INTERNATIONAL CONFERENCE ON LIVE INTERFACES

Towards a Live Interface for Direct Manipulation of Spatial Audio

Jamie Bullock[1], Tychonas Michailidis[2], and Matthieu Poyade[3]

[1] Birmingham Conservatoire, Birmingham, UK
jamie.bullock@bcu.ac.uk
[2] Birmingham Conservatoire, Birmingham, UK
tychonas.michailidis@bcu.ac.uk
[3] Glasgow School of Art, Glasgow, UK
m.poyade@gsa.ac.uk

Abstract In this position paper we present research from the AHRC-funded project Transforming Transformation: 3D Models for Interactive Sound Design. The project entails exploration of a new interface for live audio transformation whereby sound can be manipulated through grasp as if it were an invisible 3D object. This approach contrasts with existing GUI-based systems, which primarily use 2D input and 2D visualisation. In the paper we describe the first phase of this research, which enables audio sources to be positioned and moved within a 3D space by grabbing them from a palette and controlling their spatial location using hand movements. Feedback regarding spatial location is provided through a visualisation of these sources within a virtual 3D space. Spatial trajectories can be 'drawn' in the air, and sounds can be 'rolled' along these trajectories thus providing a 'direct manipulation' interface to specifying spatio-temporal dynamics. We describe the design of the system along with findings of the initial system usability tests.

Keywords: 3D; sound design; spatial audio; live; direct manipulation.

Introduction

Existing studies show a requirement for more usable and human-centred audio processing systems for live performance and highlight software design deficiencies as a barrier to musical creativity (Bullock, Beattie, and Turner 2011; Magnusson and Mendieta 2007). In response to such a requirement, we have set ourselves the ambitious goal of rethinking audio transformation such that sounds can be manipulated live, as though they are tangible objects suspended in the air. We imagine a system where a user can sit in front of an invisible, 3D space (e.g. a $1m^2$ cube), 'containing' a number of sounds that can be touched, grabbed, stretched, broken, and thrown. In terms of spatialisation, we anticipate the ability to 'throw' such sounds with inertia such that they 'bounce' off the sides of a virtual acoustic space until their mass slows them down. Our aim is thus to create a sound processing environment, the visual and interactive characteristics of which correspond more closely with 'everyday listening' (Gaver 1993) than with typical technology-centric approaches found in widely-used Digital Audio Workstation (DAW) and plugin-based environments. In this paper we describe the first stage in our research towards this goal: the representation and interaction with sounds in a virtual space such that they can spatially positioned in 3D with corresponding acoustic spatialisation using a multi-channel speaker system or binaural headphones. Our aim was thus to provide an output-independent (same 3D environment for multiple underlying spatialisation technologies) interface for 3D positional audio that feels live and direct to the user with high levels of usability and a strongly positive user experience. This research is aimed to be applicable across a wide range of application domains such as composition, live performance of new music, sound design, studio production and game audio. Furthermore, we argue below that whilst 'liveness' has performative implications is also an experiential quality of interfaces, that is interfaces can per se be more or less 'live'.

To illustrate this, we begin with the related 'direct manipulation' paradigm originally proposed by Shneiderman for interfaces that have the following properties:

1. Continuous representation of the object of interest
2. Physical actions or labeled button presses instead of complex syntax

3. Rapid incremental reversible operations whose impact on the object of interest is immediately visible (Shneiderman 1982, 251)

Norman and Draper subsequently build upon and problematise Shneiderman's concept by suggesting that *Direct Manipulation* is not a unitary concept nor even something that can be quantified in itself. It is an orienting notion. 'Directness' is an impression or a feeling about an interface capable of being described in terms of concrete actions that can be taken by system designers to make interfaces more-or-less direct (Norman and Draper 1986, 93). Norman and Draper propose two aspects of directness that can be designed into systems: distance, and engagement. 'Distance' pertains to the 'cognitive gap' between a system's representation of a task or environment and the human user's representation. 'Engagement' pertains to the 'qualitative feeling of engagement, the feeling that one is directly manipulating the objects of interest' and where 'the world of interest is explicitly represented and there is no intermediary between user and world' (Norman and Draper 1986, 94). We argue that both of these aspects of direct manipulation are, within the context of interfaces for audio transformation, key indicators of a highly positive user experience. A full discussion on direct manipulation and how it relates to user experience (UX) is beyond the scope of this paper, however, of particular relevance is the relationship between 'directness' and 'liveness' as set out by , which builds further on Schneiderman's framework:

> *Liveness* means the user interface is always active and reactive - objects respond to user actions, animations run, layout happens, and information displays are updated continuously. Directness and liveness are properties of the physical world: to examine and change a physical object, you manipulate it directly while the laws of physics continue to operate (Maloney and Smith 1995).

Our use of the word here 'live', corresponds to that of Maloney and thus differs somewhat from (although does not exclude) 'live' as in 'live performance' or 'live electronics', in that it refers to the intended liveness of the user interface as enabled through direct and pliable manipulation of the objects (in this case sounds) being interacted with.

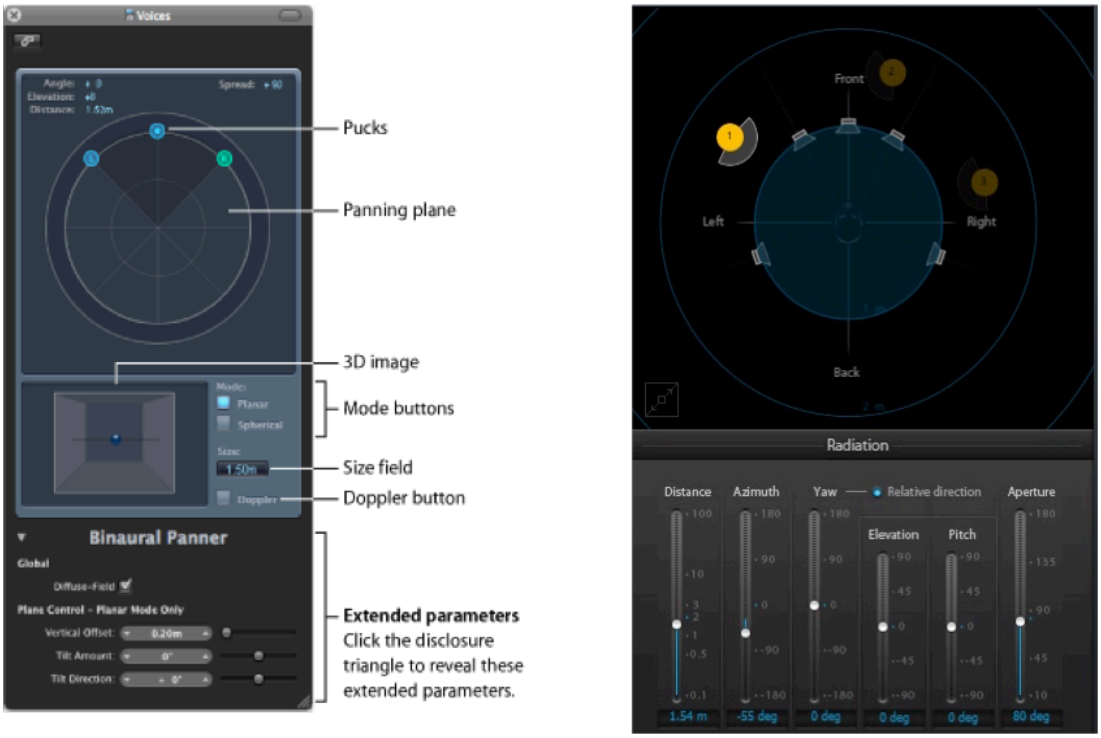

Figure 1. 2D approaches to spatialisaton UI: Logic Pro X's binaural panner (left) and Flux Spat plugin (right)

This contrasts with commonly-used commercial interfaces for audio processing in this case, spatial positioning. The use of a 2D input device such as a mouse, trackpad or joystick coupled with a 2D (single-plane) or quasi-3D visualisation

immediately creates 'cognitive distance' for users by forcing them to internally 'map' between a 2D interaction model and 3D acoustic output. A typical example of this approach is shown in figure 1. In such interfaces, the user moves a small 'puck' representing a sound source within a circle (through the horizontal panning plane) or sphere (through a 3D panning surface) by moving the cursor. As the user drags the puck in physical 2D space, the sound's apparent location within the 3D virtual acoustic space changes. For tasks, such as simulating the sound of a car moving past the listener, a 2D interface may be sufficient. However, more generally, a 2D interaction model does not sufficiently mirror the human perception (or internalisation) of sounds occupying 3D space and having 'volume'[1] (Carello, Anderson, and Kunkler-Peck 1998). Rather, sounds are represented as point-sources, with the user positioned and controls from 'above' the spatial plane. Such interfaces are also problematic because they do not provide a uniform representation of sounds and means of transformation. That is, in a plugin-based environment, spatial position is usually controlled through one part of interface (the plugin-type UI), volume through another (a mixer UI). A DAW or typical live audio environment automations of spatial parameters over time are edited through yet another interface; typically using continuous visual breakpoint functions, or 'envelopes', with one-envelope-per-dimension for spatial position, thus extending 'cognitive distance' further by reifying spatial location into discrete geometric parameters. To do something as seemingly simple as position of a sound in space, lower its volume, create a spatial trajectory, edit the trajectory, and automate the sound's motion along that trajectory, users must navigate multiple visual representations, metaphors, and levels of abstraction creating additional cognitive burden and disjointed workflow (Norman and Draper 1986; Maloney and Smith 1995). We argue through the data presented in this paper that such an approach creates a 'disembodied' UX where aspects of the workflow disconnect the body's sense of sound and movement from its representation through the interface and interaction design (Dourish 2001). This is particularly important in live performance where the performer's actions are not only functional but can signify process and intent to the audience.

The requirement for an alternative approach is corroborated by existing studies, for example Peters et. al. conducted a survey with 52 respondents in which they asked a range of questions regarding technologies and practices relating to spatialisation (Peters, Marentakis, and McAdams 2011). Respondents were primarily male (85%) from Europe and North America, university-educated with an overall average compositional experience of 20 years, 10 years of which involved spatialisation. Some key findings of relevance to our research are shown below:

- Sixty-one percent of respondents thought the time spent with spatialization tools could be reduced
- Half of the respondents use fewer features than their spatialization tools offer
- 48% cited "usability, learning curve" as motivation for using their current setup

One of their recommendations based on their findings is that 'the learning curve must be kept reasonably shallow (i.e., gradual) with good usability (e.g avoiding cumbersome command line control)'. Perhaps surprisingly, 'Visual 3D representation of a sound scene' was rated by participants as having lowest importance (less than 'fairly' important) amongst technical features. Despite this, we maintain that a 3D visual representation and interaction model is key to achieving high levels of usability and UX in spatial audio systems, and seek to explore this through empirical tests.

Existing Work

Our system builds upon existing approaches to 3D interaction in spatial audio processing and sound design. The idea of accessing temporal audio data through the mapping of a spatial interface was explored by Kobayashi and Schmandt (1997). The 'point-by-hand' interface provides users with the illusion of touching audio as 'objects' which increases the experience of directness within the virtual 3D space. However, the described system did not benefit from modern interactive technologies. It included head mounted sensors to track position and provided audible feedback notifications (of increasing loudness) to inform the user of their successful grabbing action. The 'un-grabbed' function required users to 'stand still' for 3 seconds. Our research takes as its starting point the classification approaches in the field of 3D interaction techniques, *Navigation*, *Selection and Manipulation* and *System Control* as described in Jankowski and Hachet (2013). Jankowski suggests that to avoid frustration amongst novice users for 3D interaction, the system should

[1] We use the word 'volume' here in the physical, not the acoustic sense.

adapt to the user's expertise and experience. Jankowski goes on to suggest that an interface with strong appeal and enjoyment factors will motivate users to perform 3D tasks. A 3D interface offers the possibility to capitalise on users' already-learned gestures such as grasp, touch and placement. This could be regarded as an application of the direct manipulation interaction paradigm described in the introduction. It also relates strongly to Löwgren's notion of interface pliability (Löwgren 2009).

This approach is further explored by Gelineck and Korsgaard (2015) who devised a system to evaluate user interfaces for 3D audio mixing. Of a particular interest is how users, in the context of 3D audio, prefer to use 3D tools e.g to grasp and position sound through hand-air gestures, rather using buttons to grasp and release sound as is currently used with Fairlight 3DAW.[2] Our system differs from and builds upon the work of Gelineck and Korsgaard in that it integrates the user within the immersive 3D environment, providing the ability to use both hands for gesture control over audio properties (for example using one hand for selecting a source via grasp, and using the other hand for manipulations such as volume change and drawing spatial trajectories) rather than imposing on hardware controls, to which the user must adapt. On a more fundamental level, whilst our proposed system builds on the virtual representation used by Gelineck and Korsgaard and , it differs in that its scope is significantly wider, seeking application not only audio mixing but any musical application requiring 3D sound spatialisation—and in the longer term a diversity of methods of audio transformation. Comajuncosas et. al. also describe a system whereby a Kinect interface is used to capture 3D gestures, which are used to perform real-time 'audio mosaicing' a process by which tiny audio fragments are concatenated according to their location on a timbre map (Comajuncosas et al. 2011). Of particular interest in this work are the perceived benefits of 3D interaction over 2D (via a mouse) in achieving 'an enhanced correlation between gestures and musical outcome [...] and increased performability' with in-air gestures 'amplifying the perception of the player manipulations'. These approaches in particular will become of increasing interest as our research progresses beyond spatialisation to other forms of direct and live sound transformation.

Prototype System Design

Our system utilises TwoBigEars 3Dception as a spatial audio engine. 3Dception is capable of rendering a set of 3D positional audio co-ordinates for given sources to binaural, Dolby Surround or Ambisonic outputs. The binaural technique uses a psychoacoustic model to create the impression of sounds originating within a 3D space using conventional stereo headphones, for example a sound can appear to be coming from above or behind the listener. Our system therefore goes beyond existing work in not only providing 3D panning (which essentially moves sound around the surface of a sphere, see Churnside, Pike, and Leonard (2011)), but also acoustic simulation including room reflections and object occlusion. The initial system (figure 2) is implemented in Unity[3], a cross-platform game engine suited to 3D graphics and interaction. A Microsoft Kinect 2[4] is used as a motion capture device in order to detect the position of the user's hands as well their hand pose. The Kinect 2 was chosen as an initial input device due to the hand pose recognition built into the SDK, therefore enabling rapid development within the short timescale of our funded project. Poses currently used are 'hand closed', 'hand open' and 'index finger'. The user's hand centre position is captured by the Kinect and translated into a point within the co-ordinate system of a virtual 3D space so it can be visualized within a virtual environment (figure 2). This visual display represents a space into which audio sources (represented as spheres) can be positioned. The centre point of each sphere controls its positional audio location, which can be rendered in real-time through headphones or to a multi-channel speaker system. For example, if the user moves a sphere from left to right in the virtual space, it will simultaneously be localised from left to right in the playback system.

[2] http://www.fairlight.com.au/product/3daw/
[3] https://unity3d.com
[4] https://dev.windows.com/en-us/kinect/develop

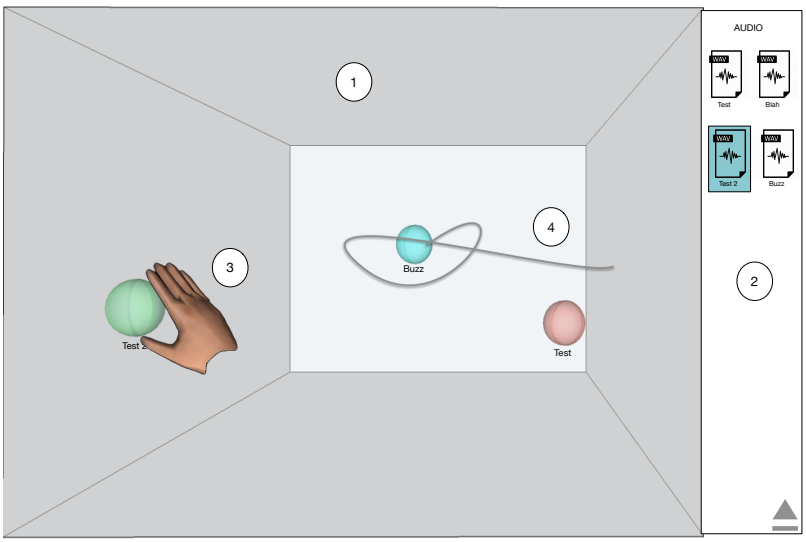

Figure 2. Original mockup of visual feedback for the system including 1) a representation of the virtual acoustic space; 2) a palette of sources to placed in the space via drag-and-drop; 3) an avatar representation of the user's interaction with the sources; and 4) a visual representation of spatial trajectories drawn by the user.

The visual display provides a cue to the user enabling them to accurately position their hand within the virtual space (using an avatar), whilst also providing a visualisation of the relative locations of the audio sources. A palette of audio sources is displayed to the user in a 2D panel on the right-hand side of the screen (labelled '2' in figure 2), where a 'source' can be either a soundfile or a live source from the computer's audio interface or built-in microphone. When the user's physical hand position is placed such that their hand position within the visual display is in front of an audio source, a 'hand closed' pose will cause the corresponding audio source to be 'picked up'. As the user moves their hand across into the virtual space, the audio source icon will change to a sphere and a 'hand open' pose will cause the audio source to be 'released' into the space. At this point, for soundfile sources, looped or non-looped audio playback will begin, with the spatialised position of the source corresponding to its co-ordinates within the space, with the listener's head effectively positioned in the centre of the space. Once an audio source is within the virtual 3D space, it can be picked up and moved around through a grab (hand closed) 'move' release (hand open) interaction. Finally, audio sources can be stopped by removing a source from the virtual space. This is achieved by grabbing a sphere and dragging it onto the 'eject' icon in the bottom-right corner of the audio source palette. Our system represents a 'spatially relative' model, that is sound sources are positioned in a fixed-size *virtual* space, with distances that are scaled up or down in relation to their actual perceived locations within a physical or binaural acoustic space. The system is also geometrically 'translated' in that although sound may appear (for example) to come from behind the user when spatialised, the bounded space with which the user interacts is always *in front* of the user. In practical terms, this means that in order to place a sound behind their head they place it in front of them, but behind the location of their head within the virtual space (i.e. behind the centre of the virtual space). This differs from the spatially 'absolute' model employed by , whereby sound sources can appear to be emitted from any point within a physical space, and users interact with the sounds directly at their perceived physical location. In a sense Müller et al's system is a more faithful implementation of the direct manipulation paradigm (see section introduction) than our design, but it is less general and less scalable. With our system, the sound could be output to binaural headphones or a large-scale ambisonic system and the interface remains consistent.

Changing Audio Source Loudness

In order to change the loudness level of a given audio source, the user's left hand can be used, employing an 'index finger' pose to raise and lower amplitude. When an 'index finger' pose is detected, the ability to control amplitude is indicated on the visual display next to a sphere using a 'slider' widget following a traditional mixing desk 'fader'

metaphor. As the user moves their finger up or down, the slider value will change causing a corresponding change in the level of the audio source. In future work alternative visual representations of audio amplitude will be considered.

Hyperreal physics and spatial trajectories

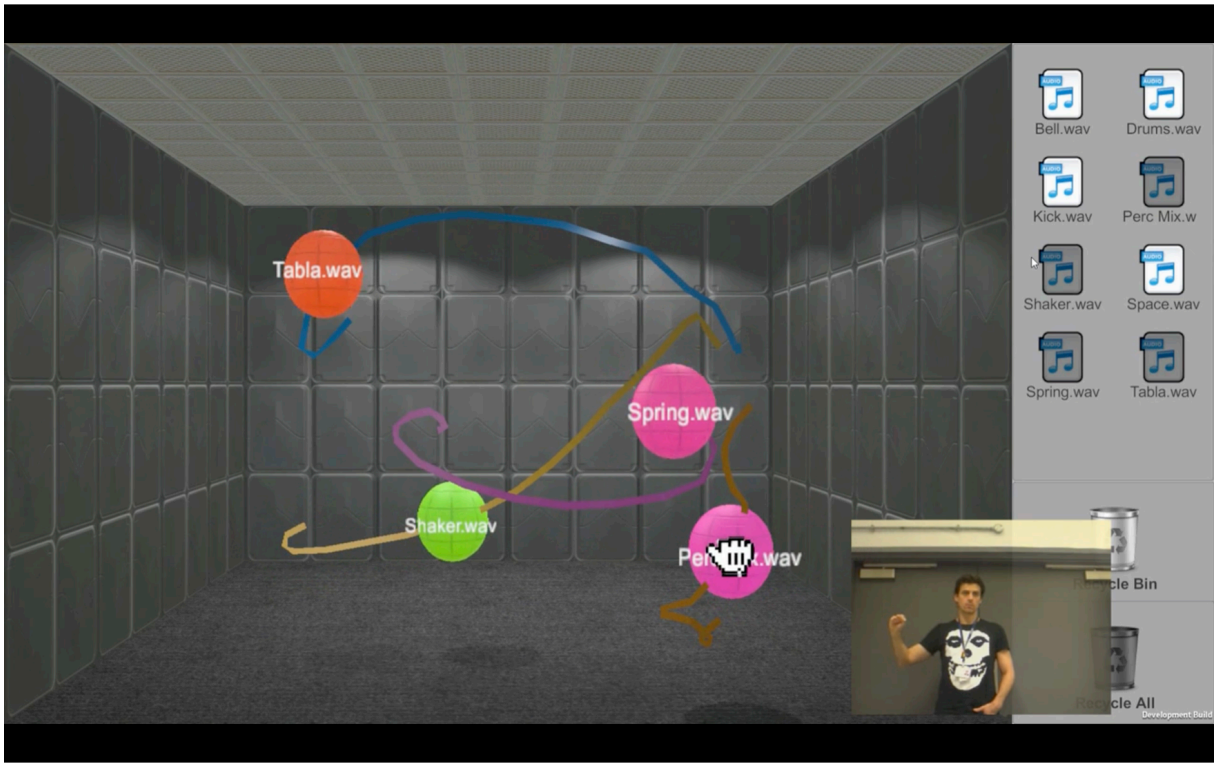

Figure 3. Screen capture showing the system in use with user input as picture-in-picture (bottom right).

Sources can also be given 'physical behaviours' within the virtual space. This includes the ability to 'throw' sounds, giving them inertia and thus moving them in the direction of the throw within the acoustic space at a rate determined by the intensity of the user's movement. Sounds given such inertia may also slow down due to 'friction', and 'bounce' off surfaces or other objects placed within the virtual space. To automate spatial movement, the ability to draw trajectories within the virtual space is provided. Such a trajectory is indicated by the coloured lines in figure 3, which sounds may then be thrown along (analogous to rolling a ball along a curved tube).

Initial Testing

Our eventual aim is to develop an end-user system that provides a positive and highly engaging user experience, surpassing that of traditional systems for spatial audio. In order to ensure this, we are undertaking an iterative process of testing and research. In the interests of rapid development initial tests were undertaken either by the authors, or using ad hoc 'hallway testing'. In performing these tests on consecutive iterations of the system, a number of issues were identified that could be rectified through small but important enhancements to the design. These included;

1. Changing the colour saturation of a sphere to visually indicate its relative volume level,
2. Greying out sources in the audio palette to show they have been added to the space,
3. Adding a visual glow to a sphere indicating proximity to another sphere, thus aiding 'same location' positioning.

However, a number of more fundamental issues were identified relating to the choice of Kinect 2 as an input device. For example, the Kinect specifications require that the user is at least 1.4m from the device. This has implications for screen positioning, size and resolution. For example, in our tests audio source names were impossible to read on a 27" display at a resolution of 2560 x 1440. Operating at the required distance from the screen also decreased the sense of user 'immersiveness' (Ermi and Mäyrä 2005). More critically, we found that the Kinect 2 is unreliable as the user's palm orientation deviates from the plane parallel to the front surface of the device. In our initial tests users began by reaching into the virtual space without any particular consideration for palm orientation. Reaching forwards to grab an object with the palm approximately horizontal is a physiologically comfortable movement (the anterior forearm muscles are relatively relaxed). However, interacting in this way resulted in a poor experience with a large number of 'grab' gestures not working due to the Kinect failing to disambiguate 'hand open' and 'hand closed' poses. This was because of the hand being too flat in relation to the horizontal plane. In order to improve system performance, we found users needed to adapt their usage by consciously orienting their hand as close as possible to the (vertical) plane parallel to the front of the Kinect. Grabbing an object in front of the user with palm facing 'forwards' in this way is physically more awkward, putting a strain on the anterior forearm muscles. It therefore feels less comfortable and consequently use of the Kinect poses a mismatch between a comfortable 'grab' movement capitalising on commonly used actions—with associated muscle memory (Ebert et al. 2009)—and the pose recognition capabilities of the device, creating a hard limit on directness and 'naturalness' of interaction. Finally, whilst we found that most users readily understood the interface and could place and move sounds (with the above caveat for hand angle), some users were confused when asked to place a sound 'behind them'. In these cases the users in question moved their hand behind their own head, rather placing it at the back of the virtual space in front of them.

Further Work

Our first priority is to conduct formal and detailed usability and user experience tests with the current system, to establish how well it meets our stated design goals using a combination of quantitative and qualitative methods. In the medium term will also seek address issues highlighted in the above findings namely: i) a lack of reliability and control in the user interaction, and ii) to clarify through the design that the acoustic space is in front of the user. On a practical level this will entail an investigation into alternative input devices, or alternative setups that address the palm orientation and proximity issues identified in our tests. Specifically, we aim to make the system operable from a seated position in close proximity to a screen and for the user to adopt a comfortable hand orientation. We will also seek to test better technologies for visualisation, these may include immersive displays and virtual reality headsets. Our research is currently at an early stage, but a future step will be to conduct further investigation into users' prior associations of gesture, action and visualisation with various forms of live audio transformation. Only with this understanding can we seek to align the representations of audio transformations provided in the system with human understanding and imagination. Such an understanding is a long-term task and will operate side-by-side with an iterative process of design, development and testing in which we will seek to implement a range of live audio transformations following a direct manipulation model. It is our hope that these will offer greatly enhanced learnability and a more positive user experience compared to existing systems.

Conclusions

We have presented the rationale, design and initial implementation of a novel system for spatial audio positioning utilising 3D interaction and visualisation of a virtual space. More generally, we have described how direct manipulation, 'live' and pliable interface design concepts may be widely applied to systems for interactive audio transformation opening up a new field of investigation in audio interface design. We have described our initial design and tests, which have shown that our proposed concept and system work, and indicate strong potential for further work. However our results indicate a number of limitations in implementation, specifically with the Kinect 2 as an input device for direct audio manipulation. In particular, there is a minimum distance requirement of 1.4 metres from the device. This means that the display used for visual feedback must be sufficiently large for users to perceive details, or alternatively other visual feedback such as a virtual reality system should be considered. Furthermore, hand poses can only be recognised if the user's hand is oriented around 60 degrees of the device's front. As part of the ongoing development of the system,

alternative sensors such as the LEAP motion are therefore being investigated. In conclusion, our research demonstrates a novel concept and provides promising implementation, serving as a strong foundation for further work.

Acknowledgements. We would like to thank the Arts and Humanities Research Council in the UK for funding this research under their Digital Transformations Small Grants scheme. We would also like to thank TwoBigEars for making 3Dception free for academic projects and supporting our work with it.

References

Bullock, Jamie, Daniel Beattie, and Jerome Turner. 2011. “Integra Live: A New Graphical User Interface for Live Electronic Music.” In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 387–92.

Carello, Claudia, Krista L Anderson, and Andrew J Kunkler-Peck. 1998. “Perception of Object Length by Sound.” *Psychological Science* 9 (3). SAGE Publications: 211–14.

Comajuncosas, Josep M, Alex Barrachina, John O’Connell, and Enric Guaus. 2011. “Nuvolet: 3D Gesture-Driven Collaborative Audio Mosaicing.” In *NIME*, 252–55.

Dourish, Paul. 2001. “Seeking a Foundation for Context-Aware Computing.” *Human–Computer Interaction* 16 (2-4). Taylor & Francis: 229–41.

Ebert, Achim, Matthias Deller, Daniel Steffen, and Matthias Heintz. 2009. “‘Where Did I Put That?’–Effectiveness of Kinesthetic Memory in Immersive Virtual Environments.” In *Universal Access in Human-Computer Interaction. Applications and Services*, 179–88. Springer.

Ermi, Laura, and Frans Mäyrä. 2005. “Fundamental Components of the Gameplay Experience: Analysing Immersion.” *Worlds in Play: International Perspectives on Digital Games Research* 37: 2.

Gaver, William W. 1993. “What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception.” *Ecological Psychology* 5 (1). Taylor & Francis: 1–29.

Löwgren, Jonas. 2009. “Toward an Articulation of Interaction Esthetics.” *New Review of Hypermedia and Multimedia* 15 (2): 129–46. doi:10.1080/13614560903117822.

Magnusson, Thor, and Enrike Hurtado Mendieta. 2007. “The Acoustic, the Digital and the Body: A Survey on Musical Instruments.” In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 94–99. ACM.

Maloney, John H, and Randall B Smith. 1995. “Directness and Liveness in the Morphic User Interface Construction Environment.” In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*, 21–28. ACM.

Norman, Donald A, and Stephen W Draper. 1986. “User Centered System Design.” *Hillsdale, NJ*.

Peters, Nils, Georgios Marentakis, and Stephen McAdams. 2011. “Current Technologies and Compositional Practices for Spatialization: A Qualitative and Quantitative Analysis.” *Computer Music Journal* 35 (1). MIT Press: 10–27.

Shneiderman, Ben. 1982. “The Future of Interactive Systems and the Emergence of Direct Manipulation†.” *Behaviour & Information Technology* 1 (3). Taylor & Francis: 237–56.