Nuvolet : 3D Gesture-driven Collaborative Audio Mosaicing

Josep M Comajuncosas
Music Technology Group
Universitat Pompeu Fabra
Escola Superior de Música de Catalunya - ESMUC
josep.comajuncosas@esmuc.cat

Alex Barrachina
Escola Superior de Música de Catalunya - ESMUC
alex.barrachina@esmuc.cat

John O'Connell
Music Technology Group
Universitat Pompeu Fabra
johngerardoconnell@gmail.com

Enric Guaus
Escola Superior de Música de Catalunya - ESMUC
enric.guaus@esmuc.cat

ABSTRACT

This research presents a 3D gestural interface for collaborative concatenative sound synthesis and audio mosaicing. Our goal is to improve the communication between the audience and performers by means of an enhanced correlation between gestures and musical outcome. Nuvolet consists of a 3D motion controller coupled to a concatenative synthesis engine. The interface detects and tracks the performers hands in four dimensions (x,y,z,t) and allows them to concurrently explore two or three-dimensional sound cloud representations of the units from the sound corpus, as well as to perform collaborative target-based audio mosaicing. Nuvolet is included in the Esmuc Laptop Orchestra catalog for forthcoming performances.

Keywords

concatenative synthesis, audio mosaicing, open-air interface, gestural controller, musical instrument, 3D

1. INTRODUCTION

Direct manipulation of sound through visual representations, either by gestural, haptic or GUI-based interaction, takes advantage of well established audio representation techniques. By incorporating this paradigm, intuitiveness and easiness of use of such interfaces are maximized. Within this context, content-based navigation and retrieval of audio through scatter plots have become commonplace in MIR-based applications. For instance, Coleman[4] proposes a method for personal sample library exploration based on the analysis of event-synchronous audio segments extracted from a user's digital music collection. Janer[6] presented a sound object browser that allows the user to preview and select the desired target to assign to each step in a looped sequence. Some unit navigation systems for concatenative synthesis and audio mosaicing environments will be reviewed in Section 2.1.

NIME'11, 30 May–1 June 2011, Oslo, Norway.

Nuvolet adds an interaction layer to CataRT[1]. CataRT is a real-time corpus-based concatenative synthesis environment for MaxMSP. It has been developed at the IRCAM by D. Schwarz[15], and allows the user to freely navigate scatter plots of sound corpuses as well as performing target-based audio mosaicing. While there are some cataRT-based systems designed for laptop ensembles (p.ex. Catork[2]), our research is focused on on-stage gestuality to convey expressiveness and intelligibility to the sound exploration process.

Nuvolet bears a strong resemblance to The Enlightened Hands proposed by Viglienson[16]. Comparing both systems, our interface provides unobtrusive multiuser three-dimensional navigation and gestural control of target-based resynthesis.

The paper is organized as follows. In Section 2 we present the most relevant advances in concatenative sound synthesis and in gesture-based interfaces. Next, in Section 3, both the concept and the architecture of the interface are presented, followed by the description of two different case examples, data cloud navigation and interactive target-based mosaicing, in Section 4. Finally, we present a discussion for the system design and interaction issues, and the final conclusions in Sections 5 and 6 respectively.

2. STATE OF THE ART

In this section, we present the most relevant advances in concatenative sound synthesis and in gesture-based interfaces for our work.

2.1 Concatenative Sound Synthesis

Concatenative Sound Synthesis (CSS) is a process whereby audio is created by the concatenation of many small segments of audio, called units, from a source unit database, called a corpus. In this process, unlike in traditional granular synthesis methods, the grain selection is not arbitrary but rather determined by the characteristics of the audio itself. This *data driven process* [14] may take a given audio input as a "target" from which a list of audio features called descriptors are derived.

Source units from the corpus are then selected based on how well they match selected descriptors of the target. Typically, the multi-dimensional descriptor space is searched using a path search algorithm (e.g. Viterbi[14]) or an adaptive local search algorithm (e.g. Zils[20]). This process is called *unit selection.* The target specification is often derived from

[1] `http://imtr.ircam.fr/imtr/CataRT`
[2] `http://www.brunoruviaro.com/catork/`

a piece of audio[20] or from user navigation through the corpus of source units.

Diemo Schwartz has been exploring real time improvisation with CataRT by analyzing and segmenting live audio captured onstage from a musician [3]. Several authors investigate as well how to navigate the multidimensional descriptor space, for example plumage [5], which uses a custom 3D interface to control CataRT. Compared to it, Nuvolet relies on direct mapping from the spatial dimensions to a three-dimensional sound space, thus achieving a touchless but direct manipulation of the virtual timbral space.

2.2 Gesture based interfaces

A pioneering three-dimensional controller was the Radio-Drum, by Boie and M.Mathews, which tracked the batons 3D location by radio-frequency. Another system closer to the interface presented in this paper is Lightning, by D.Buchla, a device which tracked the performer location in the vertical plane by triangulating the infrared transmitters built into baton-like wands. Both were introduced at the early 1990s [3].

The Theremin-like quality of such gestural devices quickly dives into the realms of dance, theater and interactive installations when the space and number of performers increase [10]. Coherently, systems that utilize video capture and IR motion capture devices had been employed since the eighties for dance driven music, as Simon Veitch's 3DIS system [2] and David Rokeby's VNS (Very Nervous System) [18, 19].

More recent developments employ a large number of sensing devices for active location and/or motion capture. A paradigmatic example is the Brain Opera [13], a large multimedia production conceived by Tod Machover and Joseph Paradiso in the late nineties, which implemented a number of open-air sensing techniques, ranging from capacitive sensing for small areas to arrays of ultrasonic range finders or microwave Doopler radars.

Figure 1: Two performers playing the *Nuvolet*, as seen in the monitoring screen

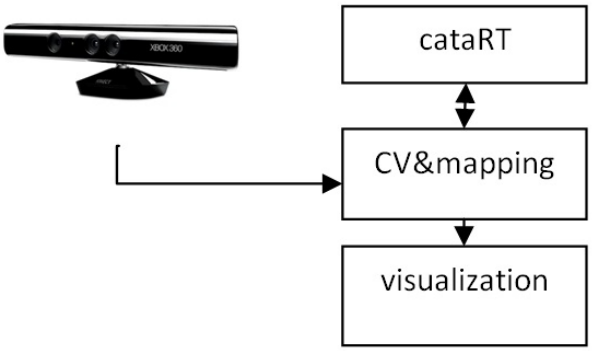

Figure 2: *Nuvolet* block diagram.

3. SYSTEM OVERVIEW

The Nuvolet, originally developed for a musical work written by the catalan composer Ariadna Alsina for singer-reciters and laptop ensemble, is designed to let a number of performers (one to four) of the Esmuc Laptop Orchestra[4] to explore a multidimensional representation of audio snippets by moving their hands in the space.

3.1 Concept

According to the aesthetic intentions of the work, the performers should exemplify some key concepts from the script by visually drawing soundscapes. As the singers evoke places attached to their childhood memories, Nuvolet players navigate a sound corpus made up of field recordings from those locations.

The sound cloud is shown to the performers as a 3D overlay on a monitoring screen, as displayed in Figure 1. Our main goal was to improve the communication with the audience by an enhanced correlation between gestures and musical outcome, achieving at the same time an increased performability compared to the CataRT mouse based GUI. Broad gestures amplify the perception of the player manipulations, which may have a positive effect on the perceived authenticity and expressiveness of the performance.

3.2 System architecture

The system consists of a capture device and a set of software modules, as displayed in Figure 2. A three-dimensional representation of the performer's location is obtained with the Microsoft Kinect[5], which provides an infrared laser projector for robust, ambient light immune depth sensing. The stereoscopic vision is thus achieved through point cloud optical triangulation and the available playing area is about $6m^2$, with a tracking range of 0.7 to 6 meters.

The computer vision software consists of the OpenNI[6] framework, which takes care of the skeleton tracking, and a custom openFrameworks[7] application which performs the required mapping. Only subject and hand tracking were necessary for this project. This application also takes care of the visualization of the sound clouds for performer feedback, as already seen in Figure 1.

Finally, the audio synthesis engine is the concatenative synthesis environment CataRT described in Section 1. Most of the synthesis features required were already available in cataRT, namely audio segmentation, corpus analysis and target driven mosaicing. Only a polyphonic synthesis engine was implemented to minimize interference between performers, and an OSC link was added to interchange data with

[3] http://www.youtube.com/theconcatenator.
[4] http://barcelonalaptoporchestra.blogspot.com/
[5] http://www.xbox.com/es-ES/Xbox360/Accessories/kinect/Home
[6] http://www.openni.org/
[7] http://www.openframeworks.cc/

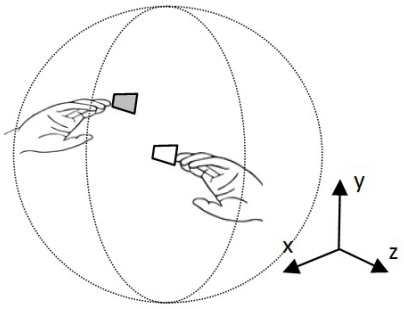

Figure 3: Collaborative sound navigation in Nuvolet.

the visual layer.

4. CASE EXAMPLES

This section shows two different mapping strategies: the first one is designed to concurrently explore two or three-dimensional sound cloud representations of the units from the sound corpus, and the second one is oriented towards collaborative target-based audio mosaicing.

4.1 Descriptor space navigation

In our first scenario, a simple mouse replacement for the CataRT data explorer was implemented, extending the original 2D mouse-driven LCD GUI to a three-dimensional data cloud with concurrent access to sound snippets. Users may thus collaboratively explore the sound corpus represented by 3 audio descriptors directly mapped to the spatial dimensions, as displayed in Figure 3.

For the performers' gestures to be perceived as musically meaningful, we mapped the vertical dimension to frequency-related descriptors (such as the spectral centroid) and the z (depth) dimension to energy-related descriptors (such as the rms), which proved to be intuitively playable and easily understandable by the audience.

Although this interaction model may seem obvious, sparsity and uneven population of the descriptor space makes navigation through the units difficult, as already noted by [14, 15].

4.2 Interactive mosaicing

A second mapping strategy was implemented to allow the performers to explore target-driven audio mosaicing. It is necessary to incorporate a virtual time pointer in the interface, for example as a given path which the user should follow to retrieve the minimum distance audio units for a faithful reconstruction of the target. This pointer is then sent to CataRT as the desired target position, and CataRT itself then chooses the closest matching units from the corpus.

The spatial trajectory associated with this target should be preset in advance. If it is the user who chooses a suitable, continuous path, it could, for example, define a pictorial shape which may have a semantic relationship with the target. For example, when reconstructing a whispered voice from sea recordings, drawing a shape halfway between a lip and a wave turned out to be rather suggestive.

The new mapping is schematized in Figure 4. Note how the cloud along the predefined XY path (the target path) now consists of slices in YZ, normal to the path plane. For each target frame, the YZ plane displays all the corpus units located according to the descriptor distances.

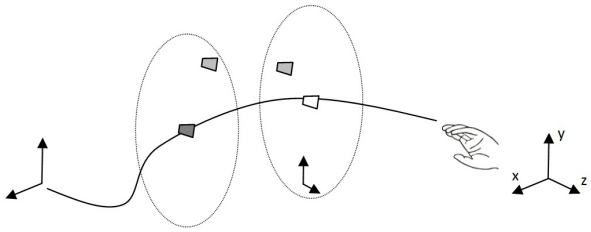

Figure 4: Gesture mapping for realtime target-based mosaicing

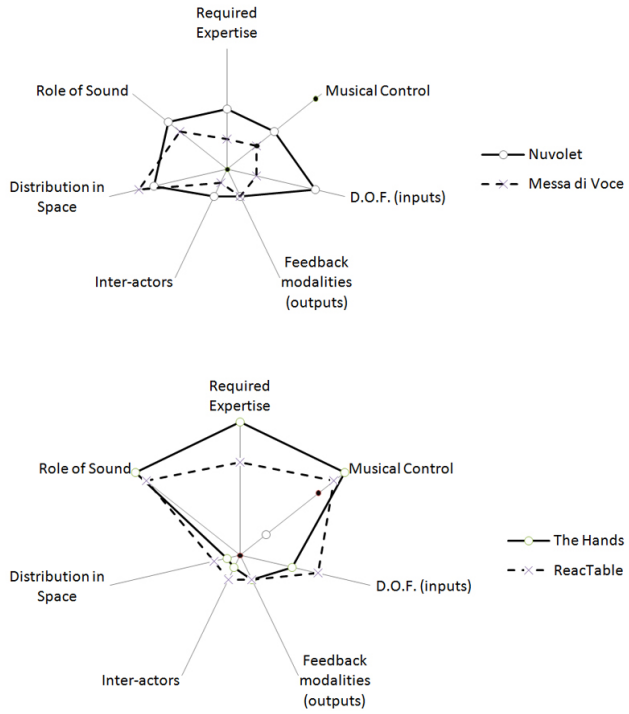

Figure 5: dimension space plots for four different gestural interfaces

This new mapping provided a richer experience to the performers, but concurrent use of the interface was less obvious.

5. DISCUSSION

The presented interface lies between the archetypes of a musical instrument and an interactive installation. Depending on the specific focus, its reference paradigm can range from sonification of gestural information to gesture-based sonic exploration. Figure 5 displays the 7-axis dimension space plots (see [1]) of Nuvolet, a virtuosistic interface like M. Waisvisz's The Hands, Jaap Blonk's solo from Messa di Voce, a performance and installation for voice and interactive visuals [7] and the collective tangible interface ReacTable, from Sergi Jordà. Nuvolet lies between The Hands and Messa di Voce, which also explores the realms of voice visualization, but is clearly shifted towards a sound installation profile.

The very nature of the interface imposes severe constraints on a number of desirable features for musical instruments, as listed in [12], like generality or perceivable correspondence between performer *effort* and sound quality. Moreover, the lack of the haptic channel for feedback in open-

air controllers increases the cognitive demands in such interfaces [17], which makes the trade-off between precision and agility even harder. These issues are most successfully addressed in smaller, instrumental interfaces, as with the Silent Drum Controller [11].

We observed however that through practice, performers were gradually less dependent on the visual cues and relied more on proprioception and kinesthetic feedback, but at the cost of being too static onstage.

6. CONCLUSIONS

We presented Nuvolet, a gestural interface for collaborative exploration of sound clouds and interactive target-based audio mosaicing. Nuvolet offers direct manipulation of multidimensional representations of sound corpuses by moving the hands on the space. A number of mapping strategies were studied, as well as an evaluation of the defining features of the new interface. Despite its rather satisfactory behavior in a restricted context, the challenges inherent in the design of open-air interfaces pose some unavoidable issues which deserve further research.

A logical addition to the interface could be a simultaneous gestural control of the sound spatialisation and an adoption of the general GDIF OSC namespace for exchanging of gesture related information between the software modules, in the line of [9], as well as the definition of a semantic mapping layer if more complex and multiuser input gestures are adopted [8] .

7. ACKNOWLEDGMENTS

The authors would like to thank all the members of the Esmuc Laptop Orchestra for their collaboration and continuous feedback throughout the development and testing of Nuvolet.

8. REFERENCES

[1] D. Birnbaum, R. Fiebrink, J. Malloch, and M. M. Wanderley. Towards a dimension space for musical devices. In *Proceedings of the New interfaces for Musical Expression (NIME)*, pages 192–195, 2005.

[2] W. Burt and A. Thompson. Fair exchanges'. *Writings on Dance V*, 1990.

[3] C. Casciato. On the choice onf gestural controllers for musical applications: An evaluation of the lightning ii and the radio baton. Master's thesis, McGill University, 2007.

[4] G. Coleman. Mused: Navigating the personal sample library. In *Proceedings of the International Computer Music Conference (ICMC)*, Copenhagen, Denmark, August 2007.

[5] C. Jacquemin, R. Ajaj, R. Cahen, Y. Olivier, and D. Schwarz. Plumage: design d'une interface 3d pour le parcours d'échantillons sonores granularisés. In *Proceedings of the 19th International Conference of the Association Francophone d'Interaction Homme-Machine*, pages 71–74. ACM, 2007.

[6] J. Janer, M. Haro, G. Roma, T. Fujishima, and N. Kojima. Sound object classification for symbolic audio mosaicing: A proof-of-concept. In *Proceedings of the Sound and Music Computing Conference (SMC)*, pages 297–302, Porto, Portugal., July 2009.

[7] G. Levin and Z. Lieberman. In-situ speech visualization in real-time interactive installation and performance. In *Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering*, pages 07–09, 2004.

[8] J. Malloch, S. Sinclair, and M. Wanderley. From controller to sound: Tools for collaborative development of digital musical instruments. In *Proceedings of the International Computer Music Conference (ICMC)*, Copenhagen, Denmark, August 2007.

[9] M. Marshall, N. Peters, A. Jensenius, J. Boissinot, M. Wanderley, and J. Braasch. On the development of a system for gesture control of spatialization. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 260–266, New Orleans, USA, August 2006.

[10] J. Mustard. *The integrated sound, space and movement environment: The uses of analogue and digital technologies to correlate topographical and gestural movement with sound*. PhD thesis, Western Australian Academy of Performing Arts, 2006.

[11] J. Oliver and M. Jenkins. The silent drum controller: A new percussive gestural interface. In *Proceedings of the International Computer Music Conference (ICMC)*, 2008.

[12] G. Paine. Towards unified design guidelines for new interfaces for musical expression. *Organised Sound*, 14(02):142–155, 2009.

[13] J. Paradiso. The brain opera technology: New instruments and gestural sensors for musical interaction and performance. *Journal of New Music Research*, 28(2):130–149, 1999.

[14] D. Schwarz. *Data-driven concatenative sound synthesis*. PhD thesis, Université Paris, 2004.

[15] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton. Real-time corpus-based concatenative synthesis with catart. In *Proceedings of Digital Audio Effects (DAFx)*, Montreal, Canada, September 2006.

[16] G. Vigliensoni. The enlightened hands: navigating through a bi-dimensional feature space using wide and open-air hand gestures. In *Proceedings of the New interfaces for Musical Expression (NIME)*, Sidney, Australia, June 2010.

[17] G. Vigliensoni and M. Wanderley. Soundcatcher: Explorations in audio-looping and time-freezing using an open-air gestural controller. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 100–103, New York, USA, 2010.

[18] T. Winkler. Making motion musical: Gesture mapping strategies for interactive computer music. In *Proceedings of the International Computer music Conference (ICMC)*, pages 261–264, Banff, Canada, 1995.

[19] T. Winkler. Motion-sensing music: Artistic and technical challenges in two works for dance. In *Proceedings of the International Computer Music Conference (ICMC)*, Ann Arbor, USA, 1998.

[20] A. Zils and F. Pachet. Musical mosaicing. In *Proceedings of Digital Audio Effects (DAFx)*, Limerick, Ireland, December 2001.