as MPEG2, will become a standard.

To cite another possible example, one might construct a client sampling instrument (either in JAVA or as a browser plug-in) to which a short sample could be sent followed by instructions to loop, change pitch, change the envelope, reverse the sample, much in the same way that hardware samplers work. This approach could be developed to allow some musical treatment of sampled sounds with a reasonable amount of transmitted information - one large burst, followed by much smaller messages.

Having a server actively operate on sounds prior to sending anything is as much a conceptual hurdle as a technical one– this becomes a form of composition and, consequently, will take time for people to develop musically successful ways of having what amounts to an agent make some basic musical decisions.

It is worth mentioning JAVA in this regard as well, since JAVA remains one of the great white hopes of raising the level of programming in networked multimedia to the standards set by local multimedia. Its promise is clearly its ability to be hardware-independent, as well as the fact that its design seeks to make it easy to incorporate graphical and networking functions by encapsulating them in standardized libraries At present, however, JAVA still looks and acts primitive. History has shown it often takes the weight of substantial commercial software development to produce proprietary interface and window toolkit elements (such as those developed by Microsoft and Adobe) that raise the ante for what software should look like.

JAVA at the moment has only the ability for a server to load and send a sound file to a client. What it promises, however, is the flexibility to design the behavior of clients to which information can be sent as well as the structure of the information to be sent One important strength of JAVA that could prove

to be very relevant in the evolution of sound-action paradigms is its multithreaded nature. A thread can be thought of as a sub-program that runs concurrently in its own space, once it has been spawned by another program. A server could use threads, for example, to create a program that might process some sounds or sound abstractions in the background, while continuing to listen for client requests and whatever else it is doing. In essence, this style encourages parallel thinking where different programs handle different things - in a client/server environment, for example, there could be several server threads each of which is relating to a client thread. This flexibility means it would be possible to create the idea of an orchestra of different threads each contributing to the overall soundscape.

Some other models to consider include the relationship between sonic and visual objects. For example, what if sound and visual objects are considered as aspects of one unified object? A transition from one visual object to another might entail a corresponding transition from one related sonic identity to another.

The issue I am interested in stressing here is that models like this have yet to be explored. Conceptualizing these models as abstractions permits us to make a comprehensive survey as to what possible models might be, prior to considering their implementations. Once the range of models has been established, it remains to be seen what the perceptual efficacy of the models might be.

Conceiving of a unified visual and sonic identity opens up a wide range of analogical possibilities. The challenge is to find suitable relationships between primary types of evolutionary behavior for visual objects (such as scaling, rotation, color, illumination, translation, topological distortion) and primary (e.g. pitch, amplitude, stereo placement) and secondary (e.g. style, density, momentum) musical parameters. It also remains to be seen what alternatives there could be to transitions based on proportional relationships between visual and musical or sonic parameters.

Another possibility in the realm of unified visual and sonic identity concerns the interaction of two such objects.

In this case, the relationship between visual and sonic parameters is harder to imagine, governed by modes of interactive behavior rather than a discrete logic of perceptible transition. Behaviors--such as personality development, conflict, cooperation, and imitation--which are at times opaque when represented musically, are readily apprehensible visually. Psychological studies show that people are quick to associate personality and psychological archetypes with even simple visual shapes engaging in basic behavior patterns. Music would most often been seen as an ancillary foil to this visual behavior, a role which it could choose to accept or challenge.

Implementing this behavior is possible in both local and networked multimedia using JAVA or Director-related software. JAVA makes it straight forward to create a unified visual-sonic object as a single class. Typically, interaction would be monitored by a separate Manager class that watches to see when interaction occurs and determines how to define the interaction. Unlike Director, JAVA makes it easier to conceive of different objects as separate programs running concurrently in parallel.

In closing, I should like to make a few comments about the distinction between the models I have just presented and the naive basic model of response feedback. One way of viewing this distinction is to view the naive model as essentially deterministic - that is, there is little or no ambiguity about the relationship between action and aural response. Adding complexity to this situation increases the level of ambiguity, which is an important musical construct that needs to be handled musically. For example, Leonard Meyer in his work Music, The Arts, and Ideas argues that ambiguity is most often used in the initial presentation of a musical idea to extend the horizon of potential futural options. Ambiguity is then diminished as definite choices are made, their latent ramifications realized or confronted. At certain moments, when a certain choice would seem blatantly predictable, ambiguity might be reintroduced to rekindle tension, establish contrast, or begin a source point for a new musical direction.

I have outlined a basic set of five oppositions:

- sound/unified sound-video
- primary/secondary
- simple/complex
- stream-based/set-based
- atomic/interrelated

I hope that these might serve a useful taxonomic and/or descriptive purpose. One could describe a sound-action example as having a certain set of these qualities--such as interrelated, unified, primary, simple, set-based--which would at least provide a clear implementation-independent description of what that sound/action model does. It is my hope that some of these models might inspire people to develop models of their own and to extend the language with which they can be described.