Approaches to Real Time Ambisonic Spatialization and Sound Diffusion using Motion Capture

Alyssa Aska
University of Calgary
Calgary, Alberta, Canada
alyssa.aska@ucalgary.ca

Martin Ritter
University of Calgary
Calgary, Alberta, Canada
martin.ritter@ucalgary.ca

ABSTRACT

This paper examines the use of motion capture to control ambisonic spatialization and sound diffusion parameters in real time. The authors use several software programs, developed in MAX, to facilitate the gestural control of spatialization. Motion tracking systems using cameras and peripheral devices such as the Leap Motion are explored as viable and expressive means to provide sound localization. This enables the performer to therefore use movement through personal space to control the placement of the sound in a larger performance environment. Three works are discussed, each using a different method: an approach derived from sound diffusion practices, an approach using sonification, and an approach in which the gestures controlling the spatialization are part of the drama of the work. These approaches marry two of the most important research trajectories of the performer practice of electroacoustic and computer music; the geographical dislocation between the sound source and the actual, perceived sound, and the dislocation of physical causality to the sound.

1. INTRODUCTION

This paper explores live spatialization in three current implementations, suggesting that motion capture is an effective and appropriate means for incorporating gestural sound spatialization into electroacoustic and computer music. For the purposes of this paper, the term gesture will refer to physical actions produced by the performer that have some basic trajectory and emotional intent. Such gestural control of spatialization has been explored quite a bit throughout electroacoustic music history; one could look to early efforts such as Pierre Schaeffer’s potentiomètre d’espace as precedents for gestural control of sound [1], with more recent historical applications including gestural controllers such as Michel Waiswicz’s Hands, and advancing to include multi-touch interfaces and further gestural control [2, 3]. Needs for control over spatialization continues to increase as concert halls accommodate more speakers and more complex speaker arrays. There are two primary ways sound can be perceptibly spatialized, with variances existing between each and on a continuum. The first is that of sound diffusion, a technique originating from the GRM involving the placement of a sound within a multi-channel system [4]. The second concept is sound choreography, which involves spatial trajectories and placements existing as a primary component of the composition [5]. The term “sound choreography” was firmly established at IEM in Graz during a 2010-14 research project entitled “The Choreography of Sound”, and is therefore quite new. However, distinction between diffusion and choreography of sound is essentially that one uses the alteration of audio signals to create an illusion of sound movement, whereas the other actually focuses on using a digital mechanism to localize the sound in space, and then send it to a speaker. Sound diffusion has been primarily associated with acousmatic music performance practice, whereas sound choreography is more associated with live performance of electronic music (although it could apply to acousmatic music as well). This paper explores the gestural control of spatialization using both sound diffusion and sound choreography, as well as a third approach that lies some-where between the two.

Gestural control of spatialization has been executed many different ways before; in fact, one could argue even that the movement of a fader is a performed gestural action. However, for the purposes of this paper, the term gestural control will apply to those systems, which have been developed specifically with gesture of a performer in mind as the primary driving force between data retrieval and output. Jan Schacher, for example, undertook extensive research on gestural control of 3D sound, which he discussed at length in his 2007 paper “Gestural Control of Sounds in 3D Space”. [6] describes a group of modules that can be used control sound in a 3D environment using gesture. However, what differentiates this system from that of the authors is that physical interfaces are used to track gesture in Schacher’s software, and that each of the systems described in this paper was designed, at least originally, with a very specific performance practice intention, unlike the generalized systems he discusses. Schacher distinguishes two modes of interaction, top-down and bottom-up. The former mode involves the performer having direct control over the properties of the sound, whereas the latter enables the performer to interact with a sound that has its own physical properties within the virtual space. The systems discussed within this paper explore both structures described by Schacher, the first system using a bottom-up approach, and the final a top-down.

Prior research has also been conducted regarding the perception of spatialization relating to gesture. Marentakis
and McAdams studied the effects of gestural spatialization on the audience and found that visual cues improve perception of spatial trajectories, but that such cues result in audience members focusing more on the visual stimuli than the aural [7]. This could be considered a negative for the use of gestural control, especially for music in which the intention is to strip visual stimuli and focus as much as possible on the sound. However, it is a force that could also be harnessed for dramatic or narrative intent, either by increasing the focus on the visual with gestures, or by obscuring the focus on the visual by removing gestures. This paper focuses on using gestures to spatialize sound that have (deliberately) varying degrees of visibility from the audience’s perspective. As long as the intent is clear, all are acceptable as meaningful performance techniques.

2. AAAMBI SOFTWARE

All of the approaches that are explored in this paper involve the use of a software module, developed by the author in Max [8, 9], designed to interface with the ambisonics software developed at ICSF by Jan C. Schacher and Phillippe Kocher [10]. This software, AAAMbi, provides a simple and intuitive graphical user interface intended for use by composers and performers. It allows the user to easily localize sound using both Cartesian and polar coordinates by connecting modules that are designed to send and receive messages from each other. In addition, users can group points, and automate panning features with these easy to use interfaces. AAAMbi has a modular and infinitely expandable design. The user is allowed to specify the number of inputs and outputs to the bpatcher as arguments. This number can be locked and the bpatcher embedded so that connections between the module and other objects are not lost. There is also a user friendly edit panel which can be used to specify speaker configurations, including the azimuth, elevation, and distance of each speaker output. These configurations can be saved and recalled later, which is extremely useful for users who need to rehearse and perform in multiple spaces. This simple dynamic control of input and output arrangement was the original intent of AAAMbi, and the early modules were basic in design and very general in use. However, as new works were created and differing spatialization needs arose, new software modules were created to connect with the AAAMbi software. Three such implementations are discussed below.

3. LIVE DIFFUSION USING MOTION TRACKING

The first approach discussed in this paper involves the live diffusion of a stereo work using computer vision tracking in Max. Jean-Marc Pelletier’s Computer Vision toolbox objects for Jitter [11] were developed into specialized tracking modules by the author. These tracking modules are easily linked and serve many functions; for the purposes of the diffusion system, the authors use modules that track bounding rectangles of movement within specified areas; the derived data is then used to determine spatial location and speed of movement. This diffusion software was designed for the performance of the specific piece discussed below, but can be used to diffuse any stereo file, if desired.

3.1 Why motion tracking?

With so many easy and accessible systems available, including analog and digital mixers and other devices, it is potentially questionable why one would use motion tracking to diffuse a piece in real time. While there has been some research involving the creation of new diffusion systems that provide a more feasible environment when numerous faders are required, such as the M2 developed by James Mooney, Adrian Moore, and James Moore in Sheffield [12], even such newer systems use a model based on the fader system rather than one on gestural control. However, past gestural controllers required large setup and could potentially take up much space, only adding difficulty to the diffusion process and detracting from the purpose of listening intently. The choice to diffuse this piece using motion arises from research into the performance aesthetics of electroacoustic music, and the particular pieces of hardware used (cameras, Leap motion) were selected based on their accessibility; both are relatively cheap, easily available anywhere, and have regularly updated and improved software and hardware interfaces.

Marko Ciciliani, in his 2014 paper “Towards an Aesthetic of Electronic-Music performance practice” [13], described two models of performance aesthetics, which he termed centripetal and centrifugal. These models are based on the measurement of how visible and central the performing body is to the sound source. Centripetal models consist of those in which the performer is the focal centre and whose movements are very perceptibly linked to the sound. Centrifugal performance, on the other hand, includes works in which the sound enacting body is removed or obscured from vision. Acousmatic works, such as the author’s City of Marbles, are by their very nature centrifugal if using Ciciliani’s classifications; the sound sources are intended to be unseen by the audience, and traditional diffusion performance would place the composer or other diffusing body hidden from view behind a mixer in a dark room. Using visible gestural diffusion in an acousmatic work marks a big change to the performance practice; the sounds themselves still have no visible source, but a human performer is present, centrally located, and present during performance. For City of Marbles, this is a programmatic intention. One of the primary sources of the work is a voice speaking the Latin phrase: “Marmoream se relinquere, quam latericiam, accepisset,” a quote by Augustus Caesar that translates into “I found Rome a city of bricks and left it a city of marbles.”
Additional recordings consist of a marble being dropped into various objects, and cicadas captured in Rome at the Palatine hill. The role of the sound diffuser is intended to be central and imposing because the quote itself infers creation and manipulation of several sources by a singular body. However, the agents that produce the actual sound sources are not visible during the performance. Performance of City of Marbles, therefore, uses visible gestural control of spatialization to both foreground the link between gesture and sound, and the obvious lack of visibility of the sound producing agent.

3.2 Tracking of motion to diffuse sound

As mentioned earlier, this system uses computer vision in Max to track the movement of a performer, which translates the data into messages for AAAmbi. The camera outputs a frame, which is then inverted on the horizontal axis (for the purposes of easier visual feedback for the performer) and split into ten separate matrices to be separately analyzed. The first four matrices correspond to faders controlling the output to channels 1, 3, 5, and 7, and the final four the output to channels 2, 4, 6, and 8. The matrices in the middle are not used; the division into ten components allows the body to be cut from the processed image, preventing unwanted fader movement. Two implementations from the MR.jit.toolbox [14] are used to analyze the motion: MR.jit.motionComplex and MR.jit.bounds. MR.jit.motionComplex is used to obtain general information about the amount of motion between successive video frames. It allows the user to specify motion smoothing, binary thresholding, morphological operations, and motion thresholding. The output consists of a binary frame, which can be used for further analysis (see below) and the current total active pixel count. The analyzed frame is then sent to MR.jit.bounds, which calculates a region of motion. These values are used to calculate the area of the motion, horizontal size, vertical size, and velocity and direction of both horizontal and vertical values.

The output is then used to move the sound sources around the performance space. The horizontal parameters are analyzed successively to capture delta values, and these delta values are subsequently used to control the Cartesian coordinates of the sound in space. Once the delta value exceeds a certain threshold, gates are opened within Max, which allow the delta value to control the “Y” parameter on a Cartesian plane, and the X (horizontal) value itself to control the “X” value on a Cartesian plane. The vertical movement data is analyzed in a similar manner, and the resulting delta values control the raising and lowering of virtual faders in the software. These faders are selected using the matrix splitting process described above. Therefore, the performer’s hand location in horizontal space selects which fader to affect, and vertical hand/arm movement subsequently modifies this fader value. A high vertical velocity with an upward motion, for example, will result in a strong increase of gain for the fader(s) selected by horizontal positioning. This gesture control system therefore spatializes the sound in two ways: 1) the input sources are moved using horizontal movement data, and 2) the perceived location of sounds is affected by using the vertical movement data to increase or decrease the volume that is output to specific speakers. This second method is much like the original approaches to panning, or the diffusion technique. Since it actually takes quite a bit of motion to trigger the movement of the sounds, the spatialization primarily uses the raising and lowering of virtual faders to diffuse the sound, with some areas that allow for input sources to be moved for an emphatic effect.

3.3 Perception – visual cues in acousmatic music

This piece was initially described as an acousmatic work, and it is such when performed in a dark room on two stereo loudspeakers (or many loudspeakers in stereo pairs). However, when performed with the diffusion software, there is a visual performance element present that changes the performance aesthetics of the work. Gestural control of spatialization has been proven to enhance the audience’s awareness of spatial trajectories (as mentioned briefly above), but to also decrease their focus on the sound itself. This runs contrary to the philosophy of acousmatic music, which is derived from the Pythagorean concept of listening to a lecturer behind a wall to focus solely on the sound (this is also where acousmatic music gained its name). City of Marbles, however, uses this live diffusion to illuminate a programmatic element: the narrative of the piece is about an individual creating something (in this case a city), with all of the development controlled by this being. Therefore, the diffuser in this case acts as metaphor for Augustus Caesar, controlling all of these elements to “create” a “city of marble”. The lack of visibility of the sound sources is also indicative of the way that great civilizations and cities are created; the “workers” are essentially unseen throughout history, with the ruler of the civilization essentially credited for growth and construction.

4. PIANO HAMMER TRACKING

The second approach to spatialization discussed in this paper involves using a camera to track the mechanical movement of the hammers inside a piano, and software to determine the amount and location of hammer movement. This data is then transmitted to a module called AAAmbiPianoHammers, which translates the hammer movement information to messages that can be sent to AAAmbi, ultimately altering the location of the sound.

4.1 Precedents -IMuSE

The Integrated Multimodal Score-following Environment (IMuSE) was a project at the University of British Columbia, Vancouver, Canada [15, 16]. The system was primarily designed to aid in the rehearsal and performance of score-based interactive computer music compositions. IMuSE incorporates several different software components such as NoteAbilityPro [17, 18] for the notated score (both the traditional score for the performer as well as the score for the electronics) and Max/MSP or pd [19] for the performance of the computer generated sounds as well as the analysis, matching, and networking of the tracking data to NoteAbilityPro. In order to track and score-follow a piece of music, a performance of said piece has to be recorded and the tracking information obtained aligned to the score in NoteAbilityPro.
Tracking data may include any discretely sampled data and the system has been tested with information obtained from accelerometer data, pitch tracking information, amplitude envelopes, spectral information (e.g. spectral brightness), and visual tracking using cameras, the Kinect [20] and the LeapMotion device [21–23]. A variety of instruments have been studied for tracking during and after the project’s official timeframe, which include: Viola, Cello, Clarinet, Trombone, Piano, and Accordion.

4.2 Tracking of Hammers

As mentioned above, the authors’ approach in tracking the hammer movement is to capture the mechanical action of the instrument, rather than the physical action of the performer itself. This is a direct extension of the research carried out during the piano portion of IMuSE, where auxiliary movements by the performer would on occasion interfere with the tracking as the head or shoulders would enter the tracking area. Capturing the mechanical action inside the piano can eliminate such obstructions. The current version of the software allows the user to crop and rotate the incoming video stream before it is analyzed. This dramatically increases CPU efficiency, as large portions of the video frame do not include any information of importance. This cropped frame is then subdivided into customizable tracking regions. One region per hammer is the default configuration. Each region is then analyzed for movement and if a (user-) defined threshold is exceeded, the region is marked as active, meaning a note was likely played. A list of ON-notes is compiled for each analysis frame and sent to Max for use in the above described software modules.

![Image](image1)

Figure 2: Top: Cropped and rotated image with superimposed movement data (white); Bottom: tracking region divided into 44 discrete areas with currently tracked region in black.

![Image](image2)

Figure 3: Customizable tracking regions. (a) = 3 active tracking regions; (b) = 45 active tracking regions;

4.3 Application Purpose

Tracking the hammers of a piano and subsequently applying this data to spatialization mappings allows for two primary functions: 1) the frequency space of the piano (i.e. low to high) is translated into spatial trajectories, which enables a perceptible aural correlation between the pitch and spatial location, and 2) tracking inside of the piano removes extra-musical performance gestures, such as seat adjustment, expressive movement, and page-turning, from the motion tracking data. This in turn makes the data more stable and strictly connected to the piano’s sounding results. This was the initial intention of the software, which was designed for use in a piano trio by the author that used performance practice as a compositional technique [24]. The piano trio contains motion tracking of all of the instruments, but the tracking data derived from the violin and cello performance is used to modify audio effects. The tracking data from AAAmbiPianoHammers is used to spatialize the piano sounds, and also sounds of prerecorded bells that play back. This is an attempt to connect the bell sound source, which is not visible, with tracking of a mechanism that is not visible (the hammers), but occurs as the result of a visible action (the performer playing the piano). The difference in the relationship between gesture and sound production of a cello and a piano is also pertinent to the choice of hammer tracking. The gesture a cellist makes, for example, occurs right at the physical location that produces sound; a bow is dragged across a string, which then vibrates to produce sound. There is no disconnect between the bow and the hand. A piano is somewhat different in that the performer strikes keys, which are then used as levers to enact hammers to strike strings. There is a visual disconnect between gesture and sound; the frame of the piano obstructs the hammers visually. What we are seeing relates to the sound, but not exactly. If a cellist makes micro-movements in his hand as a result of nervousness, this will translate into the sound, as the bow will make micro-movements. This is not the case for the piano, and this presents another reason why the authors have chosen to implement the software in this way.

5. USING THE LEAP MOTION TO CREATE A GESTURAL NARRATIVE

Finally, the use of dramatic and/or programmatic gestural movement to control spatialization is discussed. Fayum Fragments, part of a larger poly-work by the author for soprano, flute, cello, piano, and electronics, uses a Leap Motion device to capture the gestures of the vocalist. This gestural interaction is primarily used throughout the work to determine the overall form and narrative, as the structure of the work is aleatoric and dependent on the singer’s gestures to advance each section. The Leap Motion also serves the purpose of triggering and spatializing sounds, and it is this particular use of the Leap Motion that will be discussed here.

5.1 Background – MRleap

The MRleap object [25] serves as an interface between the Leap Motion USB peripheral device and Max/MSP. It was created specifically to give the user the ability to precisely
enable and disable the device’s over 90 different gestural data acquisition options. This is extremely useful to performers and composers as it allows them to choose which data streams to apply to audio, visual, or control parameters and avoid unwanted CPU usage by disabling unwanted functionality.

5.2 Using a glyph system for gestural notation

*Fayum Fragments* uses a system of twelve graphic glyphs the vocalist must interpret with her hands, which loosely translate into gestural movements with certain emotional intent. This is the system the vocalist uses to interface with the Leap Motion throughout the work. The Leap primarily controls the playback of sound files during performance; motion triggers the onset of a sound, and the shape of the motion affects the parameters of the sound as a continuous control. The spatialization of these sounds works in much of the same way, with the starting location of each sound selected and distributed immediately when triggered. The movement shape determines small changes in spatialization, such as the amount of azimuth distance between sources, and the distance of the sources from the centre. Additional processes, such as filters, are added for subtle localization cues.

![Figure 4: Glyph table showing glyphs and their meanings.](image)

5.3 Gestural data and spatialization

There are only a few parameters of the Leap Motion that are translated into sonic control information during the *Fayum Fragments*: X, Y, and Z-positions of the hands, and the velocity of movement of the hands. This provides a very simple interface, but with effective musical results. The velocity parameter is used to trigger the sample, and as a result, triggers a randomly selected starting location as a centre point. The continuous X value then controls the degree of circular spread between the sources by modifying the distance value between them. The continuous Z control affects distance from the centre, and Y is used to add a very subtle low pass filter for distance cues. The reason that the trajectory to the starting location is immediate, giving the initial idea of a more stationary localization model, is due to the nature of the samples themselves, which consist of spoken and sung Greek word. These vocal sounds are much more effective and believable when triggered at stationary locations because humans are generally relatively stationary from the listener’s perspective when they are speaking. However, once the sounds are triggered, the shape of the gesture allows for very small adjustments in movement. These minor changes in spatialization that coincide with the gestural shape create subtle sonic effects, however, they provide an effective and meaningful link between gesture, spatialization, and sound processing.

6. CONCLUSION

Sound diffusion and choreography using gesture are not novel concepts; systems have been put in place to achieve both, although the gestural control attempts have primarily used physical controllers and sensors using knobs and sliders. Some of the constraints of using such gestural systems purely for diffusion, such as lengthy setup and bulky extra material, are removed by using motion control systems that do not require external hardware controllers. The drawback to using such systems is that motion capture does not necessarily contain the amount of fine control provided by physical sensors that track gesture. 3d tracking systems could be implemented for finer control of parameters, and area consideration for future development. However, for the purposes described in this paper, which involve very general spatialization for mostly artistic and dramatic effect, the tracking of motion using these systems has been effective. Additionally, as the works described in the paper are all concert works intended for multiple performance, accessibility of technology was a large consideration. Thus far, all of the systems have been tested successfully, from technical and aesthetic perspectives, in a controlled lab environment. The diffusion system will be tested in a live performance summer 2016, and the other two systems will be tested in live performance and performance workshops throughout late 2016, culminating in my thesis performance in early 2017.

7. REFERENCES


