Composing space in the space: an Augmented and Virtual Reality sound spatialization system

Giovanni Santini
Hong Kong Baptist University
info@giovannisantini.com

ABSTRACT

This paper describes a tool for gesture-based control of sound spatialization in Augmented and Virtual Reality (AR and VR). While the increased precision and availability of sensors of any kind has made possible, in the last twenty years, the development of a considerable number of interfaces for sound spatialization control through gesture, their integration with VR and AR has not been fully explored yet. Such technologies provide an unprecedented level of interaction, immersivity and ease of use, by letting the user visualize and modify position, trajectory and behaviour of sound sources in 3D space. Like VR/AR painting programs, the application allows to draw lines that have the function of 3D automations for spatial motion. The system also stores information about movement speed and directionality of the sound source. Additionally, other parameters can be controlled from a virtual menu. The possibility to alternate AR and VR allows to switch between different environment (the actual space where the system is located or a virtual one). Virtual places can also be connected to different room parameters inside the spatialization algorithm.

1. INTRODUCTION AND BACKGROUND

Sound spatialization has been used as a resource for musical expression at least since Willaert's production at *Basilica di San Marco* in Venice (mid 16th century) [1]. More recently, since the first implementations of electronic music and especially in the past few decades, with the development of advanced sound spatialization algorithms (e.g., Vector-based Amplitude Panning (VBAP) [2], Higher Order Ambisonics (HOA) [3]), spatial sound has become a key element of the compositional syntax for an increasing number of composers: "space as a finality in music expression" (Leo Kupper in [4]) and "space as a compositional language"([5]).

Since the first experiments by Pierre Schaeffer in the early 50s [1] one of the key aspects has been the control of the trajectories of sound sources (i.e., how to manipulate position coordinates through a "high-level" interface), along with the composition of many other parameters that can affect sound perception (e.g. directivity, aperture of sound source and room characteristics).

Many solutions have been developed by providing some form of graphic editing/automations. In order to achieve intuitiveness and ease of use in a context where a big number of parameters comes into play, often some specific form of gestural input has been deployed. Gestural interfaces include tablets or gamepads ([6], [7]), gesture recognition through camera input, both for visible light and infrared ([8]), [9]), or different sensors ([10]). More extensive reviews can be found in [11] and [12].

One further differentiation among systems can be identified between real-time sound spatialization systems or offline studio editing applications: in the latter group can be inscribed systems responding to the needs of computer-aided composition, i.e. intuitive controls to be connected to the development of a musical structure ([13], [6]). Real-time control systems can often be referred to as DMI (Digital Musical Instrument [11], [14]) and more specifically as Spatialization Instruments, defined as "a Digital Musical Instrument, which has the capability of manipulating the spatial dimension of the produced sound, independently of its capability of producing or manipulating the other sound dimensions" [12].

Notwithstanding the high differentiation in functionalities and implementation details, all the cited input models result in some kind of symbolic representation that does not show the sound source in its exact position in space. In other words, none of those system lets the user see and control the sound trajectory "as it is". Overcoming such limitations might provide a better control, as "[...] devices whose control structures match the perceptual structure of the task will allow better user performances." ([15], referring to [16]).

In the case of Spatialization Instruments, "matching the perceptual structure of the task" would mean to exactly see where the sound source is positioned in space [1] .

The recent advancements in VR and AR technologies provide the background for representing the sound location.

2. DESCRIPTION OF THE SYSTEM

The described tool allows to represent and control the behaviour of sound sources in a 3D immersive space, as well as to edit other sound source parameters and store, save and

[1] The limitations of direction and distance perception (that would counteract the idea of clear identification of sound source position and trajectory) will be discussed later.

recall those data. Such automations can be modified after creation. Representation of positioning is in real-world scale and has a reduced level of abstraction, prioritizing as much as possible intuitiveness and matching visual objects to sound behaviour.

The Augmented Reality implementation allows to see and place sources in the real space. The VR mode provides interaction with virtual environments. Different (real and virtual) locations can be linked to different audio room settings inside the spatialization algorithm.

The system is developed through the interaction of two main components:

- an AR/VR project developed in *Unity3D* for the *HTC Vive Pro* headset;
- a Max/MSP patch dedicated to sound spatialization by using *Spat* (Ircam tools).

The two programs talk to each other through OSC (Open Sound Control) protocol.

The system has been tested in the *LIATe* (Lab for Immersive Arts and Technology) at Hong Kong Baptist University, with a 24.2 channels setup.

Figure 1. 10 sources distributed over the Sound Spatialization setup in the LIATe shown in the Max object *spat5.oper*.

2.1 The Unity Project

The AR session is implemented in Unity for HTC Vive Pro, currently the only headset allowing both VR and AR applications.

The input comes from the two controllers for the Vive, which have 6 DOF (Degrees Of Freedom) motion tracking.

The right controller allows the positioning of one sound source at a time through *parenting* (an operation by which a virtual object is linked in position and rotation to another object). By moving the controller and pressing the back trigger, the user can create/modify the trajectory of the selected sound source. Such trajectory is shown as a line drawn in the air. As a child [2], a source can be given an offset respect to the parent controller, thus translating and magnifying the movement of the controller (for example, by shifting the sound source one meter above the controller on the Y axis, a 360 rotation of the controller would create a 2m diameter circle centered on the controller).

[2] A parent is the object providing the reference coordinate system, while a child is a virtual object whose coordinates are referred to the coordinates of the parent.

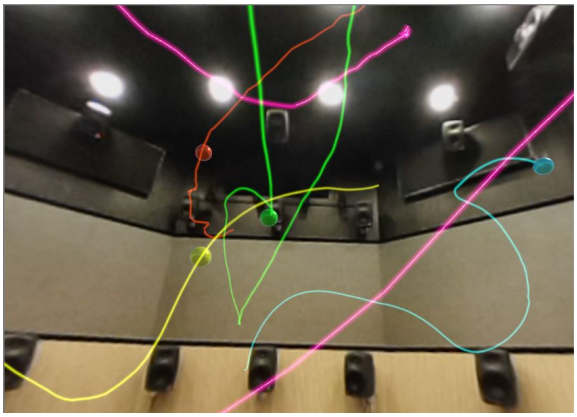

Figure 2. Point of view 1 on a combination of sources and trajectories.

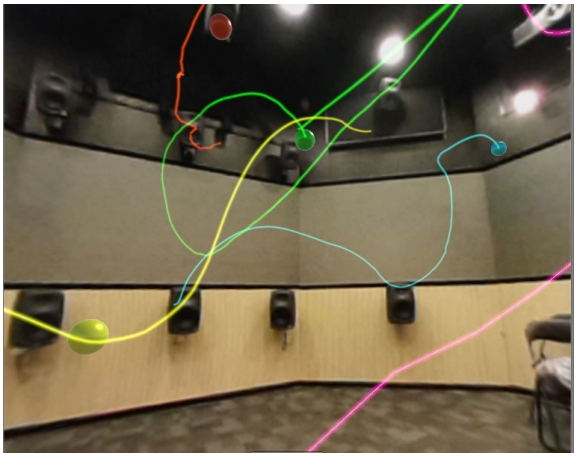

Figure 3. Point of view 2 on the same combination.

The position of sound sources (update frame by frame) is sent through OSC to Max/MSP (that performs the sound spatialization).

In the current state of development, the application allows to control up to 10 sound sources at the same time.

The left controller can move an additional sound source. Furthermore, it has a User Interface (UI) attached allowing for the selection of different tools (the UI only sends OSC commands to Max/MSP, which actually performs the tasks):

- shifting the sound source from the parent controller (over the three different axes);
- selecting and soloing (if needed) different sound sources and assigning different trajectories (recognizable by different colors);
- changing the aperture and yaw of the selected source;
- choosing the spatialization algorithm;
- changing the room (as a VR room);
- storing and recalling those trajectories; changing trajectories after drawing.

Figure 4. The Menu attached to the controller.

Sound sources are visualized as spheres of different colors; when they move, either they follow a trajectory or are moved by a controller. The trajectory is not followed with a fixed speed: speed is changed according to the original gesture (every trajectory is, originally, drawn with a gesture). If the sound source's duration is longer than the trajectory's one (e.g., the sound source is 3 seconds and the trajectory is 2 seconds long), the sound source is left static on the last point of the trajectory. However the gesture representation can be always edited in real time by pressing the trigger of the controller. Thus, the user can freely adjust a trajectory to the sound source it is related to.

2.2 Sound spatialization

OSC bundles sent out from Unity are received by a Max/MSP patch based on *Spat* (Ircam tools). As both Unity and Spat use a coordinate system where 1 corresponds to 1 meter, the passage from one system to the other does not require remapping except for coordinate systems alignment. While the AR/VR project in Unity can be considered the front-end of the application, all the core functions are actually implemented in Max/MSP and most of the functions control Spat parameters (position, sound source aperture, yaw, etc.).

The system uses different "coll" objects (each one for every different sound source), in order to store, save and recall trajectory information.

Different spatialization algorithms are available (e.g. 3D VBAP, HOA and binaural) [17], and their use is left to the discretion of the user.

Sources moving along trajectories can also be saved as audio tracks.

3. LOCALIZATION OF SOUND AND VIRTUAL OBJECTS

The presented application is based on a relation between virtual object position and sound source position; therefore a critical issue must be considered: distance estimation and respondence of visual and aural movements.

As [18] shows, the vision-based distance estimation of a virtual object presents problems in an AR environment. While the angular positioning is rather precise, the understanding of distance tends to be underestimated. The study evaluates numerous rendering strategies for virtual objects (such as aerial perspective [3], cast shadows [4] and shading [5]). The authors find, through two specifically designed experiments, that the most effective (by far) rendering strategy to reduce the underestimation of distance consists of casting shadows on the floor (rendered shadows are created by a virtual source of light perpendicular to the floor). In fact, in both experiments, cast shadows proved to increase accuracy in distance estimation respectively by 90% and 18%.

For audio discrimination, as shown in [19] and [20] many parameters and spectral cues enter into play: sound level, direct-to-reverberant ratio (DRR), spectral shape (e.g., low-pass filtering of frequencies in function of the distance), binaural cues like Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs), dynamic cues (motion) and familiarity with the sound. Even though such cues are important for giving an idea of distance, a precise estimation of the perceived distance is problematic. In fact, given the complexity of the overall perceptual system and the dependency of recognition upon many different factors, including the conformation of the venue itself, distance perception is biased and tends to underestimation.

[19] also shows that the presence of a visual cue can help in focusing the position of a sound source (sometimes producing *ventriloquism*, the phenomenon that occurs when a listener mistakenely adjusts the perception of sound localization to the position of the visual cue).

Moreover, the discrimination of behaviour of sources is made problematic by some other effects: for instance, one sound tends to be more sharply localized when its position coincides with the one of a real speaker. Another phenomenon we can take as an example, named as *flickering* in [5], consists in the impossibility for our hearing to discriminate position under a very fast source movement, or better, the tendency to ignore most part of a trajectory, by focusing only on some discontinuous points in space.

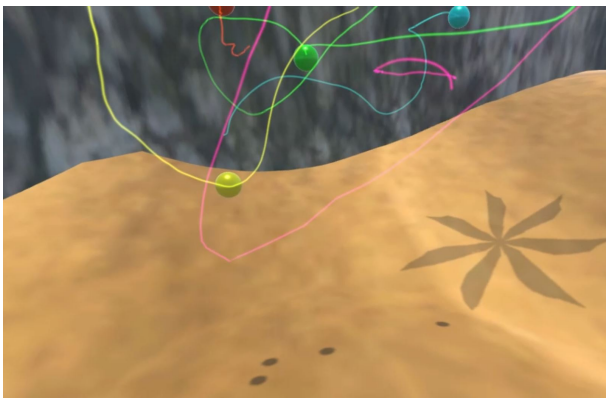

Figure 5. The same configuration of Figure 2 and 3 but in VR (and with down-cast shadows).

According to [19] and [20] the simultaneous presence of both visual and aural cues helps in discriminating position,

[3] Increased hazyness of colors with the increase of distance.
[4] Renderings of virtual shadows on the floor.
[5] Defining the reflectance properties of a virtual object.

distance and behaviour of a sound source; as down-cast shadows [6] help to have a correct estimation of virtual objects position, they further increase the precision of sources localization.

4. USES, LIMITATIONS AND FUTURE WORKS

The presented tool allows to control sound spatialization in an immersive environment, providing the visualization of sound sources' positions and trajectories. It allows fast testing of spatial compositional solutions and real-time control over numerous spatialization parameters. It can be used live as a Spatialization Instrument or off-line as a sort of (limited) Digital Audio Workstation (DAW).

As pointed out in [6] the limit of some gesture-controlled (real-time) systems might fall short for what concerns large-scale conception and compositional organization, especially in relation to musical structures that might prescind from bodily gestures. For this reason, a future improvement should include the possibility to edit trajectories even in a computer-aided composition context.

The Spatial Instrument described might seem to follow from a *naïve* approach: sound trajectories can be perceived with the same clarity of our visual perception (i.e., the two representations, visual and aural, of a movement are, to some extent, precise and identical). As already shown in [5], [19], [20] even hearing under the most ideal conditions, perceived distances appears to be "a biased estimate of physical source distance" [19]. As the perception of distance (but also of behaviour over time) is influenced by spectral characteristics of sound, the proposed system can be useful as a way for "fast prototyping", but cannot solve the problems inherent to sound spatialization, that in numerous cases require a tailored approach to different sound sources, sound fields and timbres.

In addition to the source-trajectory approach shown in this paper, another resource might be found in a spectral spatialization approach. One possible idea would consist of distributing different frequency bands of one audio file across the space as if they were different sound sources and providing each band with dynamic movements; while such approach could not have a single-bin accuracy while maintining intuitiveness of use, bin grouping based on psychoacoustic perception (such as Bark bands [21]) would certainly be possible. Therefore, it would be possible to obtain a fluctuating timbral environment by organizing the movement of different Bark bands inside one timbre.

Moreover, a future study will be addressed to the assessment of the usability and usefulness of the tool both with trained musicians and untrained people.

5. CONCLUSION

The paper has described a VR/AR immersive system for sound spatialization. It allows real-time control over position, trajectory and other parameters of different sound sources, visualized as spheres. Trajectories are visualized as virtual strokes.

The Digital Instrument mapping is intuitive, as sounds' positions and trajectories mirror the gesture of the player. These gestures can be translated in space and scaled (a small movement can result in a shift of several meters). A simple UI attached to the left controller allows the user to change different parameters and options (spatialization algorithm, sound source, aperture and yaw etc.). The application can be also used as a tool for automating trajectories and can be useful for electroacoustic composition. Data about sources movements can be stored as text in "coll" objects; spatialized soundfiles can also be exported as audiofiles.

The switch from AR to VR changes the environment where virtual sources are visualized from the real world to a VR landscape. Such possibility to switch makes it easier to render on the floor shadows of virtual objects representing sound sources. As [18] shows, such shadows, rendered under the objects with a virtual light perpendicular to the floor, increase the accuracy of estimation of virtual objects positions.

The intuitiveness of the system is enhanced by the simultaneous presence of both visual (representation of sound sources and trajectories) and aural cues. On the other side, such close mimicking between sound and visual behaviour might induce a simplistic approach (as if localization of sound sources could always be perfectly accurate). The user should always consider some degree of inaccuracy due to intrinsic characteristic of sound spatialization: the understanding of source positioning is influnced by many paraemters, such as intensity, direct-to-reverb ratio, and spectral EQ. Consequently, in numerous circumstances, a case by case approach should be considered.

[6] Down-casting shadows in AR requires a 3D scanning of the environment. HTC Pro has the capability to do so, but the range is rather limited and subject to visual artifacts. In VR shadows are easy to represent properly.

6. REFERENCES

[1] R. Zvonar, "A history of spatial music," *CEC*, 1999.

[2] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Audio Engineering Society*, 1997.

[3] D. Malham, "Higher order ambisonic systems for the spatialisation of sound," in *1999 International Computer Music Conference (ICMC)*, 1999.

[4] R. Normandeau, "Timbre spatialisation: The medium is the space," 2009.

[5] T. Schmele, "Exploring 3D Audio as a New Musical Language," Master's Thesis, Universitat Pompeu Fabra, 2011.

[6] J. Garcia, J. Bresson, and T. Carpentier, "Towards interactive authoring tools for composing spatialization," in *2015 IEEE Symposium on 3D User Interfaces, 3DUI 2015 - Proceedings*, 2015.

[7] K. Bredies, N. A. Mann, J. Ahrens, M. Geier, S. Spors, and M. Nischt, "The multi-touch SoundScape renderer," in *Proceedings of the working conference on Advanced visual interfaces - AVI '08*, 2008.

[8] D. Copeland, "The NAISA Spatialization System," 2014. [Online]. Available: http://www.darrencopeland.net/web2/?page{_}id=400

[9] W. Fohl and M. Nogalski, "A Gesture Control Interface for a Wave Field Synthesis System," in *Nime 2013 Proceedings of the International Conference on New Interfaces for Musical Expression*, 2013.

[10] M. L. Hedges, "An investigation into the use of intuitive control interfaces and distributed processing for enhanced three dimensional sound localization," Master thesis, Rhodes University, 2015.

[11] A. Pysiewicz and S. Weinzier, "Instruments for Spatial Sound Control in Real Time Music Performances. A Review." in *Musical Instruments in the 21st Century*. Singapore: Springer, 2017, pp. 273–296.

[12] A. Pérez-Lopez, "Real-Time 3D Audio Spatialization Tools for Interactive Performance," *Universitat Pompeu Fabra, Barcelona*, p. 38, 2014.

[13] R. Gottfried, "SVG to OSC transcoding as a platform for notational Praxis and electronic performance," in *Proceedings of the International Conference on Technologies for Music Notation and Representation*, Paris, 2015, pp. 154–161.

[14] J. Malloch, D. Birnbaum, E. Sinyor, and M. M. Wanderley, "Towards a new conceptual framework for digital musical instruments," in *Proceedings of the 9th International Conference on Digital Audio Effects*, 2006.

[15] M. M. Wanderley and N. Orio, "Evaluation of input devices for musical expression: Borrowing tools from HCI," *Computer Music Journal*, 2002.

[16] R. J. K. Jacob, L. E. Sibert, D. C. McFarlane, and M. P. Mullen, Jr., "Integrality and separability of input devices," *ACM Trans. Comput.-Hum. Interact.*, vol. 1, no. 1, pp. 3–26, Mar. 1994. [Online]. Available: http://doi.acm.org/10.1145/174630.174631

[17] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty years of Ircam Spat: looking back, looking forward," *International Computer Music Conference Proceedings*, 2015.

[18] C. Diaz, M. Walker, D. A. Szafir, and D. Szafir, "Designing for depth perceptions in augmented reality," in *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct 2017, pp. 111–122.

[19] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, 2005.

[20] A. J. Kolarik, B. C. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, and Psychophysics*, 2016.

[21] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.