Functional Specification of a Distributed and Mobile Architecture for Virtual Sound Space Systems

Stéphane Natkin, Florent Schaeffer
CEDRIC/ CNAM 292 rue St Martin 75141 Paris Cedex 03, France
natkin@cnam.fr, schaeffer@cnam.fr

1 ABSTRACT

This paper develops a functional analysis of the augmented reality system presented in [Natkin00]. It also presents the first elements of an experimental architecture and a real example where the system would be used.
The system is based on virtual sound reality : spectators are walking into a real space, indoor or outdoor, wearing headphones. They see the real space and at the same time hear a virtual sound space, homeomorphic to the real one. This means that there is a continuous function which maps any trajectory in the real space to a trajectory in the virtual space, thus determining which sound is heard along this trajectory.
The synthesis of the virtual sound along a trajectory may depend on many factors : speed of the spectator, past movements of the spectator, current or past position of others spectators, random events and so on. Moreover special rules or constraints will be added, considering the kind of application desired : quality of sound needed, maximum number of spectators using the system at the same time, interactions between spectators (or lack of), complexity of the sound synthesis…
Possible fields of application include art installations, personal guided visits through a space, audio help to drivers in a reduced visibility area, audio help in the maintenance of industrial plants….
In this paper we present a functional analysis of the system and a general distributed architecture using both ground and mobile computers. We address in details the localization, transmission and spatialization functions and the time-space-bandwidth complexity of these functions. This leads to a classification of the possible distributed designs according to application constraints. Then we consider an art installation application "The Persian Carpet", proposed by the composer Cecile Le Prado. The specific aesthetic constraints lead to a particular solution of the general design proposed.

2 GOALS

The goal of the system described in this paper is to play on a perceptual paradox : a set of spectators is walking through a real space, seeing this space and hearing the sound of a virtual space through headphones. The topology of the visual and audio space can be arbitrary as long as both spaces are homeomorph. Roughly speaking each trajectory of a spectator in the real space must be mapped by a continuous application into a trajectory in the virtual sound space. In the simplest case the sound space is determined, in a more complex case it can depend on random events, physical data (such as the visual space brightness), memory of past events (like the spectator or other spectators trajectories) or even use different acoustic laws (for example linear attenuation of sound instead of a logarithmic one)... The only constraint is that the sound space must be defined in each reachable point.
We also decide, as a first hypothesis, that the determination of the sound in a point of the virtual space does not depend on the sound in the corresponding point of the real space (at least not in real time) in order to avoid a real time computation of sounds recorded in the real space.
Such a system has numerous applications. Our initial interest was suggested by a composer (Cecile Le Prado) for a sound installation, but it can also be used for guided visits of museums, to provide help to drive through a reduced visibility area, for augmented reality help for industrial supervision and cooperative work, for augmented reality games…

3 GENERAL CONSIDERATIONS

In [Natkin00], the application needs were discussed, in order to draw some conclusions regarding the computation complexity of the system.
The factors influencing the complexity are : the number of users in the system, the complexity of the virtual space (number of sound sources, moving sources…) and the quality of the spatialization (especially the reverberation model). But if the virtual sound space can be divided in several airtight areas, each area can be considered as a separate and simpler system.
These factors depend on the goal of the system which ranges from one person and poor sound quality up to dozens of spectators (for a guided visit) and CD quality (in the case of an art installation).
Therefore we will define in this article a scalable and configurable system in order to cope with all the possible kinds of applications.

4 FUNCTIONNAL SPECIFICATIONS

4.1 Introduction

When a spectator enters the installation, he picks up a headphone and is localized at a given reference position. The system creates a new entry for him and allocates the needed computer resources to be able to cope with the corresponding computation.
The figure 3 is a rough SADT specification of the system functions in steady state behavior with the data flows between them and the external inputs.

The functions to be computed are:
- The coordinate determination
- The management of the memory of the process
- The localization of the moving sound sources and the determination of the set of sources which can be heard by the spectator (Cinematic computation and zone determination)
- The synthesis of the sound for each source
- The spatialization of the sound

In this section we refine the analysis of each function

4.2 Coordinate determination

Each spectator uses a wireless headphone with a position captor.

The coordinate system includes:
1. The tracker identification I, as each signal sent to the system by a given captor must be referenced by a logical address to allow the individual and continuous tracking of the spectator.
2. The position of the spectator s in the real space X(s,t)
3. And according to the application:
   - The relative position of sound sources to the head of the spectator is not used. In this case the captor does not give any additional information :
     $$\Phi(s,t)=\{\varnothing\}$$
   - The system uses a spatialization of the sound on the plane. The captor must give the relative position of the spectator's ears in the plane :
     $$\Phi(s,t)=\{\theta(s,t)\}$$
   - The system uses a 3D spatialization. Then the slope and the elevation of the head must also be detected, leading to the determination of three angles :
     $$\Phi(s,t)=\{\theta(s,t),\rho(s,t),\varphi(s,t)\}\ .$$

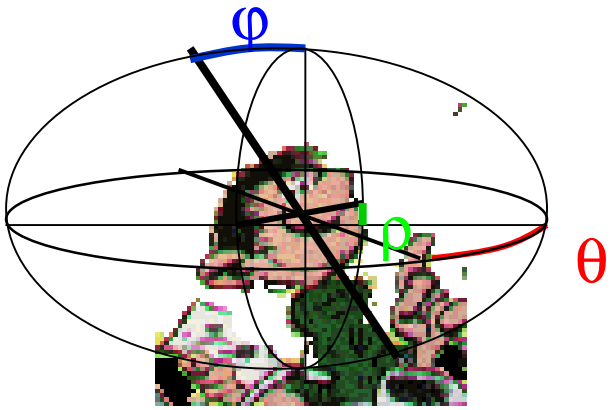


Figure 1: Head coordinate system

The coordinate determination is essential to the generation of the artificial sound. But what precision is required ?
We will assume that it is the smallest movement that would generate a difference in the sound heard by a spectator. According to [Blauert], this "localization blur" depends on the intensity and frequency of the sound, on the plane of movement (θ, ρ or φ) and on the position of the sound source at the beginning of the movement (front, side or rear). Following Blauert's advice, we will consider that the minimum perceptible change in optimum conditions should be our constraint : it is about 1° for a change of angle (head movement) and, in the case of a close sound source, 25 cm (about one step).
Of course, these constraints can be relaxed, depending on the application.

4.3 Process memory management

If the evolution of the sound in the virtual space is memoryless this function is empty; but it is generally not the case. The memory of the system can then be split in two classes of state variables : global memory and personal memory.

Some state variables are identical for all spectators. They determine the global dynamic of the sound field. We call global memory this set of variables and denote M(t) its value at time t. For example, if sound sources are moving according to a deterministic or random process independent of the spectator location, all the instantaneous cinematic parameters (position and speed for example) of the sound sources must be stored in M(t). It is also possible to allow a spectator s to leave a message or a trace, which will be used in the subsequent sound generation. Each time s leaves a trace tr, tr is in M(t). Moreover all the dynamic parameters which are needed to synthesize and spatialize the sound sources for all the spectators are in M(t). It can be pointers in Midi parameter tables, time codes in audio files…
Other state variables are used only to compute the sound heard by a given spectator s. We call personal memory this set of variables and denote m(t,s) its value at time t.
For example if a given sequence of sound has to be started when the spectator s enters in a region of the real space, the date of entry is in the memory in m(t,s). This is an important feature for a guided museum application. It is also possible to allow a spectator s to leave a trace which will be used to compute the sound for another given spectator s'. In this case when a spectator enters in the system he must give his name s. The couple (I,s) must be in m(t,s). And each time s leaves a trace tr for s' the triple (s,s',tr) is in m(t,s). All the dynamic parameters which are only used to synthesize and spatialize the sound sources for s are in m(t,s).

It is of course impossible to give a general specification of the memory management function, one may think about the state variables of an arcade game applied to a sound

installation. The important feature in terms of architecture is the relative space and time complexity of the two sub-functions: manage the global memory and manage the personal memory.

4.4 Cinematic computation and zone determination

This function must compute all the current positions of the moving sources and then determine the set of sound sources which can be heard by s at time t : so(s,t,X(s,t)).
Let's analyze more precisely how we can use this result.
We denote by R the real space, V the virtual space, f the homeomorphism from R to V (a bijection which maps any continuous function in R into a continuous function in V).
Let $v(s,t,X(s,t)) \in V$ be the smallest part of the virtual space such that the sound heard by all spectators in $r(s,t,X(s,t))=f^{-1}(v(s,t,X(s,t))) \in R$ can be computed from the sound produced by so(s,t,X(s,t)). Let n(s,t,X(s,t)) be the number of spectators in r(s,t,X(s,t)).
If n(s,t,X(s,t)) is greater than one (the spectator s) then it means that several spectators hear the same set of sound sources so(s,t,X(s,t)). In this case, the synthesis of the sound for each source and a part of the spatialization of the sound (see this section) are the same for all the spectators in r(s,t,X(s,t)) and the computation results can be shared.
This means that it is possible to divide the computation needed into a common part done only once and a personal part using the common result to complete the calculation.
However, it will not always be possible : if several spectators can be at the same time in one region of the real space then $n(s,t,X(s,t)) \geq 1$. But as a counter example consider a system designed to help drivers in a low visibility zone : in a given part of the real space there is generally less than one driver, and each driver needs personal instructions. Hence n(s,t,X(s,t))=1. In this case, all the process memory is in m(s,t) and not in M(t) and the sound for each spectator is personal; no computation can be shared.

4.5 Sound Synthesis

The sound of each audio source can be produced either by pure synthesis or by picking samples in audio files. In both cases the corresponding stream can be modified using real time audio effects. But it is difficult to give a more precise specification of the sound synthesis function without considering a particular application. Two extreme cases can help to understand the diversity of this function. In the situation of an art installation like the one described later in this paper, the sound is produced by the real time modification of audio streams stored in a multi-tracks device. For each track a CD quality of sound is required. In the "help for drivers" example the sound is composed by simple messages such as "turn right", "take care, obstacle in front at 200 meters"… These messages can be either created by a standard voice synthesizer or by mixing word samples and this application needs only the sound quality of a standard phone.

4.6 Spatialization

The spatialization of the sound is divided in two parts : the directional spatialization (which allows an auditor to say that the sound "comes from behind", for example) and a non-directional spatialization (the sound is "wet"). The latter is called the "room effect" and does not depend on the position of the sound sources.
By combining the zone determination function described earlier and the "directionless" property of the room effect, it is possible to define an efficient way of dividing the computation of the spatialization :

1. Determination of a sound zone
2. Sound synthesis of all the sources which spectators can hear in this zone
3. Spatialization of the sound for a virtual spectator located in the center of the zone (figure 2)
4. Localization of the directional part of this sound for each spectator in the zone

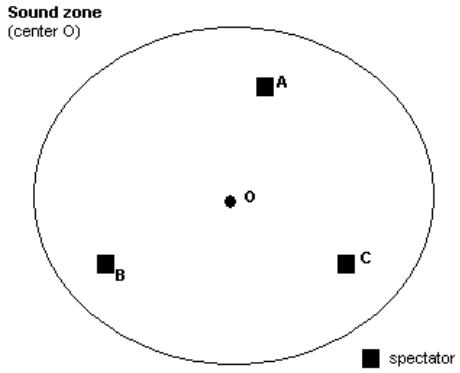
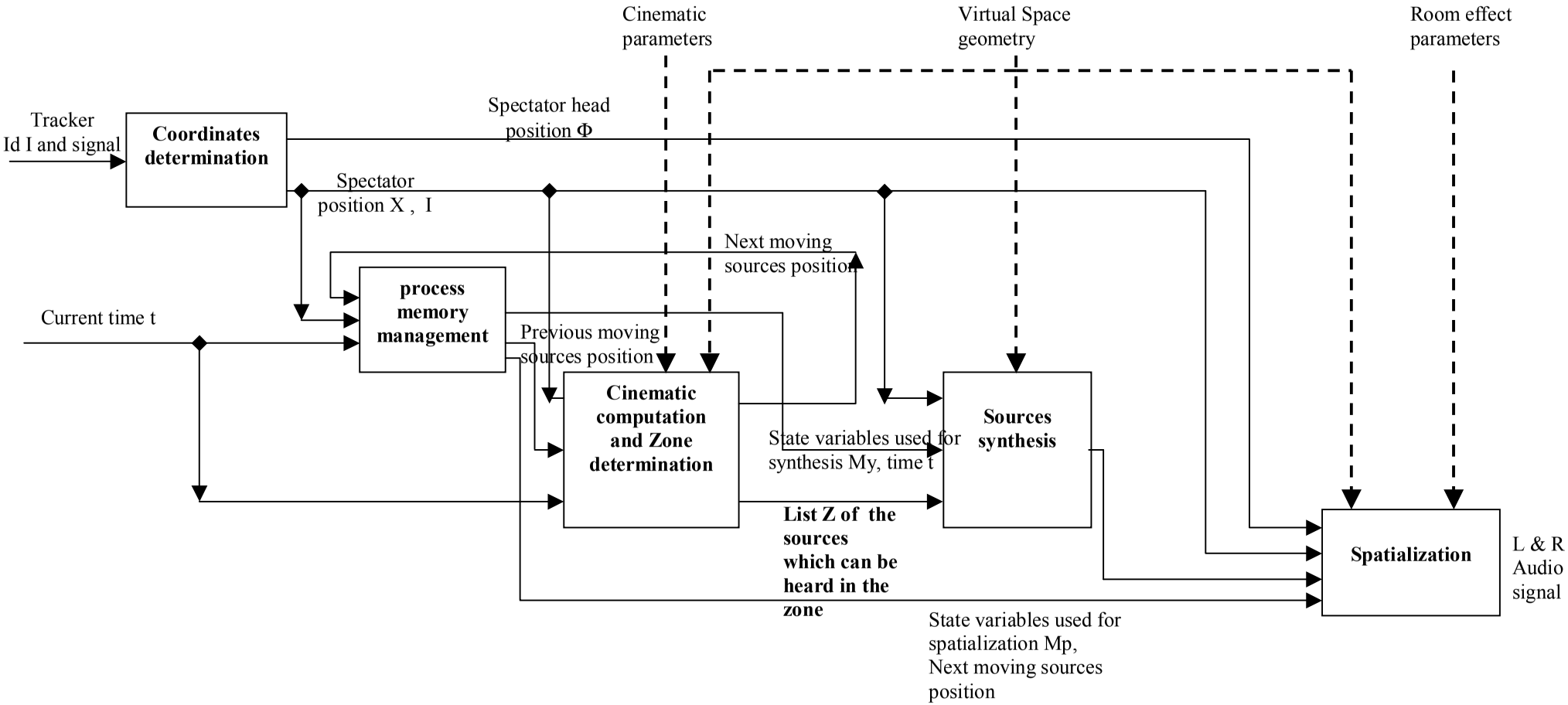


Figure 2 : Sound zone for several spectators

When there is only one spectator in the zone, the spatialization (step 3) is done directly for his position. When several spectators are in the zone, the computation needed in steps 2 and 3 is done only once and the results shared between them. Then the directional part of the spatialized sound is localized (ie modified accordingly to the position of the spectator in the sound zone).

Figure 3 : SADT specification of the system



5 ARCHITECTURE

5.1 Real time constraints

When a spectator is moving in the real space, he must "move" at the same speed in the virtual space. This means that the sound must not be heard "too late" (because his position in the virtual space decides what sound he hears). That is to say that the latency must not be too important between a movement of the spectator and a change in the sound he hears. And that despite the fact that the system must locate the spectator, compute the sound and transmit it. We assume that a maximum latency of 1 ms would be acceptable (only trained listeners can pick it up).

5.2 Logical description of the virtual sound space

This important aspect of the system is the object of a separate work, conducted by Alexandre Topol [Topol01]. We have decided to extend the VRML sound description to represent the virtual sound spaces in a form closely related to virtual visual 3D spaces. It gives us a simple and widely used way to describe a scene, using trees and nodes. This ensures that tools to create virtual sound spaces will be available (we already have a modified VRML viewer able to manipulate these new sound extensions).
Furthermore this will allow us to later use our system for fully immersive applications, taking advantage of a full VRML description of scenes including both sound and image.

5.3 Possibilities for a distributed system

The spatialization leads to two mono audio signals for each spectator (left and right). The computational cost of the function is not the only problem to consider, there is also the transmission of the signal to the spectator. Several solutions may be considered, depending on whether the computation is done by a central unit, mobile units carried by the spectators or a combination of both. The following table explains the main possibilities.

Central computation unit *(at system level)*	**Local computation unit** *(at spectator level)*	**Transmission needs**
Spatialization of left and right signals for each spectator	No local spatialization	Two channels for each spectator
Spatialization of an ambisonic signal for each spectator	Left and right ear differentiation computed from the ambisonic signal	Four channels for each spectator
Spatialization of an ambisonic signal for a zone	Localization of the spatialized sound and left-right differentiation	Four channels for each sound zone
No global spatialization	Spatialization of left and right signals	One channel for each sound source and one data channel

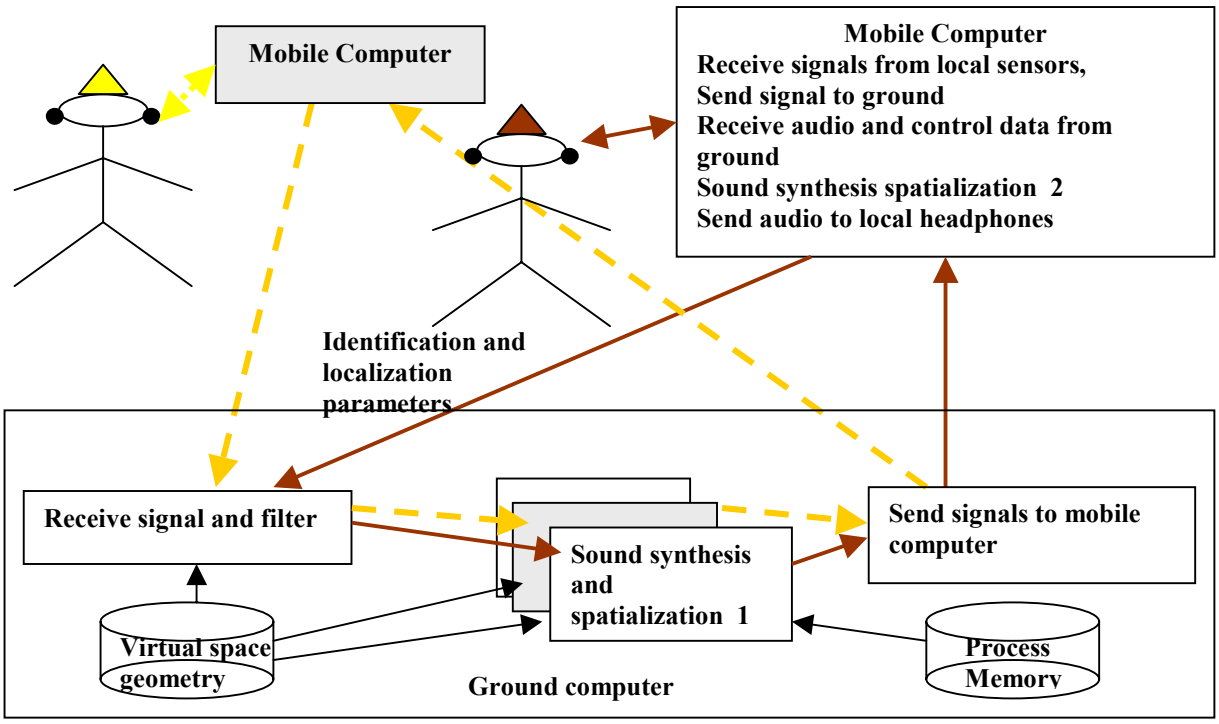
These channels are logical, they represent the "streams of data" that need to be transmitted. The actual number of physical channels needed is problem of bandwidth and multiplexion, because of the different sound qualities. It is discussed in the next section.

5.4 Transmission needs

The quality of the sound needed by the application is the most important factor to evaluate the needs regarding the transmission of the sound to the spectator.
An art installation would require a good quality of sound (say 16 bits at 22kHz) that means a rate of 43 kb/s for a mono signal.
An audio assistance system would require only voice quality (8 bits at 8 kHz) which means a rate of 7 kb/s for a mono signal.
Depending on the solution chosen, the number of logical channels needed to transmit the information is either constant (last two cases) or increases linearly with the number of spectators.
The choice of a technology to transmit the data is a part of the problem which has not been investigated yet.

5.5 Experimental architecture

We are working on a first prototype of architecture based on the IRCAM FTS/Spat [Dechelle 98] system and following the functional design presented in the figure 3. This system works on a standard Linux configuration. The space and the spectator's position are simulated on a simple graphic interface on a PC. One PC is used to simulate one spectator and up to ten spectators can be at the same time in the installation. The location of each spectator is sent to a pool of Linux stations, each processor is able to cope with one to three spectators. This first prototype will be used to implement a description of the virtual sound space, experiment some applications needs and develop the sound spatialization function.
Since in the current FTS implantation the whole sound synthesis process is sequential, one of the first tasks is to split the cinematic analysis from the room effect and sound synthesis.
Then a mobile distributed architecture using laptops and wireless communication will be developed, allowing a more realistic experimentation with one or two spectators moving in a room with helmets and headtrackers.



6 A REAL EXAMPLE : THE MAGIC CARPET

The patterns on a Persian carpet are a symbolic representation of the world. In this art installation proposed by Cécile Le Prado, the "magic carpet" is the floor of a room : it represents the world and spectators walk on it, travelling from place to place. They hear different sounds (real sounds recorded all around the world and modified by computer or artificial sounds) depending on their movements and the position of the other spectators. Each spectator can leave a sound trace and the virtual sound space evolves with the different visits, following predefined rules.
This application, installed in a single room, would require CD quality sound and allow up to ten spectators at the same time.

7 CONCLUSION

In this paper we continue the work started in [Natkin 00] : the definition of a class of sound systems which can be used in various areas, from music installation to industrial augmented reality systems.
We now have a functional specification of the system (in its general form) and a particular application with it specific needs. The next step will be to develop an experimental architecture able to meet the requirements of Cécile Le Prado's project but also scalable, in order to move towards a general workable solution.
Our prototype, using a pool of Linux computers and later laptops, will use a description of the virtual space in VRML with sound extensions and the IRCAM/Spat software. It is currently under development.

8 REFERENCES

Blauert J;, *Spatial Hearing*, MIT Press, Cambridge, MA, 1983.

Dechelle, F., Riccardo Borghesi, M., Maggi E., Rovan B. and Schnell, N.,*Latest evolutions of the FTS real-time engine: typing, scoping, threading, compiling*, International Computer Music Conference, October 1998.

Dechelle, F., Riccardo Borghesi, M., Maggi E., Rovan B. and Schnell, N., *jMax: a new JAVA-based editing and control system for real-time musical applications*, International Computer Music Conference, October 1998.

Natkin S., *Mapping a virtual sound space into a real visual space*, International Computer Music Conference, Berlin, September 2000.

Topol A., *Enhancing sound description in VRML*, work in progress, 2001.