SMuSIM: a Prototype of Multichannel Spatialization System with Multimodal Interaction Interface

Matteo Bozzolan
Department of Electronic Music
Conservatory of Music G.Verdi
Como, Italy
matteo.bozzolan@alice.it

Giovanni Cospito
Department of Electronic Music
Conservatory of Music G.Verdi
Como, Italy
giovanni.cospito@fastwebnet.it

ABSTRACT

The continuous evolutions in the human-computer interfaces field have allowed the development of control devices that let have a more and more intuitive, gestural and non-invasive interaction.

Such devices find a natural employment also in the music applied informatics and in particular in the electronic music, always searching for new expressive means.

This paper presents a prototype of a system for the real-time control of sound spatialization in a multichannel configuration with a multimodal interaction interface. The spatializer, called *SMuSIM*, employs interaction devices that range from the simple and well-established mouse and keyboard to a classical gaming used joystick (gamepad), finally exploiting more advanced and innovative typologies based on image analysis (as a webcam).

Keywords

Sound spatialization, multimodal interaction, interaction interfaces, EyesWeb, Pure data.

1. INTRODUCTION

Technology and music have always had a particular relationship and affinity. In particular, the researches and experimentations in the fields of electricity first and then of the electronics and informatics have allowed, in the last two centuries, the birth of a series of instruments for a new musical expressivity.

Besides, thanks to a more and more available computational power associated with the development of new technics and technologies for the human gestuality acquisition and analysis, new ways have opened in the field of the human-computer interaction, allowing so the birth of a new generation of interfaces that find a natural employment even in music applications.

As reported in [11], the most widespread interaction devices currently used are (in an increasing order of complexity): PC keyboard, mouse, joystick, MIDI keyboard, video camera, touchpad, touchscreen, 3D input devices (data gloves, electromagnetic trakers) or haptic devices.

This paper shows the results of the experimentation of some of these interfaces for the realization of a system for the multichannel spatialization of sound sources. In particular the devices explored in this work are: *mouse and keyboard* (very simple and primitive), a *gamepad* (classical gaming joystick) and a *webcam* (low cost USB camera, that allows, through image analysis techniques, a totally non-invasive and free-hand interaction).

In respect of the sound spatialization, the proposed prototype provides a quadriphonic sound diffusion and allows to control up to four independent sound sources. The spatialization technique implemented is the well-known Amplitude Panning extended to the multichannel case. This choice of simplicity find its motivation in the fact that the primary aim of this work is the investigation on the interaction interfaces rather than the implementation of advanced spatialization algorithms. The sound projection space can be artificially altered by controlling the direct to reverberated signal ratio.

2. RELATED WORKS

Although the use of spatial sound is present since from the origins of the music and it appears many times in classical western music, it becomes a fundamental practice and a key aesthetical element mainly from the second half of the past century (first thanks to the development of sound diffusion electrical devices and then because of the revolution of electronic and digital sound systems). For brevity, in this section are presented only some of the most recent works in the field of real-time sound spatialization digital systems.

A first example is *MidiSpace* [6], a system for the spatialization of MIDI sound files in virtual environments realized at the end of the '90s at the Sony Computer Science Lab in Paris. It is one of the earliest sound spatialization experiments in 3D worlds and it gives the user two distinct graphic interfaces to control the application: the first one (bidimensional) allows to displace the various sound sources (identified by a set of musical instruments) in the projection space, while the second one (three-dimensional and realized in VRML) controls the movements of an avatar in the virtual world. The spatialization technique is the two-channel Amplitude Panning and the interaction devices are mouse and keyboard.

A more recent work is represented by *ViMiC* [3], a real-time system for the gestural control of spatialization for small ensamble of players. It belongs to the wider project *Gesture Control of Spatialization* started in the 2005 at the McGill University IDMIL Lab (Montreal, Canada). It's very interesting because it allows the user to control the displacement of the sound sources simply by moving his hands in the air (thanks to a complex apparatus for movements interpretation and codification called *Gesture Description Interchange Format*). A set of 8 sensors (connected with an electromagnetic tracking system) is applied to the two

NIME08, Genova, Italy

hand of the player.

Zirkonium [9] is a software implemented to control the spatialization within the *Klangdom* system at the ZKM (Germany). The *Klangdom* is formed by 39 speaker and it can be controlled by *Zirkonium* through mouse and joystick. It implements various spatialization algorithms (Wave Field Synthesis, Ambisonics, Vector Base Amplitude Panning e Sound Surface panning) and it allows the user to define an arbitrary number of resources[1] to spatialize in the concert hall. The system is controlled through a simple graphic interface.

Challenging Bodies [4] is a complex multidisciplinary project for live-performances of disabled people realized at the Informatics and Music Department of the Regina University (Canada). Within this wide project, the *RITZ* system, through various techniques, allows to frontally spatialize up to 10 input signals coming from musical instruments with 7 loudspeaker placed in front of the players. Its control interface is made up by two windows: the first one, implemented in GEM[2], supplies a graphical feedback of the loudspeakers configuration and it allows to modify the position of the sound sources in the space, while the second one, the main control patch implemented in Pure Data, gives the user the possibility to set the relative and absolute sound levels. The system is hardly oriented to scalability and usability.

The last example is the work recently proposed by Schacher [10] at the ICMST of the Zurich university (Switzerland). It consists of a design methodology and of a set of tools for the gestural control of sound sources in surround environments. The spatialization is made through a structured and formalized analysis that allows to map the player gestures on the sources movements by applying various typologies of geometric transformations. From the point of view of the input devices, the system does not have a consolidated structure, but the interfaces used up to now spaces from data gloves equipped with multiple sensors (pressure, position, bending) to haptic arms and graphic and multitouch tablets. The spatialization algorithms used are the Ambisonics and the Vector Based Amplitude Panning.

3. IMPLEMENTATION: *SMUSIM*

SMuSIM is a multichannel sound spatialization system with a multiple and multimodal interaction interface. It is designed for real-time applications in musical expressive contexts (electronic music spatialization, distributed and collaborative network performances).

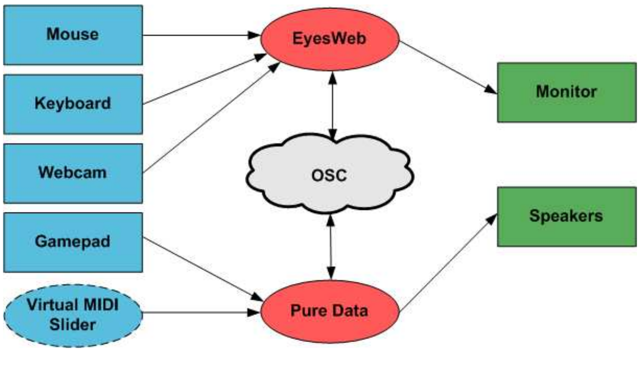

Figure 1: The system's architecture.

In this first implementation, the speaker are supposed to be arranged in the spatialization room in a typical quadriphonic configuration with the 4 loudspeaker placed at the corners of the room. The projection space can be artificially extended and modified by controlling the direct to reverbereted signal ratio (for the creation of illusory acoustic spaces).

The spatializer allows the player to control up to four simultaneous sound sources and a graphical feedback gives the instantaneous state of the system.

The system offers a set of functionalities that allows a complete and efficient control of the spatialization and in particular: a punctual and precise placement of the sound sources in the space, the control of relative and absolute volume levels, the automatization of the movements, a nonlinear interpolation of the position of the sources in time and the possibility to load pre-recorded sound files or to acquire signals coming from a microphone or any audio device.

As shown in Figure 1, the system has been implemented in *Pure Data*[3] (and its graphical interface *GrIPD*[4]) and *EyesWeb*[5] (with the creation of *ad hoc* additional blocks) communicating through the *OSC*[6] protocol, making so *SMuSIM* a native network distributed application (both with one ore more instances on several machines, allowing multiple distributed configurations).

3.1 Interaction interfaces

The prototype offers three different typologies of human-computer interaction devices for the spatialization's control. *Keyboard and mouse* are the simplest and the most widespread ones. The user controls the diffusion of the sound sources in the space through a combination of actions and commands coming from the PC keyboard and from the mouse. In this case the system provides (in addition to the visual feedback window) a bidimensional graphic environment where the player can put and move some graphic objects representing the different sound sources.

Figure 2: Input devices used for *SMuSIM*.

The second device is a *gamepad*, a classical gaming controller with two axis and ten buttons freely configurable. The very compact dimensions and the ergonomicity make the devices very usable and allows a great playability.

The last interface is a standard low-cost USB *webcam* that acquires the movements of a set of colored objects. Each physical object (through a color-based tracking algorithm) is associated to a sound object in the sound projection space.

The player can use one ore more devices at the same time (allowing a collaborative and multi-user performance). The proposed interfaces are deliberately simple, cheap and

[1] A resource is a set of one or more audio sources coming from an audio file, a network stream or any audio device.

[2] http://gem.iem.at

[3] http://www.puredata.org

[4] http://www.eyesweb.org

[5] http://crca.ucsd.edu/~jsarlo/gripd/

[6] http://opensoundcontrol.org

widely available on the market in order to let the system easily usable and accessible to any user level.

3.2 Software components structure

As shown in Figure 3, the application is composed by some functional units that perform the various needed tasks.

Data coming from the input devices are acquired, formatted and analyzed by the *Device controller* unit that is constituted by other 4 sub-units, one for each input device.

In particular *Mouse/Keyboard controller* supplies a graphic window (interaction environment) where the user can displace the four objects representing the sound sources with the mouse. A set of keyboard key combinations allows to perform a set of predefined actions (shifting of single or groups of sources, maintaining or not their topological configuration, loading/saving default configurations, etc.).

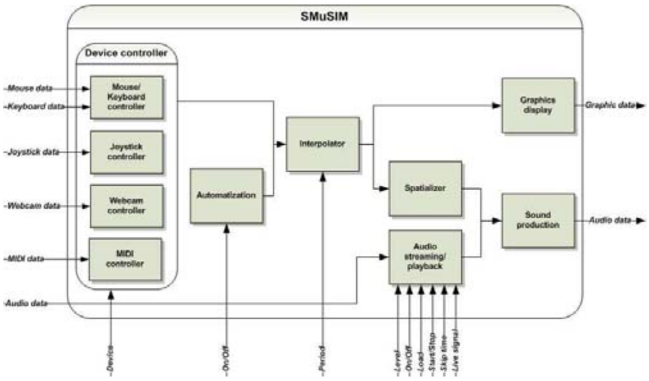

Figure 3: Diagram of the software functionalities implemented in *SMuSIM*.

Joystick controller allows to control the spatializer with a standard 2-axis and 12-buttons gamepad. The interface between the gamepad and the spatializer is managed through GrIPD, that provides all the needed functionalities. Buttons are used to select the sources to be controlled, while the two analog mini-sticks determine the changes in their position and volume. With this device it is easy to control more than one source at the same time[7]. *Webcam controller* manages data coming from the video acquisition device. The interaction paradigm in this case is the following: the webcam films a plane and neutral colored surface on which are placed the objects to be tracked; the webcam's field of view correspond to the diffusion space and the position of the colored objects determines the displacement of the sound sources on the sound projection space. The unit provides a set of tools for the real-time selection of the desired color to track (simply by picking it out on a window showing the webcam video stream) and for the extraction of centroids and bounding boxes of the color blobs. Bounding boxes are used to set the volume levels of each source (a vertical position stands for maximum volume, a horizontal one for mute). The *MIDI controller* block processes data coming from an optional MIDI device (both hardware and software).

The source movements can be automatized thanks to the *Automatization* unit, while the position's changes of sound sources are made not instantaneous through *Interpolator*, that generates a smoothed and decelerated motion by a non-linear interpolation of subsequent positional data.

Spatializer is the unit that performs the computation of the attenuation levels to apply to the audio signals on each channel. The spatialization technique is the Amplitude Panning extended to the multichannel case. On the basis of the positional data of the virtual sources coming from the input devices, a monophonic signal (considering a single source) is applied to the various channels with a gain factor as follows:

$$x_i(t) = g_i x(t), \qquad i = 1, ..., N$$

where $x_i(t)$ is the signal to apply to the loudspeaker i, g_i the gain factor of the correspondent channel, N the cardinality of the loudspeaker and t the time. The gain factor g_i has a non-linear proportionality with the position (x, y) of a single sound source in the space. To overcome the 6dB attenuation at the center of the projection space, a quadratic sinusoidal compensation curve is applied along the two dimensions. By considering all the sound sources involved, the resulting signal $X(t)$ can finally be defined as:

$$X(t) = \sum_{j=1}^{K} x_i(t)$$

where K is the maximum number of sound sources involved in the spatialization ($K = 4$ on the specific case of *SMuSIM*).

The graphic and audio feedback production is managed respectively by the *Graphics display* and *Sound production* units. The last one prepares the audio stream to send to the loudspeakers. It essentially manages the reverberation algorithm by applying it to the resulting signal coming from the combination of the original audio stream (furnished by the *Audio streaming/playback* unit) and the spatialization data, allowing in this way the creation of illusory acoustic spaces. By controlling the balance between the direct and reverberated signal independently for each channel, it is possible, besides increasing the overall distances perception, to deform the sound projection environment (by acting along one ore more dimensions of the room). Currently the functionalities of this unit are extremely limited in view of a future integration of a sound synthesis engine for the real-time generation of sounds.

4. FUTURE WORK

The system developed is still in a prototypal phase and has some limitations that can be easily improved. First it can be interesting to test some other interaction interfaces (to enlarge the multimodality issue) such as more performative cameras (higher frame-rate, infrared lighting) or other technologies for the exploitation of the gestural control of the instrument (electro-magnetic or ultra-sound tracking systems, data gloves). A study is currently active for the exploration of touch-sensible interfaces (graphic tablets, multitouch and painterly interfaces).

A second improvement refers to the spatialization technique, given that in this first phase of the project it has not been the crucial aspect of the work. The simple Amplitude Panning technique can be replaced by more complex and efficient algorithm such as the Vector Based Amplitude Panning, Ambisonics extended to a multichannel configuration and Wave Field Synthesis.

Another key issue is represented by the performances of the system that are the main requisite of the application in contexts of real-time musical performances. In fact there are actually some latency problems in the configuration with the webcam running particularly on not high performances machines or notebooks. This could be resolved by improv-

[7] Thanks to the compact size and to the ergonomicity of the gamepad, that allows the contemporaneous pressure of more than one button at the same time.

ing and optimizing both the tracking algorithm and the visual feedback production (in case abandoning the EyesWeb and Pd platforms and realizing an integrated, stand-alone and dedicated software application).

From the point of view of the automatization, it does not provide any way of interaction with the player, but it is an autonomous and isolated modality. It could be interesting the implementation of rules for pattern learning and reproduction in order to let the system able to imitate and continue a performance initially guided by a human user.

Other possible developments could refer to the diffusion system (increasing the number of loudspeakers and their configuration) and to the integration of a sound synthesis engine within the application.

During this first phase of the work there was not enough space for an intensive and structured test session on a large and heterogeneous set of users. However a hypothetical evaluation experiment has been predisposed for a future use.

The experiment has a total duration of about 45 minutes and it is composed by six sections:
1) free trial of the instrument (10 min) without any explanation about the working principles of the system (the user has previously read a short user manual)
2) supervised test (10 min) in which the user has to execute some tasks evaluated by the operators
3) explanation of the working principles (5 min) by an operator in order to increase the consciousness of control of the spatialization instrument and to accelerate the learning process
4) repetition of the test (10 min) after the explanations of the operator
5) questionnaire (5 min) of evaluation compiled by the user
6) interview (5 min) in which the operators deepen some aspects appeared during the test.
The two proposed tests contains list a of 21 tasks (for each test) that the user has to execute. Each task receives a mark according to a five point Likert-scale (1: not executed, 5: executed at the first trial). The tasks are sorted by the increasing level of difficulty and they are intended to test most of the functionalities of the instrument and its expressive possibilities. The questionnaire presents 22 questions divided into 5 categories: usability of the system (8), learnability (3), audio feedback (3), visual feedback (4) and overall opinion (4). Also in the questionnaire the players has to give a mark according to a five point Likert-scale (1: bad, 5: very good).

5. CONCLUSIONS

A real-time sound sources spatialization system with a multimodal interaction interface has been developed.

The interaction interfaces have been realized with very simple and inexpensive technologies and devices, that have nevertheless shown satisfactory expressive and interaction possibilities. In particular the best results came out, as expected, with the gamepad and the webcam, devices that allow more freedom in movements and a more intuitive and natural interaction. Moreover the webcam let the user move independently each sound source (action impossible with both the mouse and the gamepad). On the other hand, the performances are one of the key aspects associated with this last kind of device, because of the computational load of the image analysis techniques that make the real-time issue a crucial aspect of the application.

In general even all the graphic rendering operations for the creation of the visual feedback are particularly onerous for the overall performances of the system. Under this consideration, the graphic feedback proposed to the user is quite simple and thin, but it results very efficient and let have the actual state of the sound sources in the diffusion space always under control.

From the point of view of the sound spatialization, the Amplitude Panning technique produces the expected results. It is very efficient, it does not have problems of computational complexity and it is easily configurable to the various executive and technical contexts (customization of the panning curves and of the number of diffusion channels).

Even if an intensive and large scale test session has still to be conducted, *SMuSIM* has shown good results in terms of learnability, intuitivity and expressiveness. There are various possible developments of this work and they refer both to software and hardware issues (input devices, diffusion system) and applicative and musical aspects.

6. REFERENCES

[1] A. Camurri et al. Toward real-time multimodal processing: EyesWeb 4. In *Proceedings of the Convention on Motion, Emotion and Cognition (AISB04)*, Leeds, UK, 2004.

[2] J. M. Chowning. The simulation of moving sound sources. In *Journal of the Audio Engineering Society*, volume 19, pages 2–6, 1971.

[3] M. Marshall, Wanderley, et al. On the development of a system for the gesture control of spatialization. In *Proceedings of the 2006 International Computer Music Conference (ICMC06)*, pages 360–366, New Orleans, USA, 2006.

[4] J. Nixdorf and D. Gerhard. Real-time sound source spatialization as used in Challenging Bodies: implementation and performance. In *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06)*, pages 318–321, Paris, France, 2006.

[5] N. Orio, N. Schnell, and M. M. Wanderlay. Input devices for musical expression: borrowing tools from HCI. In *Proceedings of the 2001 International Conference on New Interfaces for Musical Expression (NIME01)*, 2001.

[6] F. Pachet and O. Delerue. A mixed 2D/3D interface for music spatialization . In *Proceedings of the First International Conference on Virtual Worlds*, pages 298–307, Paris, France, 1998.

[7] M. Puckette. Pure Data: another integrated computer music environment. In *Proceedings of the 1996 International Computer Music Conference (ICMC96)*, pages 269–272, Hong Kong, China, 1996.

[8] V. Pulkki. Spatial sound generation and perception by amplitude panning techniques. Graduation thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2001.

[9] C. Ramakrishnan, J. Gossmann, and L. Brummer. The ZKM Klangdom. In *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06)*, pages 140–143, Paris, France, 2006.

[10] J. C. Schacher. Gesture control of sounds in 3D space. In *Proceedings of the 2007 International Conference on New Interfaces for Musical Expression (NIME07)*, pages 358–361, New York, USA, 2007.

[11] L. Schomaker, A. Camurri, et al. A taxonomy of multimodal interaction in the human information processing system. Technical report, Nijmegen University, 1995.