GESTURAL AUDITORY AND VISUAL INTERACTIVE PLATFORM

B. Caramiaux, S. Fdili Alaoui†, T. Bouchara‡, G. Parseihian§ M. Rébillat¶*

LIMSI-CNRS, Ircam-CNRS, LMS-École Polytechnique

caramiau@ircam.fr

ABSTRACT

This paper introduces GAVIP, an interactive and immersive platform allowing for audio-visual virtual objects to be controlled in real-time by physical gestures and with a high degree of inter-modal coherency. The focus is particularly put on two scenarios exploring the interaction between a user and the audio, visual, and spatial synthesis of a virtual world. This platform can be seen as an extended virtual musical instrument that allows an interaction with three modalities: the audio, visual and spatial modality. Inter-modal coherency is thus of particular importance in this context. Possibilities and limitations offered by the two developed scenarios are discussed and future work presented.

1. INTRODUCTION

This paper introduces GAVIP (Gestural Auditory and Visual Interactive Platform), an interactive and immersive platform allowing for audio-graphical virtual objects to be controlled in real-time by physical gestures. GAVIP is based on a *control unit* driven by *gesture processing* monitoring the behaviour of the virtual objects. A global *control* of the virtual world ensures the *inter-modal coherency* of the proposed environment. The SMART-I^2 audio-graphical rendering engine achieves the *spatialization* of the virtual objects in a 3D audio-graphical scene [1]. The *spatialization* increases the coherence and thus the user's feeling to be «present» in the virtual scene [2]. *Gesture processing* allows for a wide range of natural *interaction* enhanced by the *presence* sensation [3].

GAVIP can be either considered as a sound installation, a platform of development for experimental protocol or an *extended musical virtual instrument*. By «extended», we mean that the created virtual musical instruments do not only allow an interaction with the audio modality, but an interaction with three modalities: audio, graphic and spatial. Declinations of GAVIP mainly depend on scenarios that are implemented. In this paper, focus is particularly put on two scenarios exploring the interaction between a user and the audio, graphic, and spatial synthesis of a virtual world where the interaction is thought in the sense of virtual musical instrument design. Previous works have dealt with the interaction between gesture and sound for the design of virtual musical instruments [4]. Early works have led to low interaction expressivity caused by a direct mapping between gesture parameters and sound synthesis engine parameters. Extensions have led to take into account more expressive interactions by considering higher level descriptors or mapping based on physical models [5]. Progressively, virtual musical instrument became more linked to the sound perception and immersion. This has led to consider the spatial diffusion of sounds produced by such instruments [6], thus changing the terminology from *virtual musical instrument* to *sound installation* in which interactions are multimodal [7]. In this paper, our contribution is to propose a platform that combines sound spatialization and audio-graphical systems. To enhance interaction, our platform aims to guarantee inter-modal coherency (see Sec. 2.2).

The paper is structured as follows. In the section 2 we present the concept of GAVIP (the general platform for designing new extended musical instruments). Section 3 describes a first scenario based on a simple interaction approach. To increase the degree of inter-modal coherency, a second scenario was developed based on a physical model. It is described in Section 4.

2. THE GAVIP CONCEPT

2.1. Concept

The goal of GAVIP is the conception and the exploitation of an immersive, interactive and multimodal platform. Such a platform is built as a support for different scenarios. These scenarios will typically plunge the participants into a virtual environment populated of several audio-graphical entities among which they can freely evolve. Participants can interact in real-time with these entities through their gestures. Gesture analysis is used as a natural way to control the behaviour of audio-graphical entities and their spatialization in a 3D scene. Thus, a great effort is made to design a general architecture that allows for wide interaction strategies.

2.2. Interaction and Inter-modal coherency

In this paper, *interaction* means the bidirectional relationships between human gestures and virtual elements. Our hypothesis is that perception of interaction is enhanced by the coherency between audio and graphical objects (the so-called *inter-modal coherency*). Without being stated, this idea is shared across complementary research fields like Human Computer Interactions (HCI) [8], New Interfaces for Musical Expression (NIME) [9], or Virtual Reality (VR) [2]. By *audio-graphical object* we mean a virtual object with consistent audio and graphical components. Inter-modal coherency of the *audio* and *graphical* components of the virtual object (semantic and spatial) is required to integrate the audio and graphical streams into one unique percept [10]. However the classical approach is to synthesize separately the *audio* and *graphical* streams. Mismatches between these two information streams typically exist and are thus perceived. This can significantly degrade the quality of the audio-graphical rendering and the resulting interaction. To enhance *presence* and *interaction*, *Inter-modal coherency* has to be ensured.

* Ircam-CNRS

† LIMSI-CNRS, Ircam-CNRS

‡ LIMSI-CNRS, Univ. Paris 11

§ LIMSI-CNRS, Univ. Paris 11

¶ LIMSI-CNRS, LMS-École Polytechnique

2.3. Architecture

In order to achieve interaction with consistent audio-graphical objects in a virtual 3D scene, we designed a general virtual space where the audio and graphical parts of each entities are governed by the same control unit. Concretely, one unique control model is mapped with both modalities (the sound synthesis and the graphical synthesis) and the virtual space (spatialization) according to three consistent intra-modal mapping strategies (see Fig. 1). This allows the immersed user to influence through his/her gesture the control model itself and, consequently, the audio-graphical objects behaviour and their position in the virtual space. The consequence of an action of the user is by nature multimodal (audio/graphical/spatial) and a careful design of the mapping strategies ensures, by construction, the preservation of a strong *inter-modal coherency*.

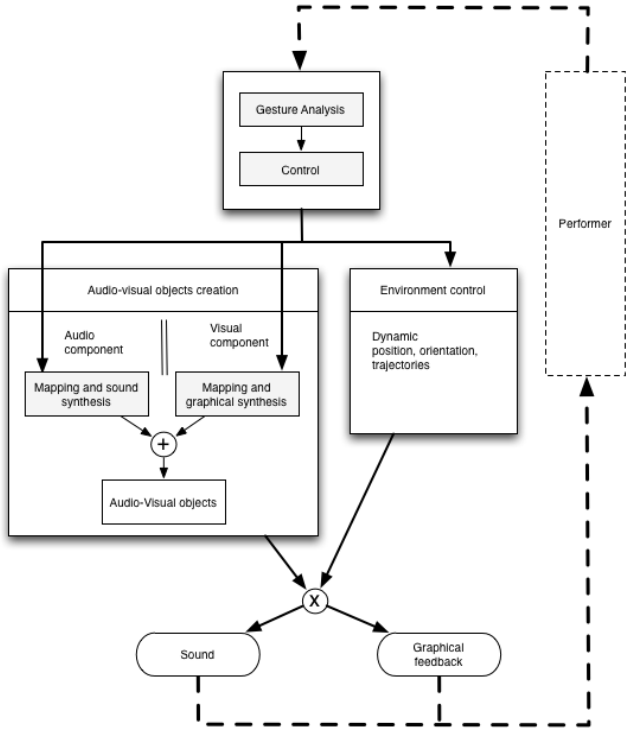

Figure 1: Overview of the general architecture of the GAVIP

2.4. Audio-graphical spatial rendering

The SMART-I^2 (Spatial Multi-user Audio-graphical Real-Time Interactive Interface) [1] provides an immersive, real-time and consistent audio-graphical spatial rendering. The spatial sound rendering is achieved by means of *Wave Field Synthesis* (WFS) [11]. The WFS technology physically synthesizes the sound field corresponding to one or several virtual audio sources (up to 16 sources) in a large rendering area. The 3D graphical rendering is made using passive tracked stereoscopy. In the SMART-I^2 system, front-projection screens and loudspeakers are integrated together to form large multi-channel loudspeakers also called *Large Multi-Actuator Panels* (LaMAPs). The rendering screens consist of two LaMAPs (2 m $\times$ 2.6 m with each supporting 12 loudpeakers) forming a corner (see Fig. 2). Overall, the SMART-I^2 allows to generate 16 independent virtual objects. Each virtual object is a sound source and a graphical object. See [12] for more informations regarding audio-visual immersive systems using WFS.

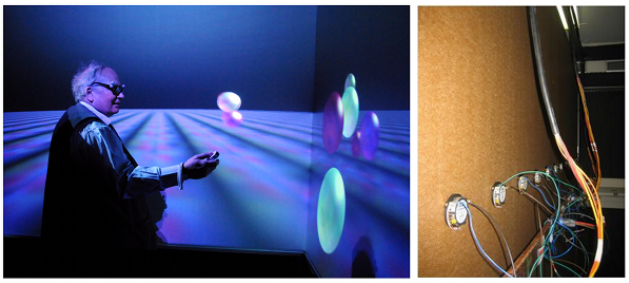

Figure 2: Front (left) and back (right) views of the SMART-I^2 audio-graphical rendering system. The left picture shows a participant using the early approach of interaction.

3. EARLY APPROACH

This approach was presented as a demo during the european meeting European VR-EVE 2010 that took place at the LIMSI-CNRS, Orsay, France.

3.1. Overview

In this first approach, each graphical object was a bubble and each sound source was a single sound sample taken from a given database (see Fig. 2). The challenging task was to control the set of 16 independent sound sources generated by the SMART-I^2. We chose to synthesize sounds of water drops. Two modes of interaction correspond to either a sequential or a single triggering of sound samples. Since a simple one-to-one mapping was impossible, we chose to have automatic controls of some of the sound parameters. We separated the *sound controls* (sample volume, sample playback speed, period between two successive sounds) from the *spatial controls* (voice number, distance, azimuth). Each of these parameters was controlled by a specific probability density function from which each parameter's value was drawn. This allowed to control the parameters' variations and to have less predictable behavior.

3.2. Gesture Control

The gesture control process is twofold. First, an Optitrack motion capture system composed of six circular cameras is used to track the user motion and adapts the graphic feedback to his/her spatial position. The calibration of the cameras is performed with the help of the Optitrack Rigid Bodies software. This software locates the markers dispatched on the user stereoscopic glasses and sends their 3D positions and orientation via a VRPN communication protocol.

Second the sound control is achieved by means of a WiiMote controller. Hence, the gyroscopes are used for driving azimuth and distances: the user can choose to be either closer or further from the sound sources and put the sources either on the right side or on the left side. Accelerometers are used for triggering control. We refer the reader to the previous section (Sec. 3.1) for the presentation of the two triggering modes. In the first mode (triggering mode), the user can use the controller to trigger percussive sound of water drop. In the second mode (sequential triggering), regular hits with the controller at a certain frequency trigger sequentially

played sounds with the same frequency The frequency range allowed for synthesizing sound textures from separate water drops to heavy rains.

3.3. Limitations

The gesture control of audio-graphical objects proposed by this approach is straightforward. This enables a direct understanding of the general behavior by the user that can experience the scenario in a ludic way. However, this approach features several limitations. First, the global behavior is mostly controlled by probability density functions and the user's degree of control is limited. Hence the user can not really consider the interface as a virtual musical instrument as its control in sound production and positioning in space is limited. Second, the audio and graphical elements have a basic discontinuous behavior that consists in *appearing – disappearing* which does not simulate sound texture.

4. CURRENT APPROACH

To overcome the limitations of the first scenario, focus has been put on the control aspects of the platform for two main reasons. First, we aim at providing a smoother control of sound synthesis and sound spatialization. Second, in our quest for *inter-modal coherency*, we were looking for a more coherent relationship between audio and graphical parts of the virtual entities.

4.1. Overview

We refer the reader to the general architecture of GAVIP depicted in Fig. 1. The control process in the early approach was based on accelerometer–gyroscope sensor data analysis and simply linked triggering and orientation messages between gesture and audio modalities (also called *direct mapping*). Here we propose to define a more complex control unit based on a physical model simulating the interactions between elements in the scene. Audio elements are sound grains obtained by segmenting recorded environmental sounds. Since we aim to recreate sound textures we basically segment environmental textures like rains, winds, etc. Graphical sources are deformable sphere whose dynamics was inspired by magma lamp. In the following we mainly focus on audio control and rendering.

4.2. Physical Model

Here we define a physical model governing the dynamic of N abstract spherical particles in a 2D space. These particles are in a mutual interaction and in interaction with their environment. A physical model of N punctual masses linked together with springs governs their movements in this environment. Let us denote $\vec{a}_n(t)$ the acceleration of particle n at t, each element n has a mass m_n and follows the Newton's second law of motion defined by:

$$m_n \vec{a}_n(t) = \sum_{k \neq n} \vec{F}_{k \to n}(t) + \vec{F}_{\text{centre} \to n}(t) + \vec{F}_{\text{ext} \to n}(t)$$

Where at time t, $\sum_{k \neq n} \vec{F}_{k \to n}(t)$ is the sum of the forces applied from the other particles on the particle n, $\vec{F}_{\text{centre} \to n}(t)$ is the force applied from the centre on the particle n, and $\vec{F}_{\text{ext} \to n}(t)$ are the external forces exerted on n. Within the framework of this physical model, the forces are defined as follows:

- $\vec{F}_{k \to n}(t)$ accounts for: the elastic force (characterized by a parameter α_k), the viscous friction (characterized by μ_k), the electrostatic attraction (characterized by the product of their electrostatic load $q_n q_k$)
- $\vec{F}_{\text{centre} \to n}(t)$ accounts for: the elastic attraction (characterized by K) and the central viscous friction (characterized by η)
- $\vec{F}_{\text{ext} \to n}(t)$ accounts for: the global rotation (characterized by β_n) and a magnetic field that is linked to the electrostatic load of the particle q_n.

This physical model allows to generate the relative movements of masses from a given initial state. At each time step t, the physical model returns the particles' index together with their Cartesian coordinates.

4.3. Sound Synthesis

The sound synthesis is based on the CataRT software developed by Diemo Schwarz at Ircam [13]. The software takes a set of sounds and segments them given a grain size (the default grain size is 242ms) and/or a segmentation algorithm (for instance based on onset detection). Since we work with sound textures with no onset, we choose to segment the input audio stream every 242ms. CataRT analyzes each grain by computing a set of audio descriptors. The resulting grain representation is a vector whose elements are the descriptors' mean values. CataRT then visualizes those grains in a 2D space (also called descriptor space) where the dimensions are chosen by the user among the available descriptors, for instance the *loudness* along the x-axis and the *spectral centroid* along the y-axis. Depending on an input path on this 2D space, sound grains are selected: at each time step t, the selected grain (i.e., played) is the closest to the current position x_t, y_t in the path in terms of a Mahalanobis distance. More precisely, this distance is computed between x_t, y_t and the grain descriptors mean values that correspond to the 2D space axis (e.g., loudness and spectral centroid). Note that a path in the descriptor space can be given either by using a gesture interface such as a mouse or by an input sound described in the same descriptor space. The latter case is sometimes called *mosaicing*.

Here, the sound grains are selected by the abstract particles' positions defined previously. To that end, we match both 2D representation used for the physical model and the sound corpus. Fig. 3 depicts the process. The sound corpus remains unchanged but the output of the physical model (which consists in each particle index and Cartesian coordinates) is scaled to fit the sound corpus dimensions. Hence we have two layers in the same Cartesian space (see the image at the middle of Fig. 3). An abstract particle position selects the closest grain (according to a certain neighborhood radius) at each time step. Concretely, our physical model is composed of 16 particles navigating in the sound corpus and selecting 16 grains (i.e., sound sources). Finally, this virtual plane is the sound field produced by the WFS.

4.4. Gesture control

The user's head is still tracked using an Optitrack motion capture system that allows to adapt the graphic feedback to his/her spatial position. Then, the gesture control of the virtual environment and objects corresponds to the possibilities of action and influence that a participant will have on them. One of the main challenges is to offer the participant an important power of expression in his control of the environment and the virtual objects.

At this step, the mapping between gesture analysis outputs and physical model parameters is not implemented but the global con-

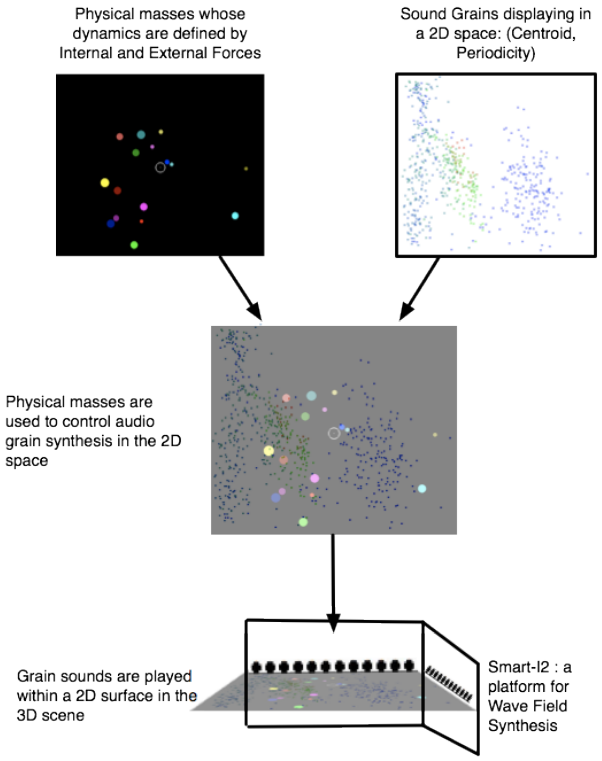

Figure 3: Sound Synthesis. Physical model environment and the sound corpus environment are matched and assigned to the sound field produced by the WFS.

cept is as follows. Two main gesture parameters will be computed (periodicity and energy) and the mapping is one-to-many. A periodic gesture with low energy will involve a stable behavior of the system with low changes in its parameters and high viscosity, and high magnetic rotation forces. A periodic gesture with high energy will decrease the viscosity and the elastic force between particles leading to a set of particles with higher quantity of motion. Finally, a non-periodic gesture with high energy will produce a highly entropic and unstable behaviour of the physical system. Hence, energy will be computed on data coming from the 3D accelerometers and periodicity on the data coming from the 3D gyroscope (using an autocorrelation criterion for example).

4.5. Implementation

A first version is already implemented in the Max/MSP real-time programming environment. A demonstration video of the phyiscal model controlling CataRT is available online[1]. A smaller version using binaural sound spatialization and simple 3D graphical rendering will be shown during the conference.

5. CONCLUSION

In this paper, a platform for the conception and exploitation of immersive, interactive and multimodal scenario were presented. An effort has been made to define a modular architecture that allows for various interesting interaction designs with respect to the coherency between audio and graphical objects. Two concrete scenarios were presented. The first one made use of simple triggering and orientation interactions between gesture and audio synthesis and spatialization. The second scenario proposed a general physical model that governs the behavior of audio-graphical objects in the scene. Gestures analysis controls the physical model. This new approach offered better coherency between audio and visuals as well as a smoother gesture control of them.

Future works will consist in (1) completing the implementation of the gestural control of the physical model and (2) designing perceptual experiments to evaluate the platform.

[1] `http://imtr.ircam.fr/imtr/GAVIP`

6. ACKNOWLEDGMENTS

We are grateful to both LIMSI-CNRS and UMR STMS Ircam-CNRS for the lending of equipment. We also would like to thank Matthieu Courgeon for his contribution and technical assistance on the 3D graphical programming.

7. REFERENCES

[1] M. Rébillat, E. Corteel, and B.F. Katz, "Smart-i^2: "spatial multi-user audio-visual real time interactive interface", a broadcast application context," in *Proceedings of the IEEE 3D-TV conference*, 2009.

[2] K. Bormann, "Presence and the utility of audio spatialization," *Presence: Teleoperators & Virtual Environments*, vol. 14, no. 3, pp. 278–297, 2005.

[3] M.J. Schuemie, P. Van Der Straaten, M. Krijn, and C.A.P.G. Van Der Mast, "Research on presence in virtual reality: A survey," *CyberPsychology & Behavior*, vol. 4, no. 2, pp. 183–201, 2001.

[4] A. Hunt, M.M. Wanderley, and M. Paradis, "The importance of parameter mapping in electronic instrument design," in *Proceedings of the 2002 conference on New interfaces for musical expression*. National University of Singapore, 2002, pp. 1–6.

[5] M.M. Wanderley and P. Depalle, "Gestural control of sound synthesis," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 632–644, 2004.

[6] G. Leslie, B. Zamborlin, P. Jodlowski, and N. Schnell, "Grainstick: A collaborative, interactive sound installation," in *Proceedings of the International Computer Music Conference (ICMC)*, 2010.

[7] A. Camurri, "Multimodal interfaces for expressive sound control," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFX-04), Naples, Italy*, 2004.

[8] R. Vertegaal and B. Eaglestone, "Looking for sound?: selling perceptual space in hierarchically nested boxes," in *CHI 98 conference summary on Human factors in computing systems*. ACM, 1998, pp. 295–296.

[9] X. Rodet, J.P. Lambert, R. Cahen, T. Gaudy, F. Guedy, F. Gosselin, and P. Mobuchon, "Study of haptic and visual interaction for sound and music control in the phase project," in *Proceedings of the 2005 conference on New interfaces for musical expression*. National University of Singapore, 2005, pp. 109–114.

[10] C. Spence, "Audiovisual multisensory integration," *Acoustical science and technology*, vol. 28, no. 2, pp. 61–70, 2007.

[11] A.J. Berkhout, D. De Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journal of Acoustical Society of America*, vol. 93, pp. 2764–2764, 1993.

[12] D. de Vries, *Wave Field Synthesis*, Audio Engineering society Monograph, 2009.

[13] D. Schwarz, G. Beller, B. Verbrugghe, S. Britton, et al., "Real-time corpus-based concatenative synthesis with catart," in *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx), Montreal, Canada*. Citeseer, 2006, pp. 279–282.