BEYOND SYNCHRONIZATION
towards formal coherence in computer-generated visual music

Overview and Synopsis

Dr. Mark Zaki
Assistant Professor of Music
Director, Rutgers Electro-Acoustic Lab (REAL)

Chancellor’s Faculty Research Day
Rutgers University - Camden

November 4, 2010
This work has been generously funded in part by

The Andrew Mellon Foundation
The Penn Humanities Forum at the University of Pennsylvania
Beyond Synchronization: Towards Formal Coherence in Computer-Generated Visual Music

The digital “revolution” of the 1990s and convergence of video and audio software has given composers wide access to the rapidly expanding field of computer-generated visual music. Despite established conventions borrowed from film and other visual media, there are creative challenges for composers interested in the intersection between musical and visual languages. Compositional strategy often relies on higher-level processes and “virtual creativity” as part of the composer’s methodology. In most cases, audio and video are created according to separate aesthetic and pre-determined hierarchies, and combined using varying levels of synchronization. This study looks at the possibilities describing opportunities for more sophisticated coherences between audio and video, leading towards conceptual models that support highly integrated, abstract visual music.

Visual music can be described as abstract fine art animation in which the visual elements are informed and complemented by musical processes. These works are typically realized through varying degrees of digital processing onto fixed media, and in real-time using hardware control mechanisms and interactive software.

The combination of audio and video is certainly not a new phenomenon, but it is only relatively recently that we have been able to create and manipulate both the audio and visual data streams within the same software environment without

“In the impossibility of replacing the essential element of color by words or other means lies the possibility of a monumental art. Here, amidst extremely rich and different combinations, there remains to be discovered one that is based upon the principle [that] the same inner sound can be rendered at the same moment by different arts. But apart from this general sound, each art will display that extra element which is essential and peculiar to itself, thereby adding to that inner sound which they have in common a richness and power that cannot be attained by one art alone.”

— Wassily Kandinsky (1912)
considerable hardware overhead. The personal computer has become the \textit{de facto} conduit for media manipulation and, by extension, making structural connections between music and image. The potential of the computer as an agent for visual music creation was recognized long before the widespread gains and accessibility of computational and processing power we see in today’s cutting edge systems. Pioneering film maker John Whitney Sr. notes that, “the computer [is] the only instrumentality for creating music inter-related with active color and graphic design, and though the language of complementarity is still under-examined and experimental, it foretells enormous consequences and offers great promise.”\footnote{Whitney, John Sr. “To Paint on Water: The Audiovisual Duet of Complementarity,” Computer Music Journal, Vol. 18, No. 3, Autumn 1994, 46.}

Given the vast abundance of decisions facing the visual music composer, there are a seemingly infinite number of strategies that may be employed in composing for video and sound. Concern for audio-visual coherence may uncover possibilities in terms of a consistent compositional language for the creation of experimental new-media works. The ultimate aim is to locate meaningful and appropriate connections in order to produce logical and tightly integrated visual music.

Existing literature on the types of connections between audio and video sources in a dual media context often frame a subjectively defined context or agenda. Casually perusing these resources also seems to reveal more commentary from a visual perspective looking into the world of music composition with respect to forming audio-visual relationships. However, from a musical perspective visual-centric relationships can break down with broader scrutiny or aesthetic bias. Looking at the issue from the other angle, purely music-centric relationships may not always prove fruitful from a visual perspective either. It’s probably a reasonable assumption that a “grand unifying theory” isn’t completely possible,
Beyond Synchronization: Towards Formal Coherence in Computer-Generated Visual Music

nor may it always be desirable. It’s also likely that this particular study will simply describe another personal agenda as well. However, developing a means to understand and determine specific relationships between these two disparate and sometimes autonomous media sources can allow for a structural logic that potentially justifies digitally generated mixed-media works.

Generally speaking, most mixed-media composition is created using one of two models. The first is similar to conventional film practice, where video is initially created and edited with an audio score matched later. In the second, video is cut and arranged to a pre-existing soundtrack in the manner of popular music video. Both of these models may be thought of as “responsive” or “reactive” systems. In other words, one media domain is always created in response to the other. Contrary to these reactive practices, one goal is to uncover processes that generate a combinatoire that is more than mere synchronization.

In order to explore the creation of integrated mixed-media objects and experiment with their behavior, a suitable programming environment is needed to set up appropriate conditions for observation. This environment can be found in a widely available visual programming language called Max/MSP/Jitter. Max/MSP/Jitter is an interactive programming environment, which allows the user to create software using a visual toolkit of objects, and connect them together with virtual patch cords. The basic environment, which is known simply as Max, includes MIDI information and messaging, control routing, user interface elements, mathematical and logical functions, and timing objects. Layered on top of Max are hundreds of objects, including two powerful specialized collections: MSP, a set of audio processing objects that do everything from interactive filter design to hard disk recording, and Jitter, a set of matrix data processing objects optimized for video and 3-D graphics. The Max program itself is highly modular, with most routines existing in the form of shared libraries. An application programming interface allows further development
of new routines (called “external objects”). Because of its extensible design and graphical interface, Max has become a leading platform for developing interactive music and mixed-media performance software with many of today’s composers, artists and multimedia programmers.

There are several pre-existing paradigms that suggest ways to combine audio and video beyond simply synchronizing them through temporal alignment. In the first approach, we can look at the so-called “visualizations” that we find in software such as Winamp or iTunes. A still image, video or generated geometry is used as raw material, which is then subjected to video processing such as geometric deformation, feedback, motion blur and so forth. This processing is modulated based on information taken from analysis of an accompanying sound file. In this paradigm, an audio analysis object can be created to look for specific parameters or characteristics of a given sound file. These parameters are mapped to a processing action on a corresponding video clip. In this manner, the video processing has a direct correspondence with control information originating in the audio stream. Control actions may be discrete one-shot triggers, or they may exhibit continuous behavior (Figure 1).

![Figure 1 - “Visualization”](image-url)
This is a prevalent paradigm where audio information is used to control video processes. While this is easy enough to achieve, the overall effect is much like a glorified light organ. What is ultimately problematic is that regardless of the sound source, static mappings appear to generate similar visual results.

A corollary to the visualization model, and seemingly less prevalent approach is a similar “sonification” model. In this case, it is the image that is analyzed in order to provide information that is mapped to a receiving object that controls a processing or generative action on a corresponding audio clip (Figure 2).

This is also the basis for various types of motion tracking objects that can be created in Max such as Eric Singer’s Cyclops object. Initially designed for live video input, these objects can analyze each frame of a video in real-time. They divide the image area into a grid of rectangular zones and analyze each zone for color values, difference (motion tracking), threshold and grayscale. The user may specify the grid resolution, target zones for analysis and indicate the type of

![Figure 2 - “Sonification”](image)
analysis to be performed in each zone. Messages are output for each analyzed video frame that can be used to trigger any Max/MSP processes or control patch parameters for audio processing (Figure 3).

Another approach begins with the assumption that audio and visual data streams can share control information for DSP structures in parallel, and use this shared information to create linked output. For example, changes in cartesian coordinates could be mapped and sent to receiving objects that simultaneously control processing action on both audio and video files (Figure 4).
This is the rationale behind a device such as the Kaoss entrancer. This is a controller that is made by Korg, and it is designed to provide real-time control over digital effects processing for both audio and video sources, either together or separately. Users move their finger along a touch sensitive X/Y pad to control the way in which effects are applied. This touch (or sweep) of a finger allows a number of parameters to be manipulated, with easy synchronization of the video and the audio. This is an interesting and viable approach, particularly with respect to live performance. It does seem to have limitations however when it comes to making formal connections between the data streams. If anything, the linkages are “superimposed” based on external considerations (namely the user’s finger gesture). What also remains to be seen is how the proliferation of multi-touch input devices, such as the IPad and other tablet computers, change the dynamic of interactivity, performance gesture and multimedia control.
The primary concern remains, and this is the degree of formal and syntactical coherence that can be achieved between two separate and disparate data streams. From an examination of the paradigms described earlier, a set of goals have been idealized which could potentially yield a system of coherence and allow for formal control. These goals define a system with the following general properties:

- The system enables the creation of dynamic, integrative multimedia, ideally in real time.
- The means of input is sophisticated enough to accommodate a compositional logic.
- Audio and video characteristics are commensurately malleable.

With simple adjustments of the serial models previously described, depth and flexibility can be achieved by passing control and analysis information in both directions. At the least, this bi-directionality greatly increases integration and the “inheritance” that the output can exhibit. This system is called **CROMA** (CROss-MApping).

**CROMA** extends the modulation “metaphor” shown earlier in the visualization example by setting up bi-directional control structures; integration is achieved by enabling digital signal functions in one media stream to be controlled and modulated by information derived from the other stream. These sources are mapped to specific control targets that influence the processing in the other domain. In this case, both data streams are analyzed and used to control the other stream simultaneously. The resulting combinative object is largely determined by how much information from both streams is used (Figure 5).
Cross mapping of signal processing and its control becomes an underlying principle on which compositional logic can be built. This assumption is based on the idea that interactive processing control achieves a much tighter degree of integration between the media streams than is possible through independent and unrelated processing. Forming a logical strategy that governs the behavior of the processing opens up the possibility to systematically integrate and link audio and visual streams. There is a “genetic expression” that occurs as a matter of course, with the resulting object exhibiting inherited characteristics from both data streams.
So how can this inform a formal coherence between the data streams? To start, we can look at the parameters that are easily accessible in each of the streams. On the audio side, we can determine “beats” or transients by specifying a threshold that can trigger an event as the amplitude of the signal passes through it. It’s also quite easy to set up a signal “reader”, which can report the changing value of the amplitude over time. This is useful for specifying continuous control values. On a somewhat secondary level, we can perform a Fast Fourier Transform, which can describe the values or rate of change of specific frequency bands over time. This is useful if we want to divide up the sound file into different frequency ranges, with each range perhaps controlling a different target. Also of some use, is the ability to locate a sound spatially, or by determining its pan position. If the sound is moving, we can use its changing amplitudes in space to “track” the sound and convey this information to an appropriate target.

On the video side we have a similar situation, although digital video information is significantly more complex. A general list of video parameters could include color values (one each for the RGB values and one for the alpha channel). We can also extract control information from motion detection, threshold values (similar to transient detection) and the level of grayscale information present.

Once we’ve determined where our control information is coming from, we can then set our target parameters. Respective mappings can be indexed within an audio-video parameter matrix (Figure 6). This shows a simple matrix with some arbitrary parameters being targeted, but again this can get quite complex depending on the type of processing that is employed and the level of detail that is desired.
Using a matrix like this, relationships can be structured to implement various compositional strategies. Perhaps the most significant axis on which to base these strategies is the fact that both media streams are time dependent. Temporal relationships can be defined quite accurately with respect to processing schemes, in turn strengthening the overall compositional structure. While it is possible to fold parameters back towards targets within their respective domains (the grey-shaded areas), this may not always prove to be desirable or successful as the case may be.

Since both data streams are temporally based, various algorithms may be employed to formally position “objects” as well as govern their respective processing and behavior over time. There has been some success in this regard using compositional principles from serial time-point techniques to systematically organize the temporal dimensions of a piece. This gives a fundamental level of coherence between the data streams and further helps to define the structural form of both the combinative output and the work as a whole.
This process can introduce a fair amount of latency, which makes it somewhat impractical for live performance. There is a parallel project which seeks to incorporate many of the findings into a system which can be performed in real-time. Currently, there is a prototype of a system built in Max which uses the violin as a front end. This system is called ViPerS which is another acronym for Violin Performance System (Figure 7). ViPerS has an analysis layer which tracks the violinist’s pitch, bowing, articulation and gesture. This information is converted to control messages which then are targeted in a similar fashion as CROMA.

![Figure 7 - GUI for ViPerS](image-url)
On the positive side, ViPerS gives a great deal of control to a performer, and retains the natural gestures of the given instrument. On the downside, this removes an important assumption about coherence, namely that music and video information are on equal footing. The resulting mixed-media object is an artifact of the player’s musical gesture, rather than a logical component of a compositional system. The same situation holds true, if to a lesser degree, if one were to “play” the system using a specialized interactive instrument or custom-built controller.

There is one fundamental conclusion that can be drawn, regardless of the approach being used. Most artists and composers tend to approach the issue of coherence in visual music from the perspective of their discipline, and define their methodology in those terms. It seems apparent however, that the real action exists in some meta layer – an umbrella layer of information and logic which serves to unite the two streams by presiding over them. We can see this in the idea of gestural control and coherence in the case of ViPerS, and in the imposition of compositional logic in the cross-mapping object that links audio and video in CROMA.

Regardless of the approach, there are three creative questions that need to remain in focus:

- Do these approaches lead to expressive forms which lend themselves to “narrative” and not feel contrived? Or is that the point — the message of the medium?
- Can these systems allow access into more developed and engaging perceptual experiences?
- If so, how can these engage, without sacrificing creative aesthetics or abstraction? In other words, how do we reach the viewer’s imagination and emotional center beyond abstract eye and ear candy?
Developing procedures that inform the coherence of abstract media composition is attainable. Through the establishment of models based on intermedia relationships, distinctive media works can be created that exhibit a consistent logic throughout their structure.

This work is an important step towards continued development of a compositional language for experimental new media. While the artistic signature of the work is naturally subjective and reflective of personal aesthetics, this approach allows flexibility in a variety of styles and contexts.

The hope is that this synopsis has given the reader some insight into some of the challenges and possibilities available in forming strategies for abstract new media composition. Using systems such as the ones described address several issues that concern organization and the integration of dual media streams. Finally, there is also a personal acknowledgment that this is a connection to past traditions regarding form and organization, as well as a starting point for exploration of a medium that is very much a part of the future of artistic expression.
SELECTED BIBLIOGRAPHY


