Norwegian Artificial Intelligence Symposium, Gjøvik, 22 November 2010

A video based analysis system for realtime control of concatenative sound synthesis and spatialisation

Alexander Refsum Jensenius,[♮♯] Victoria Johnson[♯]
[♮] fourMs, Department of Musicology, University of Oslo
[♯] Norwegian Academy of Music

Abstract—**We report on the development of a video based analysis system that controls concatenative sound synthesis and sound spatialisation in realtime during concerts. The system has been used in several performances, most recently *Transformation* for electric violin and live electronics, where the performer controls the sound synthesis and spatialisation while moving on stage.**

1. Introduction

Motion capture technologies are increasingly becoming popular in various types of interactive systems. Here *motion capture* is used to denote systems that can in different ways track information about a person's position and motion in space over time. Such systems can be anything from affordable sensor devices (e.g. accelerometers) to electromagnetic, mechanical and optical infrared laboratory equipment.

In our research we use all of the above mentioned motion capture systems, and explore how they can be used for analysis of music-related body movement, and the control of sound synthesis from such movements. This includes everything from how instrumentalists may control sound effects modifying their own sound while playing a traditional instrument, to purely electronic sound generation through motion in the air.

At the core of many of these activities is the need for understanding more about the relationships between complex body motion on one side and complex sonic material on the other. Complex is here used to denote the multidimensionality of both motion and sound, both in analysis and synthesis. Many interactive systems are often based on simple one-to-one mappings between action and response, something which is also the case in most commercially available music technologies. Keyboard based synthesisers, for example, typically work by creating a simple relationship between the velocity of the key being struck to the loudness envelope of the sound. The limitations of commercial music technology may to a large extent be caused by the limitations in the MIDI standard [6], but also due to the lack of better conceptual frameworks for understanding the relationships between sound and motion in music.

We are interested in exploring how complex human motion can be used to control complex sound synthesis in realtime, something which requires more advanced feature extraction, classification and machine learning techniques. This paper reports on a pilot study where we have explored the use of feature extraction and classification techniques to create a richer sonic interaction than would have been possible with more traditional synthesis and mapping methods.

2. Artistic idea and method

The starting point for the work being presented here was the artistic idea of letting the performer (the second author) navigate around in a large soundscape on stage, as shown in an early sketch of the project (Figure 1). Performing with an electric violin using a wireless sound transmitter makes her able to move freely in space while not having to think about issues such as sound feedback. The idea was therefore that she could trigger sonic material while moving to different locations on the floor. This sonic material could then be used as the basis for further improvisation.

The current exploration has been carried out employing an artistic working method, an iterative process following a systematic "trial and error" approach. For each attempt, a critical evaluation of the success of the tested parameters have been carried out at three levels: *technology*, *interaction* and *sonic output*. The evaluation has been done with respect to these criteria:

- Stability
- Reproducibility
- Complexity
- Creativity

The two first points should be obvious, a system that shall be used in public performances needs to be

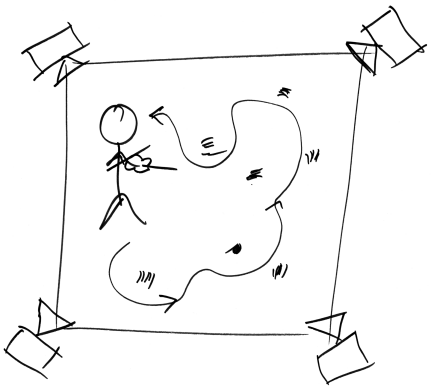

Fig. 1: An early sketch of the artistic idea, where the violinist could move on stage and trigger sonic objects.

both stable and able to reproduce results explored in rehearsal. The two latter points are connected to the artistic needs of having a system that is both complex enough to be artistically interesting to play with, and also creative enough for musical explorations.

3. Setup

Figure 2 presents an overview of the system. The following sections will describe the three main parts: motion capture, sound synthesis and sound spatialisation.

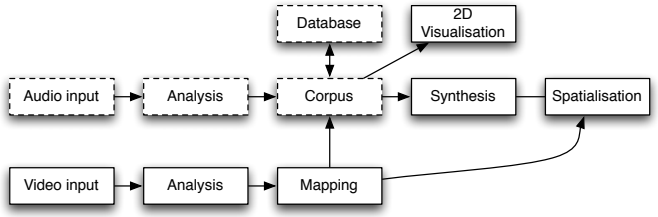

Fig. 2: An overview of the system (boxes with dotted lines show the non-realtime parts).

3.1. Motion capture

Several different types of motion capture solutions were tested. Using an infrared optical motion capture system (Qualisys Oqus 500) was abandoned early in the process. While such systems provide for accurate, precise and fast tracking of absolute position, they are not ideal for concert use. This is due to the large amount of equipment needed, as well as challenges when it comes to calibration, reflections, etc. outside of a lab environment.

We also tested an Xsens MVN motion capture suit, which can capture absolute position based on sensor fusion from sets of accelerometers, gyroscopes and magnetometers. Since this is an on-body system it is easier to carry and set up in a concert hall, while still providing fairly high speed, accuracy and precision, as well as wireless connectivity. As such, it solves many of the problems that are found with optical infrared systems, but it creates others. The most important challenge when using the suit, is that it is uncomfortable to wear, something which makes it less ideal to use for a performer.

Parallel to experimentation with the Qualisys and Xsens motion capture systems, we have explored various types of smaller accelerometer based sensor systems [5], [3]. One problem here has been instability issues with bluetooth wireless connections, especially in some concert halls. For this reason we are currently exploring ZigBee communication instead [11]. However, even with stable wireless communication, an accelerometer based solution would not be ideal, since we are interested in finding the absolute location in space.

Finally, we decided to use video based analysis for the motion capture, with a camera hanging in the ceiling above the stage. While video analysis is comparably slower than the other types of sensing systems, and requires more CPU, it has the advantage of easily being able to track the absolute position in space. By placing the camera in the ceiling, we also effectively removed challenges when it comes to separating the performer from the background.

The video analysis was implemented as modules in the Musical Gestures Toolbox [4] for the open framework Jamoma,[1] which is based on the graphical music programming environment Max [8]. This is a realtime system optimised for sound and video, and is also easily reconfigurable in rehearsal and performance.

Figure 3 shows a screenshot from the Max patch, where three different modules are used for getting video from the camera, calculating the *motion image* (the running frame difference), and using this as the basis for finding the area and centre position of the motion image. Then we know the location of the performer in space, and how much she moved.

3.2. Sound synthesis

The sound synthesis is based on a technique called *concatenative synthesis*, using the CataRT library[2] for Max [10]. The method is based on cutting up a collection of sound material into small sonic fragments, each of which can be recombined in different ways in realtime.

We have explored many different types of CataRT settings, with a huge collection of different sound material. In the end we have come up with a sample library of approximately 10 minutes of different violin sounds (mainly pizzicato and flageolets) which is used as input to the system. The sounds are fed to the analyser, which starts by slicing them into pieces of 232 ms, a

[1] http://www.jamoma.org
[2] http://imtr.ircam.fr/imtr/CataRT

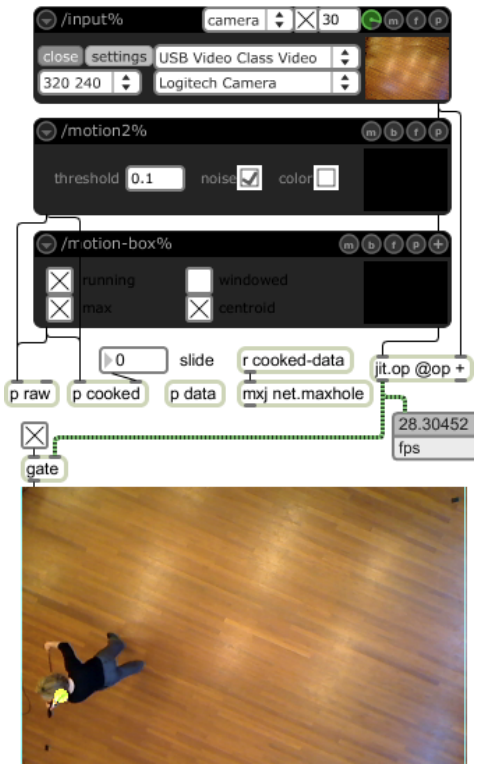

Fig. 3: Screenshot of the video analysis modules developed in Max 5. The input module reads video from the camera, and passes on to the motion module which calculates the motion image. Finally the box module calculates the area and centre of motion, the latter visualised as a circle on the head of the performer.

duration which is perceptually relevant but still short enough for being able to be spliced with other sounds. Each of the sound slices are analysed using a subset of MPEG-7 low level audio features [1], including pitch, loudness, periodicity, spectral flatness, spectral centroid, high frequency energy, mid frequency energy, high frequency content and energy. The final result is a database containing pointers to each of the original sound files, the start and stop position of each segment, and the results for all the extracted features. These features are then used for plotting the relative distance between the sound fragments in a 2-dimensional display, as can be seen in Figure 4.

CataRT allows for quickly, easily and efficiently changing between different organisation of the sounds in space. By mapping the position coordinates of the performer to CataRT's 2D display, sound playback can be controlled in realtime by the performer.

After experimentation we have found that a sonic distribution with spectral centroid on one axis and periodicity loudness on the other is the most interesting combination for interacting with our sound database. This setting gives the performer (and the audience) a

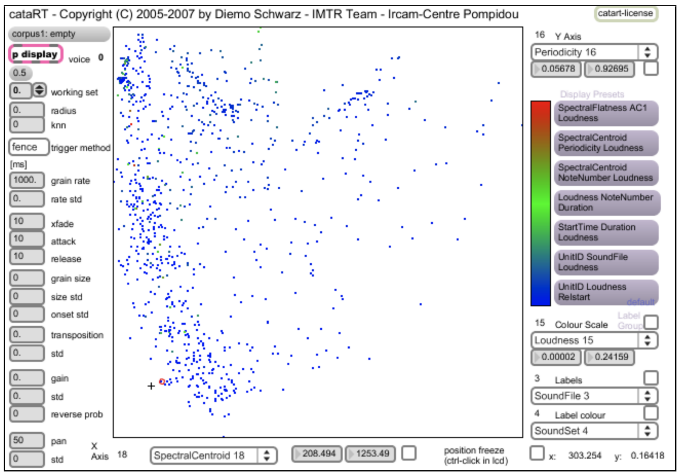

Fig. 4: Screenshot of the 2D display in CataRT. Each of the small dots represents a sound fragment of 232 ms, and the organisation of the dots can be controlled in realtime by changing which features should be mapped to the two axes of the display.

clear perceptual understanding of the two axes, while still allows for interesting sounds to appear close to each other on the floor. The end result is a sonic texture built up of individual fragments that are perceptually similar, but still different and thus more musically interesting.

3.3. *Sound spatialisation*

The last part of the system is the placement of sounds in space, or what is often referred to as *spatialisation*. Since the piece is focusing on exploring the physical space through a virtual space, we also wanted to distribute the sounds dependent on where the performer was moving.

Different spatialisation setups have been tested. For the last performance with the system, the room was set up with chairs on all four sides of the stage. Here we decided to place a set of smaller speakers at the corners of the stage area, and four larger speakers on the diagonals close to the walls (see Figure 5).

As spatialisation technique we decided to use *vector based amplitude panning* (VBAP) [9]. This is a CPU efficient technique that makes it possible to distribute sounds in space using simple matrix operations. For the current setup, a simple one-to-one mapping was set up between location on the floor and the placement of sounds, but this is something which needs to be explored more in future performances.

So far we have been carrying out the exploration in fairly reverberant spaces, but have still found the need to add a little extra reverb to the sound. The concatenative synthesis cuts the grains quite accurately, and even though there are no clicking or glitches in the playback, we have seen the need for some additional

Fig. 5: Rehearsal before the performance at Norwegian Academy of Music. Parts of the computer and mixer setup to the left, the camera hangs in the ceiling, and the 8 loudspeakers are placed in two squares around the quadratic stage area.

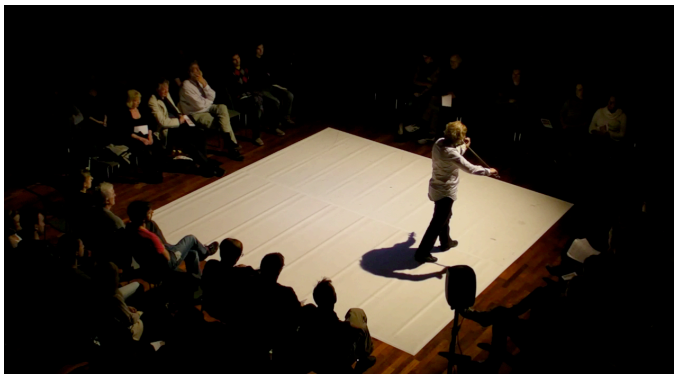

Fig. 6: An image from the concert 3 September 2010. A visual element, the white carpet also marked the boundaries for the video analysis area.

reverb to create a more holistic soundscape. This is done by using a simple mono reverb effect on each grain.

4. Discussion

The presented system has so far been used for several workshops and during three public performances :

- The foyer of the Norwegian Opera & Ballet (26.11.2009)
- The National library of Norway (4.2.2010)
- Norwegian Academy of Music (3.9.2010), (Figure 6)

There has been no software instability in neither rehearsal nor performance, and we have found the system to meet all the criteria outlined in the beginning: stability, reproducibility, complexity and creativity.

There are, of course, many possibilities for further refinement and development that will be explored in future research:

Tracking: The current tracking of position and motion on stage has proven to be stable, but also with the limitations found in video cameras: speed and resolution. We will explore using high speed and high resolution cameras to improve the response time. This will also be combined with small 6D Zigbee based sensor devices containing accelerometers, gyroscopes and magnetometers [11].

Adaptability: A drawback with the current system is the need for manual calibration. This we hope to improve by creating an auto-calibration routine so that the system can adjust itself to a new location and light condition.

Motion feature database: CataRT is based on extracting various features that are perceptually relevant. We are currently exploring similar feature extraction and classification techniques for motion capture data, so that motion features can be treated in the same way as we now work with sound data [2].

Action-sound synthesis: Based on a future database of motion capture data, we aim at creating a system with relationships between motion segments (i.e. actions) and sound objects. This will open for more complex mappings between action and sound. A prototype study of this has already been presented in [7], and will be further refined in future experimentation.

References

[1] M. Casey. General sound classification and similarity in MPEG-7. *Organised Sound*, 6(2):153–164, 2001.

[2] K. Glette, A. R. Jensenius, and R. I. Godøy. Extracting action-sound features from a sound-tracing study. In *Proceedings of Norwegian Artificial Intelligence Symposium, Gjøvik, 22 November 2010*, 2010.

[3] K. Guettler, H. Wilmers, and V. Johnson. Victoria counts – a case study with electronic violin bow. In *Proceedings of the 2008 International Computer Music Conference*, Belfast, 2008.

[4] A. R. Jensenius, R. I. Godøy, and M. M. Wanderley. Developing tools for studying musical gestures within the Max/MSP/Jitter environment. In *Proceedings of the International Computer Music Conference, 4-10 September, 2005*, pages 282–285, Barcelona, 2005.

[5] K. A. McMillen. Stage-worthy sensor bows for stringed instruments. In *Proceedings of New Interfaces for Musical Expression*, pages 347–348, Genova, 2008.

[6] F. R. Moore. The dysfunctions of MIDI. *Computer Music Journal*, 12(1):19–28, 1988.

[7] K. Nymoen, K. Glette, S. A. Skogstad, J. Tørresen, and A. R. Jensenius. Searching for cross-individual relationships between sound and movement features using an svm classifier. In *Proceedings of New Interfaces for Musical Expression++*, pages 259–262, Sydney, Australia, 2010.

[8] T. Place and T. Lossius. Jamoma: A modular standard for structuring patches in max. In *Proceedings of the 2006 International Computer Music Conference*, pages 143–146, New Orleans, LA, 2006. San Francisco: ICMA.

[9] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, pages 456–466, 1997.

[10] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton. Real-time corpus-based concatenative synthesis with Catart. In *Proceedings of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*, Montreal, 2006.

[11] J. Torresen, E. Renton, and A. R. Jensenius. Wireless sensor data collection based on zigbee communication. In *Proceedings of New Interfaces for Musical Expression++*, pages 368–371, Sydney, Australia, 2010.