The following is a reviewed excerpt from the 1988 MIT Media Laboratory M.S. thesis

Perceptual Correspondences of Abstract Animation and Synthetic Sound

by Adriano Abbado

Thesis Supervisor: Tod Machover

Reviewed in collaboration with Melinda Mele

"Certainly I have an aversion to everything that is demonstratively programmatic and illustrative. But this does not mean that I am against music that calls forth associations; on the contrary, sounds and musical coherence always arouse in me ideas of consistency and colour, of visible and recognizable form. And vice versa: I constantly combine colour, form, texture and abstract concepts with musical ideas. This explains the presence of so many non-musical elements in my compositions." György Ligeti, 1968

Organizing synthetic sounds according to the conventional principles of Western music is generally quite problematic; in order to compose with synthetic sounds a different approach must be sought. A possible approach may be to use features taken from visual language and apply them to the aural domain. If it is possible to establish links between aural and visual events, then visual language can help create a synthetic-sound composition, as well as an audiovisual piece.

I have always been interested in the associations between different areas of thought. It is part of my natural way of thinking to see if certain ideas can function when transposed to a different context. I often attempt to abstract concepts from their original contexts and to apply them in others. This method often calls forth new ideas, and sometimes I use it as a tool for the creation of artworks. In particular, I often seek to establish parallel behaviours between aural and visual events, between the elements of music and visual arts.

I think that it is possibile to establish links between perceptual categories of synthetic sounds and abstract visual forms. This conviction has led me to consider different areas of application, chiefly through the creation of a new type of audiovisual work, based on the interaction between sounds and visual objects rather than music and images. The method proposed in this thesis is based on visual language, and can lead to a new way of thinking and creating music compositions based on timbres or tone-qualities.

1 "Dynamics"

"Dynamics" is an audiovisual piece conceived and realized with digital media, and originally shown with a videoprojector, a big screen and four speakers located at the corners of the screen. It was produced at the MIT Media Laboratory, using technologies belonging to the Visual Language Workshop, the Animation Group, the Film and Video Group and the Music and Cognition Group. Three correspondences between the audio and visual events occur in this work: timbre/shape-surface attributes, spatial localization and intensity.

2 Timbre/Shape-Surface Attributes

A sound can be abstracted as an aural object, as something that may be complex but has a precise identity. This identity is defined by the sound's spectrum, its energy. Similarly, visual events can also be considered as independent objects. The shape and the kind of surface define a visual object's identity. Timbre in sound, and shape-surface attributes in images, are the most powerful perceptual determinants. Their correspondence should, in my opinion, form the basis of any new sound-image language.

Traditional Western music is fundamentally based on pitch and rhythm. In the music of other cultures, however, timbre sometimes plays an equally important role. For example, tabla players create complex tone-qualities that recall the sound of voice (they even learn a vocabulary of syllabes before learning how to play the instrument) [S. McAdams and K. Saariaho: Qualities and Functions of Musical Timbre, Proceedings of the 1985 International Computer Music Conference, The Computer Music Association, San Francisco]. Similarly, Tibetan monks' singing is based on the modulations of vocal timbre, rather than pitch. It is interesting to observe that in both cases just mentioned, musical timbre is strictly related to singing, since vocal expression takes place through timbre rather than pitch.

During this century timbre has become increasingly important also in Western music. The first composer to assign a major role tone-quality was Arnold Schönberg, who coined the term "Klangfarbenmelodie" (melody of timbres). Schönberg more precisely stated: "I think that sound reveals itself by means of the timbre and that pitch is a dimension of the timbre. The timbre is therefore the whole, the pitch is part of this whole, or better, pitch is nothing but timbre measured in just one dimension" [A. Schönberg: Manuale di Armonia, Il Saggiatore, Milan 1963, pag. 528-529]. However, the first musician to create compositions indeed based on timbre was Edgar Varèse, who emphatized the importance of the so called “rumor” of percussion instruments: his “Ionisation” is a masterpiece. Timbre has become one of the areas of major research in contemporary music [T. Machover: The Extended Orchestra, The Orchestra, Joan Peyser editor, Charles Scribner's Sons, New York 1986], [M. Stroppa: L'esplorazione e la manipolazione del timbro, Quaderno 5, Limb/La Biennale, Venice 1985]. One of the great innovations that computers brought to music is the possibility to synthetize new sounds, dramatically increasing the number of instruments a composer can work with. Jean Claude Risset and David Wessel stated that "with the control of timbre now made possible through analysis and

synthesis, composers [...] can articulate musical compositions on the basis of timbral rather than pitch variations [...] It is conceivable that proper timbral control might lead to quite new musical architectures" [J. C. Risset and D. Wessel: Indagine sul timbro mediante analisi e sintesi: Quaderno 2, Limb/La Biennale, Venice 1982, pag. 28-29].

As was said, composing with synthetic sounds, and in general with timbres, requires a new approach. One of my objectives in creating "Dynamics" was to understand in which way visual language can enhance and clarify the process of composition of a piece of music based on the complex notion of timbre rather than on pitch.

Recently, Fred Lerdhal proposed two methods with which to organize new timbres, one using hierarchic structures and the other using traditional musical concepts [F. Lerdhal: Timbral Hierarchies, Contemporary Music Review, Vol. II n. 1, S. MacAdams editor, Harwood Press, London 1987]. Although both methods are powerful tools for dealing with complex compositional structures, as tools to conceive a music composition hierarchies seem to be too inflexible, and new timbres cannot be treated as timbres were in the past. Since a level of comprehension is attained when the perceiver is able to assign a precise mental image to what he perceives, visual representation of tone-quality can help the process of understanding timbres, and can substitute a specific hierarchic organization of the timbral elements themselves. Visual representation of timbres was therefore used as the basis for the organization of the musical composition.

In the visual domain, shape is probably the element that best defines an object. However, shapes alone cannot fully confer the perceptual characteristics and the richness of timbres. Consequently I also used other properties - such as color, lighting, and the variation of specular reflectance of the surface - to enhance the expressive power of a visual object. Texture mapping, a technique used in 3D computer programs, would have permitted me to further enrich the visual appearance of the objects I created. Unfortunately the program used for this piece, Anima (written by Bob Sabiston and running on an Hewlett-Packard "Renaissance Box" at the MIT Media Laboratory), although extremely powerful, did not permit the use of this interesting feature.

In this project the procedure for associating shape to timbre was of primary importance. Although there is a certain degree of agreement, at least among Westerners, regarding attributes that relate sounds and visual objects (as for example a metallic sound and a metal object), no deliberate attempt was made to

conform to such attributes; the relationships used in this composition were completely subjective. I associated harmonic sounds with smooth shapes and inharmonic sounds with jagged shapes, because I usually hear partials in harmonic ratio as non-aggressive, and in turn I identify them with the idea of smoothness, while I hear inharmonic sounds as irregular, aggressive objects. For example, in this work white noise was represented as a highly irregular, bumpy and shiny object, while filtered white noise (with high frequencies filtered out) was represented as a much more regular shape, although still shiny. It is interesting to report Wassily Kandinsky's observations in "Concerning the Spiritual in Art": "On the whole, keen colors are well-suited for sharp forms (i.e. a yellow triangle), and soft, deep colors by round forms (i.e. a blue circle)" [W. Kandinsky: Concerning the Spiritual in Art, Dover Publications Inc., New York 1977, pag. 29]; "Yellow, if steadily gazed at in any geometrical form, has a disturbing influence, and reveals in the color an insistent, aggressive character" [Ibid: pag. 37].

This is not a general rule, however. It is extremely difficult to establish criteria and formalize processes such as the one I employed in the creation of objects, as here the links between an aural and a visual object, rather than following a precise scheme of representation, were established also through feeling and intuition. Today, the neural network simulation used on computers could be a tool to understand the process of choice. Yet establishing a rigid method of correspondence can become a mechanical procedure which limits the creative process and is inevitably reflected in the finished work. Therefore I never analyzed the spectral content of the sound with absolute precision, but always conceived the objects closely referring to perceptual aural attributes (also used in a visual context) such as harsh, wet, cold, metallic, viscous, spongy, granulous, opaque, and so forth [D. Ehreshman and D. Wessel: Perception of Timbral Analogies, IRCAM, Paris 1978], [J. Grey: An exploration of Musical Timbre, Doctoral dissertation, Stanford University 1975], [D. Wessel: Low Dimensional Control of Musical Timbre, IRCAM, Paris 1978].

3 Spatial Localization

Another correspondence I established was between the positions of the audiovisual sources in space. Spatial localization and the use of space in general, are one of the most important and innovative features of contemporary music. Compositions using space as a parameter were already found in the Middle Ages, although with simple techniques. Among the important composers of this century to take advantage of the notion of space in a more complex way are

Edgar Varèse and Henry Brant [R. Erickson: Sound Structure in Music, University of California Press, Berkeley CA 1975, pag. 141]. Many postwar composers, such as Karlheinz Stockhausen and György Ligeti, further extended this idea into electronic compositions.

Computers now permit the simulation of position and movement in 3D space, usually by means of a number of speakers. In general, music should be considered as an artform that uses not only the dimension of time, but also the dimension of space, in the same way as visual arts should be thought of involving time as well as space. This prospect has not been developed in any significant way, having been used mostly as an effect rather than as an actual musical parameter. Even though the ear is not as sensitive to sound movement as it is to pitch change, or as the eye is sensitive to visual movement, it is nonetheless worth exploring the new possibilities that modern instruments, such as computers and DVD for instance, offer in this field. One idea may be to create a music composition starting just from a visual input, that is, from the placement in space of visual objects that represent sounds.

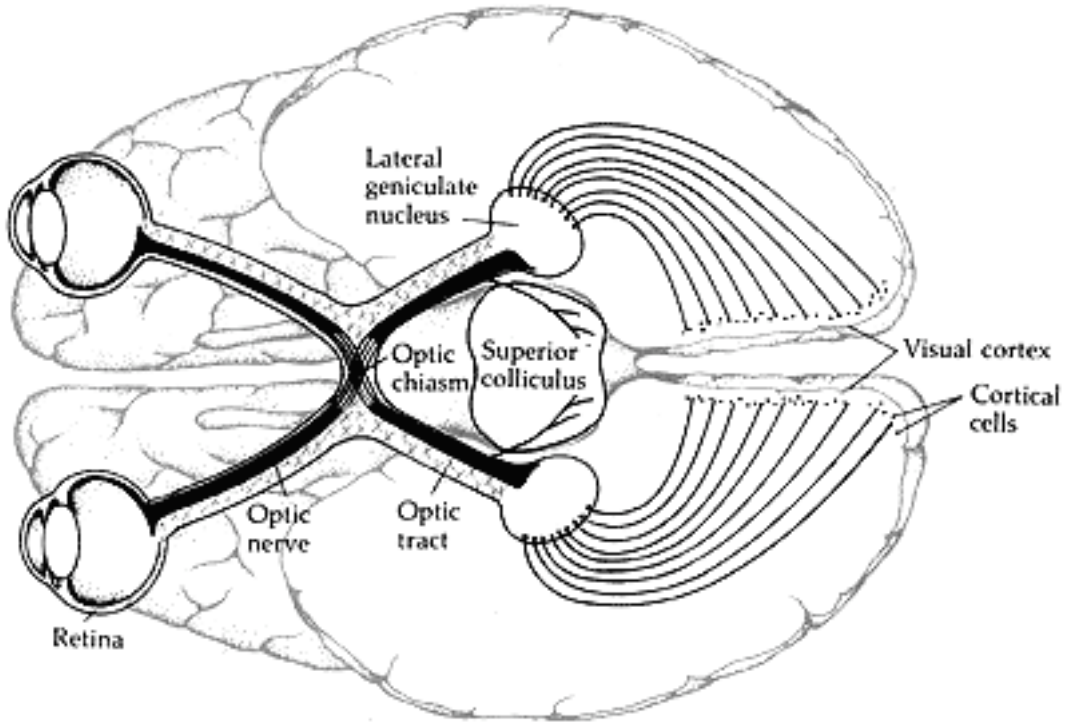

Fig. 1 - Superior colliculus - from R. Sekuler and R. Blake: Perception, Alfred A. Kopf, New York 1985, pag. 100,

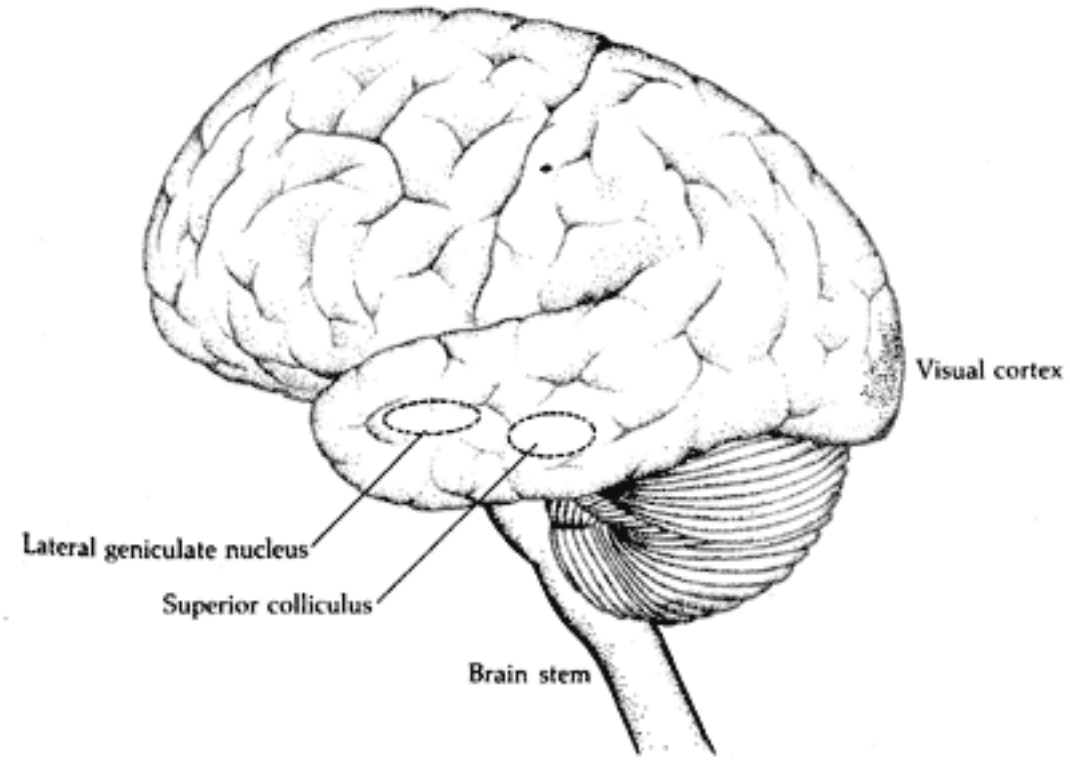

Fig. 2 - Superior colliculus - from R. Sekuler and R. Blake: Perception, Alfred A. Kopf, New York 1985, pag. 103

It is interesting to note that the human brain is structured so that there is only one area that can be truly called "audiovisual", namely, the "superior colliculus" (fig. 1 and 2). This organ is a phylogenetically older area than the visual cortex, and in certain more primitive animals the superior colliculus represents the entire brain. The superior colliculus is an area of our brain that receives input from both the ear and the eye and, because of this, its cells are called multisensory cells: "For example, if some multisensory cell responds to a light flash in the upper right portion of the visual field, that cell will respond to a sound only if it too comes from the same vicinity. Additionally, when visual and auditory inputs occur simultaneously, a multisensory cell responds more strongly than when either input occurs alone." [R. Sekuler and R. Blake: Perception, Alfred A. Kopf, New York 1985, pag. 104].

In "Dynamics" a visual object located in a certain position in space was thought of as emitting its sound from the very same location. A videoprojector, a big screen and four speakers located at the corners of the screen, were used to bring

about this identity of spatial localization. The speakers thus provided not only the usual stereo image (right-left), but also the top-bottom image, filling the area of the screen with a continuous acoustic signal. In order to achieve this, I digitized the sounds I had previouosly generated with FM machines into Csound files (Csound is a C and UNIX-based language written by Prof. Barry Vercoe, which was running on the VAX 11/70, VAXStation II and HP Bobcat workstations at the MIT Media Laboratory. It is a software package consisting of modules for audio synthesis, analysis, and processing). I then wrote two C programs that permit the input of a sound trajectory with a graphic tablet and stylus, and the use of the x and y tables which define this trajectory as control files for the sampled sounds.

Localization of sounds, however, cannot be as precise as in the visual domain, and is mainly a function of the spectral content of the sound [C. Dodge and T. Jerse: Computer Music, Schirmer Books, New York 1985, pag. 240-247]. This fact influenced the association of sizes and shapes to sounds, in the sense that the size of a certain shape was a function of the spatial extension of the corresponding sound, which was in turn a function of the spectral content of the sound itself. To be clearly perceived as localized in space, sounds have to feature some kind of noise component, and also partials above 7000 Hz [R. Erickson: Sound Structure in Music, University of California Press, Berkeley CA 1975, pag. 143]. This is especially true when vertical localization is considered, a task that our perception system does not accomplish as well as horizontal localization. The spectra of some of the chosen sounds did not contain a great deal of high frequencies, and therefore were rather vaguely defined in space; yet they had to be defined in space as specific visual objects with well-defined contours. Sometimes it became necessary to have the visual objects smoothly merge with others or with the background, in order to better simulate the effects of the sounds. A good visual example of shapes that melt with others can be found in the computer generated video "Ecology II: Float", by the Japanese artist Yoichiro Kawaguchi. It was then necessary to create a scale of sizes that could be perceived and understood by the listener. The sound that contained the highest perceivable frequency was the smallest (precise localization), while the sound with lowest frequency filled the screen, simulating its vague spatial localization. By means of compared and repeated judgement, I was able to determine a scale of sounds (from the lowest to the highest in frequency) that in turn corresponded to a scale of sizes.

4 Intensity

The final correspondence, and the simplest, was between the perceived intensities of the aural and visual events (loudness and brightness). This means that as loudness changed, following an envelope, so did the brightness of the corresponding shape. Eventually, when the visual object faded out, loudness reached zero. Although the principle is quite simple, it was not easy to relate one domain to the other, since neither of these parameters is linear. The method employed here was similar to the one used for localization. That is, I created a scale of sounds, from the loudest to the weakest, mapping the loudest as white (very bright), and the weakest as barely visible; this was done also taking into account the fact that we are not linearly sensitive to colors (for example, we are more sensitive to green than to blue). The entire process was empirical, but the tools available limited its flexibility. To establish more precise correspondences I would have needed immediate feedback , which would have been provided only by a software dealing with sounds and visual objects simultaneously.

5 Creating the Correspondences

The process of generating correspondences was carried out through the following steps:

- I conceived a sound that interested me. I found it easier to model visual objects after sounds rather than vice versa, because sounds have an extremely variable behavior over time, and one can have more control over the change of shapes than over the change of sounds, especially FM sounds.

- I sketched an outline of the temporal behavior of the timbre's main components, indicating the envelope and the sound's peaks.

- I imagined a changing shape that would match the behavior of each sound, and indicated its attributes (round, jagged, shiny and so on).

- When actually creating the 3-D model (with the program 3-dg, running on a Symbolics 3600 computer at the MIT Media Laboratory) I listened again to the sound and if necessary modified the previous idea (the sketch) so that it matched the sound. To produce the changing effect of the shapes reflecting the change in timbre, I used two different methods:

- I created two 3-D models (initial and final), that were then automatically and linearly interpolated by the animation program (in one case I used more than two models).

- I rotated the object along one or more axes. Since the shapes were irregular, they revealed other sides that had not been seen yet, behaving like a sound that, while changing, reveals unknown aspects of itself.

6 The Composition

While in the first stage I created visual objects from sounds, in the second one I preferred to invert the process, and use the visual language to create the animation, matching it with the music only afterwards. This procedure allowed me to organize the piece using visual language. Although I chose to use this compositional method to produce an audiovisual piece, it can also be useful for composing music with abstract sounds only. If the reverse process (creating the music composition first, and then matching the visual objects) had been used, the final product would have been quite different. Features such as spatial location, movement and speed, which are characteristic of the visual domain, would not have played a major role in the formulation of the music, as they did. Moreover, the process needed in order to associate a sound to a visual object (see section 2) forced me to think about the sounds themselves, to identify categories of timbres in great detail and visualize them, and led me to associate different sounds only because their corresponding visual objects were associated. Although some might argue against restricting sounds into categories, in this instance this process proved to be helpful. In fact, since the notion of timbre is quite loosely defined in terms of form, timbres had to be controlled through attributes related to perception rather than to acoustic data. Two important categories I used were those of harsh and soft timbres, visualized as jagged and rounded shapes respectively (fig. 3 and 4); other classes were shiny and opaque, visualized just as shiny and opaque objects. These categories were definetely subjective, and are not meant to be valid for everyone; I envision them as general families of aural

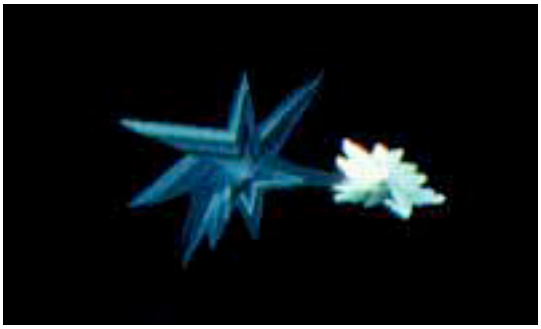

Fig. 3 - Harsh timbres

Fig. 4 - Soft timbres

and visual objects that may be comparable to the families of the orchestra. My idea of class can be related to Kandinsky's antithesis [W. Kandinsky: Concerning the Spiritual in Art, Dover Publications Inc., New York 1977, fig. 1-2].

The composition is divided into 6 major sections (fig. 5). The first one is a presentation of the nine audiovisual events, in succession. The events are shown one at a time, with cross-fading effects. The subsequent 3 sections constitute the body of the piece. Each of the sections is divided into 3, 3, and 2 episodes respectively.

1	2	3	4		5	6
intro	a b c s1	d e f s2	g h	s3	s2 s1	s1 + s2 + s3

Fig. 5 - Sections and episodes

In each episode, one or two audiovisual events perform a specific action. A final episode in each of the three sections constitutes the blending of the previous episodes of the same section. For example, in section 4 the final episode sums up the previous two episodes. In section 5, the previous final episodes are repeated. Notice that episode s3 belongs to both sections 4 and 5. The closing section is the sum of the 3 sums of each episode, s1, s2, and s3.

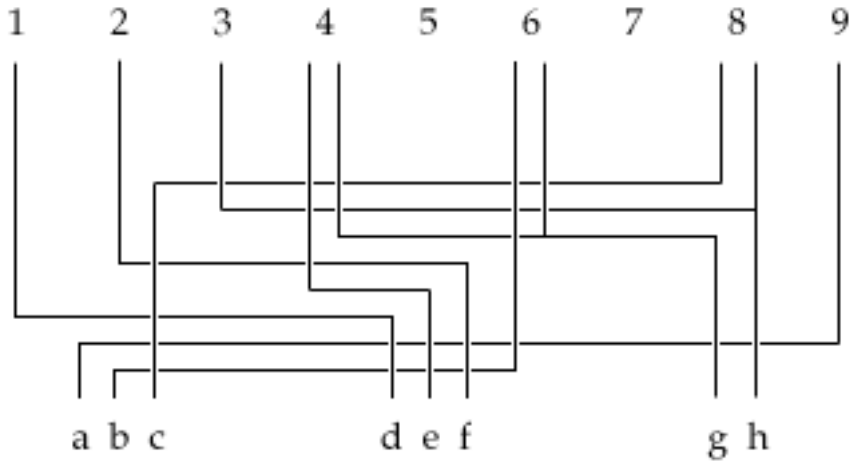

Fig. 6 - The composition scheme

I would like to briefly analyze a, b, c, d, e, f, g, h, since they represent the core of the piece. In these episodes, each object is spatially repeated (i.e. multiple copies of the same object are present). A, b, c are the sections of time, meaning that variations of time are considered, and the spatial positions are not ordered. In a, the object is shown 15 times, very fast, in succession, so that it never overlaps in time. B shows an object (and its copies) that is fades in and out, partially overlapping in time. In c, the copies of the object are perfectly synchronized. D, e, f are the episodes of space, meaning that variations of space are made.
D presents an object (and again its copies, all in synch) that is moving with a jaggy trajectory. E shows a completely static object (its shape changes, but not its position). In f a regular movement is used. Episodes g and h use 2 objects (and copies) at the same time. In these sections, different visual characteristics are opposed. The objects are all synchronized, and have a regular movement. In g, a round object is opposed to a spiny object. In h, an opaque object is opposed to a shiny object. Notice that objects 5 and 7 are used only at the beginning, in the intro section.

The organization of the piece was relatively simple. After the presentation of the audiovisual events, I created episodes that were in different ways coherent (coherence of time, space and opposition of attributes). Each of the final episodes combined the previous episodes of its section, creating a more complex audiovisual texture. The basic idea is in fact the construction of complexity,

Fig. 7 - Final sequence

Fig. 8 - Final sequence

starting from simple elements. Because of this, I repeated again the episodes s2 and s3. This way, after the presentation in which each event is seen and heard alone, and after the second part (sections 2, 3, and 4) in which 2 or 3 audiovisual objects contemporarily play different roles, I created a sequence in which several objects were considered at the same time. In other words, the piece becomes more and more dense. The final section, in which all the previous sums are again added, shows a very thick audiovisual texture, comparable to a musical "tutti".