Correlated sound and image in a digital medium
Robin Bargar

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
605 East Springfield Ave., Champaign, IL 61820
rbargar@ncsa.uiuc.edu

ABSTRACT

In multiple media formats such as cinema and video, images and sounds are traditionally correlated only at their display. Digital media allow the correlation of sound and image data prior to displaying that data as pixels or loudspeaker cone positions. Correlation and display may be distinguished along three dimensions: (1) the procedural distance from a correlation to its display; (2) the representational distance from a symbol to its digital functionality; (3) the perceived triviality of visual and auditory correlation. Using software for scientific visualization, graphic design, and sound synthesis, composition tool prototypes have been implemented on the Silicon Graphics Indigo computer. Examples of applications include compositions for video and for real time performance.

2. INTRODUCTION

The impression that a picture occupies space rather than time describes, not images, but the iterative nature of visual observation. The pattern of iteration may differ from that of listening, but motion picture and video frame rates, and CRT refresh rates demonstrate that static and moving images are observed by iteration. Viewing, like listening, may translate observations both into space (standing waves in resonant bodies) and time (one duration compared to others).

Motion picture media provide paradigms for the combination of image and sound. These paradigms rely upon synchronized sound-image display characteristics. Perhaps because image and sound are dissociated during production, great efforts are taken to ensure that strong correlations are established during their display. Synchronization is often an attempt to compensate for a lack of structural correlation.

Digital media have adopted, or emulate non-digital representations, however, computation offers symbolic capabilities radically different from the lens and the microphone. Digital media can correlate image and sound data prior to their mechanical display. Correlation does not imply synchrony.

The present compositions distinguish correlation and display along three axes: (1) the procedural distance from correlation to display; (2) the representational distance from a symbol to its digital functionality; (3) the perceived triviality of visual and auditory correlation ("trivial" in the sense of the hypothetical trivial machine - flawlessly predictable and unvarying).

3. CORRELATED SOUND FOR VIDEO PRODUCTION

Video production of computer graphics is significantly non-real time. Video displays 30 or 60 frames per second, each requiring an image rendered by the computer. Single frame computation time can easily require 20 minutes to 20 hours. Rendered frames are downloaded in non-real time from computer to videotape, then edited as standard video footage. Sound is applied to the image during post-production.

Wavefront Technologies software is designed to produce animations for the video format. Wavefront provides modular tool environments for creating rigid body animations in three dimensions. Polygonal
objects are built (Model environment), their motion over time is specified (Preview environment), and still frames are rendered and recorded to videotape (Image environment). A GUI/command line interface allows graphic construction/design using the mouse as a drawing and selection tool. Wavefront has an open architecture that encourages interface with user code for custom graphic processing.

Key-framing is an animation technique with strong ties to computer music. Key frames are values designated at specific time points along an animation path, usually an x, y, or z position, though any animation parameter (channel in Wavefront) may be assigned a key value. Key frames are guidelines; between key frames values are computed by interpolation functions. Wavefront displays graphs of channels showing these functions (splines, user-defined curves or linear interpolations) between keys. Transformation paths are computed and displayed based upon key frame positions and selected interpolations. Using custom software, channel data may be sent to sound synthesis parameters, establishing a data correlation between a graphic and an acoustic transformation, and converting the advanced Wavefront graphics interface into an hybrid synthesis interface.

During production of the video animations Garbage (Cox 1991) and The Listener (Landreth and Bargar 1991), auditory-visual correlations were created in Preview. Software modules were written connecting animation channels and sound synthesis parameters. Parameter changes were calculated according to an audio-visual storyboard/score. Pre-composed rates of change served as computational constants. Sound was computed from the same parameter changes which controlled the animation, and re-synchronized with the image during post-production. In both videos, the data-driven sounds were combined with other sounds to create compositions which interrogate the video soundtrack as a genre.

In Garbage sound samples were data-synchronized by monitoring a select motion channel: reversal of motion value indicated when an impact-event occurred, and an impact-sound was assigned a start time for the current frame. In The Listener the parameters controlling three large muscle groups in an animated face were assigned to the carrier frequency, modulating frequency, and note-on/duration of a software FM instrument. An offset factor was incorporated to prohibit synchronization with the corresponding animation. Visual rhythms were automatically translated into sound and imbedded as an echo of visual events. Post-production video edits were aligned with the continuous soundtrack.

4. CORRELATED SOUND FOR REAL TIME DISPLAY

Real time animation places restrictions on spatial dimension and polygon count. Hardware is measured in terms of polygons-per-second, either by number of computations or by display rate. Given a two or three-dimensional representation, and given the geometric transformations rotation, translation and scale, real time animation is limited by the number of polygons that can be processed while maintaining a display rate of 12 to 30 frames per second. Photorealistic images drastically increase computational costs. In order of complexity these include shaded surfaces, rounded corners and edges, directional lighting, shadows, atmospheres, textured surfaces, and ray-traced reflections. A hardware graphics pipeline increases processing speed: dedicated IC’s (integrated circuits) perform the above computations. Pipeline functions appear to the programmer as graphics library calls. In the sound synthesis domain, MIDI can be thought of as an audio hardware pipeline, with analogous tradeoffs between complexity and speed.

Using the Silicon Graphics Indigo, prototype sound-image software modules were developed for Redline 7000, a real time composition for string bass and multimedia computer. These modules perform real time animation and sound playback. Some modules synthesize sounds in real time while computing animations, other modules use sounds and animations that were pre-computed with correlated data, stored as files and displayed when triggered by a performance control routine. The live performer assumes varying degrees of power in relation to the computer; sometimes following the constraints of a traditional tape-and-instrument format, other times able to use direct gesture input. Performance display modules were constructed by modifying the software applications RTCA and NCSA PolyView.
4.1 RTICA

Mathematician George Francis created the Real Time Interactive Computer Animator for rapid display of mathematical functions and topologies. RTICA generates complex geometric figures for stereoscopic three dimensional viewing. Francis’ program displays surfaces and curvilinear paths moving across 3D displays of homotopic (one-sided) surfaces. Transformations are controlled by mouse and keyboard parameter scaling or by session script files, which may be recorded, edited and replayed. For performance, RTICA was limited to displaying lissajou lines across variable surfaces. The lissajou comprise two frequency-modulated sine curves, with variable curve resolution, carrier and modulation frequencies and amplitudes. Visual modulation frequencies (other than the display rate) are sub-audio.

Two approaches were applied to sound correlation. In the first an FM oscillator routine was added and various mappings were applied between graphic and FM parameters. In the second, the values used to draw pixels were scaled and stored in a buffer which was auditioned directly as a wavetable. The drawing routine was used as a synthesis routine, and every pixel value was translated into a single sound sample integer. Output amplitude correlated closely with the input lissajou amplitudes; output pitch was a function not of input frequency but of curve resolution: a more angular curve involved fewer values in the drawing loop, thus a shorter audio buffer resulting in a higher loop frequency. Input lissajou frequencies tended to modulate the pitch component established by the buffer length.

4.2 NCSA PolyView

PolyView, a native Silicon Graphics application, displays polygonal data sets such as those used in finite element analysis, surface modeling, and sparse matrices. Data must be stored in a Vset (vertex set) within NCSA’s Hierarchical Data Format (HDF). HDF is a multiple-object file format for the transfer of graphical data between machines (Cray, Alliant, Sun, Iris, Macintosh, and IBM PC), allowing for self-definition of data content and easy extensibility. HDF’s Vset extends HDF to allow storage of non-uniform multivariate data sets, targeted toward manipulating mesh, polygonal, or connectivity data. Like Wavefront object files, a Vset represents a polygon as a collection of vertices in 3-space, plus connectivity lists (showing which vertices are connected by edges).

PolyView display features include orthogonal and perspective projections, shading and lighting models, color palette editing and animation, animation of sequences of Vsets (object transformations), camera animation (fly-by), and script files for replaying image manipulation sequences. For Redline 7000 polygonal objects representing 3D music manuscript notations were designed using the Wavefront Model utility, converted to HDF Vsets and placed into PolyView. Real time performance sequences were created using scripts.

Two forms of sound correlation were created in PolyView: transformation of sounds using channel data for object transformations, and the triggering of sounds based upon camera proximity to objects during display. The first method resembles video production correlation: single graphic objects were deformed in a series of steps, with a new Vset created at each step, creating a sequence of Vsets to be played back as an animated deformation in PolyView. The deformation control channel data was also used to create a software synthesis score, pre-computed and stored as AIFF soundfile. The PolyView script initiates the deformation sequence and sound file playback concurrently; they display in parallel without true synchronization. Animation speeds depended on number of vertices; soundfile playback required buffer size adjustments to compensate.

The second method implemented a prototype Virtual Reality audio technique: sound files associated with particular vertices are triggered to play when the camera moves within a predefined valence around those objects. Camera motions "perform" objects while moving through space. Currently this technique requires painstaking setup of sound mappings to vertex groups and thresholds for trigger conditions. MIDI output signals can be substituted for internal soundfiles. Further development is
planned to generalize the sound-vertex mapping procedures, and to explore real time synthesis capabilities, such that vertex groups could provide more complex input and control continuously variable sounds.

5. ANALYSIS

A representation is distant from digital data by masking digital functionality. Popular interface techniques represent digital functions as familiar mechanical tools. Representational distance becomes problematic when the richness and power of digital transformations are not incorporated in a symbol. Audio contributions to the media environment tend to suffer heavily from inadequate symbolic framing.

Correlation of sound and image based upon display characteristics often occurs in direct proportion to non-digital representation of sound and to trivial correlation. Display correlations are essentially mechanical post-production models whereby sound is added to a completed image. Post-production models argue against independent organization of sound, particularly digital techniques which (1) do not produce naturalistic sounds or (2) present counter-intuitive relationships to the image.

Wavefront provides a potentially powerful hybrid synthesis interface, however, correlating sounds with fixed polygonal objects in fixed three-dimensional space tends to emphasize traditional notes and note-patterns rather than lower-level sound or image transformations. Reduced access to structural transformations emphasizes simple display synchronization rather than relative degrees of correlation. The more primitive linear creative voided single-axis/single-parameter correlations. However, it did present significant problems in audio and graphic buffering co-dependencies. (The independence of the Indigo's dedicated graphics IC and audio DSP prohibits a simple synchronous image-sound display clock.) A goal of future work is to maintain flexible low-level correlations and construct complex graphics and sounds which do not mask their correlated primitives.

6. APPENDIX: HARDWARE AND SOFTWARE

The Silicon Graphics Indigo is a UNIX-based personal computer implementation of advanced 3D graphics GUI, with video and CD-quality audio capabilities. The R30000A RISC CPU clocks at 33mHz; busses and CPU clock independently. Graphics utilize a dedicated IC for rapid processing of 3D display primitives/Smooth-shaded polygons at 40 million pixels-per-second with 24 bit difference, arbitrary clipping planes, and z-buffering for depth cues). Audio is managed by the Motonola 560001 DSP without CPU intervention, supporting 24 bit digital and 16 bit analog stereo. I/O includes both analog and direct digital coaxial serial at 24 bits at sampling rates if 8, 11, 16, 22, 32, 44.1 and 48 kHz. Digital I/O formats include consumer IEC958 and professional AES3. An optional DAT drive offers file storage, digital audio transfer, and random access audio playback. Audio library functions support AIF, NeXT, and MIDI.

The NCSA Software Development Group produces free scientific visualization software and source code, distributed to the public domain. Apple Macintosh and Mac II, IBM PC, Sun, Alliant, Silicon Graphics, and Cray are among the supported platforms.

7. REFERENCES