User interfaces for spatial audio production

Segrel Koskentausta
Aalto University
Master's Programme CCIS / AAT

sakari.koskentausta@aalto.fi

Abstract

Spatial audio poses new challenges for efficient work on producing audio experiences. There is an increased amount of high density loudspeaker setups and various home theater solutions, as well as binaural audio built into gaming and other entertainment. Commercially available spatial computing platforms, such as Meta's Oculus or Apple's Vision Pro call for immersive experiences where audio that adapts to movement in 6 degrees of freedom is needed. This review goes through some challenges in building the user interfaces that the users need. It also showcases some research on different user interaction paradigms in audio production, including mechanical, touch-based and gesture-based. Some existing spatial audio production tools and research prototypes are also presented.

1 Audio user interfaces

1.1 The challenge of designing user interfaces

Designing good user interfaces is a complex challenge. A good practice is to involve the target users in the design process, but it's often challenging to elicit valid feedback [1]. It's a common saying that the user needs to be given what they need, not what they claim to want. A good user interface allows for efficient execution of tasks and it's good to note that sometimes the visualizations provided may cause users to gravitate towards less favourable aesthetic results [2]. Therefore it is necessary to research the usability, effectiveness and the quality of results when designing audio user interfaces.

Dewey and Wakefield describe a design and evaluation process for musical user interfaces [3]. First, to be able to validate the usability, one has to define the target users and the tasks they're suppose to be able to accomplish with the interface. Then the designer has to choose good interaction styles and accompanying visual and methaphorical aids. Finally, it's feasibility has to be tested by prototyping and evaluating with users. This process carries a lot of risk and uncertainty and it's common to see products gradually converging and the design to evolve by iteration based on previous designs [4].

1.2 Audio user interface control schemes

A large chunk of audio user interface paradigms come from an analog era. The multi track tape brought with it a multitude of opportunities for audio production and by 1970s Solid State Logic mixers had polished the basic metaphors and control schemes for audio mixing [2]. A slider is by now an ubiquitous control for adjusting the level of audio and the knob often accompanies it for e.g. panning adjustment. The existence of buttons goes almost without saying and the use of a mouse, keyboard and joystick is also widespread in audio production.

The digital era brought with it some additional control schemes while copying the slider, the knob and the button along as familiar metaphors in skeuomorphic fashion. This lead to an evolution of mixing leaving the constraints of a physical room and turning more into a set of practices and processes [2]. Desired set of practices can now be conducted with one's choice of computers, tablet computers and separate controllers. The controllers are now often in themselves very basic, merely dumb interfaces communicating with a DAW via MIDI or OSC [1].

Multi-touch capable devices such as tablets have brought along a mouseless way to interact with controls, which makes some user interactions more efficient [5]. The past ten years has also brought a new rise in hand motion tracking and it's been explored for audio interfaces using movement tracking devices such as Kinect and Leap Motion [6] [7] [8]. The spatial headset interaction has lately converged towards gaze & pinch, but no publications were found to be exploring this specific control paradigm for audio interfaces as of yet. The paradigm should, in principle, improve the accuracy and comfort of gesture based user interaction [9].

It is also possible to combine some of these control schemes, for example by having a tangible knob-like widget that can be placed on a touch screen surface [10] [2]. Additionally, instead of just providing insight with visualizations and metaphors, the designed products can be "smarter" and provide suggestions about actions or even make independent decisions about tasks being performed [11]. This is in contrast to the standard way of the user adjusting the parameters such as panning and level separately one by one. There is also a rise in prompting, either through written or spoken language, which should not be discounted from viable options without further study. These systems have recently become vastly more versatile in the way they may be used for automating more complex tasks, described in enough detail in a natural language, like English.

1.3 Audio user interface paradigms

The classic mixing interface paradigm is the channel strip [10], where basic controls are duplicated along the channels and arranged in vertical strips. The still less used, but much explored paradigm is the stage metaphor. With stage metaphor one has

virtual sound sources placed in what could be called a stage, displaying a visual feedback on the placement of the sound sources relative to each other in a 2D plane. The difference between these two is depicted in Figure 1. The depth information can additionally be provided and used in a way not commonly found in channel strip paradigms, for example by adjusting the width of the sound as the source is moved closer or further [12] [10]. This already amounts to a "smart" user interface as it abstracts away a more complex set of actions instead of expecting the user to adjust each required parameter by hand.

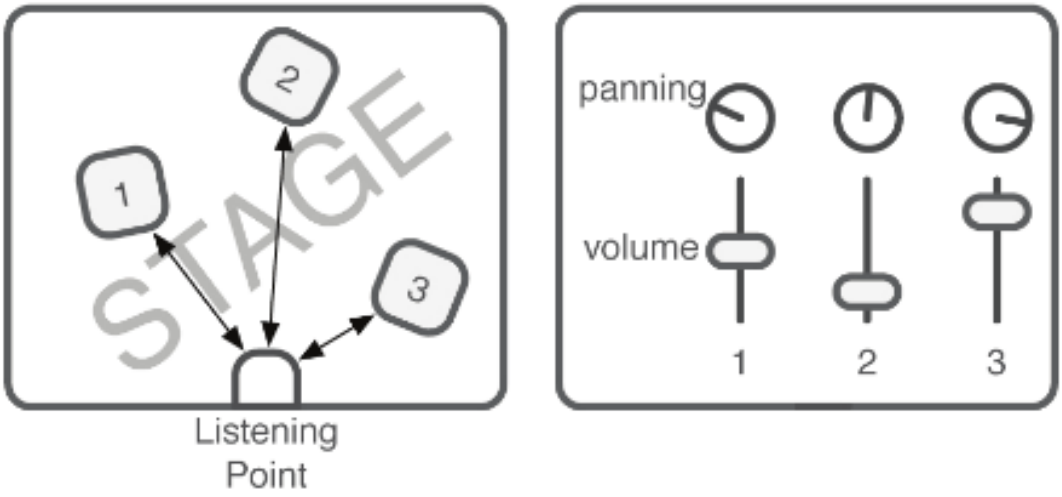

Figure 1: 2D stage metaphor (left) and channel-strip metaphor (right) depicted in a fixed listener position interface. In stage metaphor the volume and panning are defined by the distance and and angle relative to the listening position. Image from Gelineck and Uhrenholt (2016) [13].

While just about all Digital Audio Workstations are based on the channel strip metaphor [2], the stage metaphor may be more intuitive and assist with manipulating the spaciousness of the mix [14]. Several plugins targeting spatial use cases are also available for different DAWs, as explored in section 3. Different approaches to stage paradigm also exist, having different mappings to different axis. More research is likely needed in this front as no definitive mapping paradigm seems to exist yet. Combining the channel strip paradigm and stage paradigm is also possible, but little was found on this front in the existing research literature. At this time in the most common mapping of the stage metaphor interface, the X-axis represents panning (azimuth), Z-axis represents the level (distance) and Y-axis the brightness of the sound [2]. Different aspects may be assigned to different axis, however, and the appearance of the objects may be used to convey further information. Research has been conducted on providing useful and non-distracting feedback on the stage visualization and results point to carefully choosing useful data points to visualize in order to avoid distractions [13].

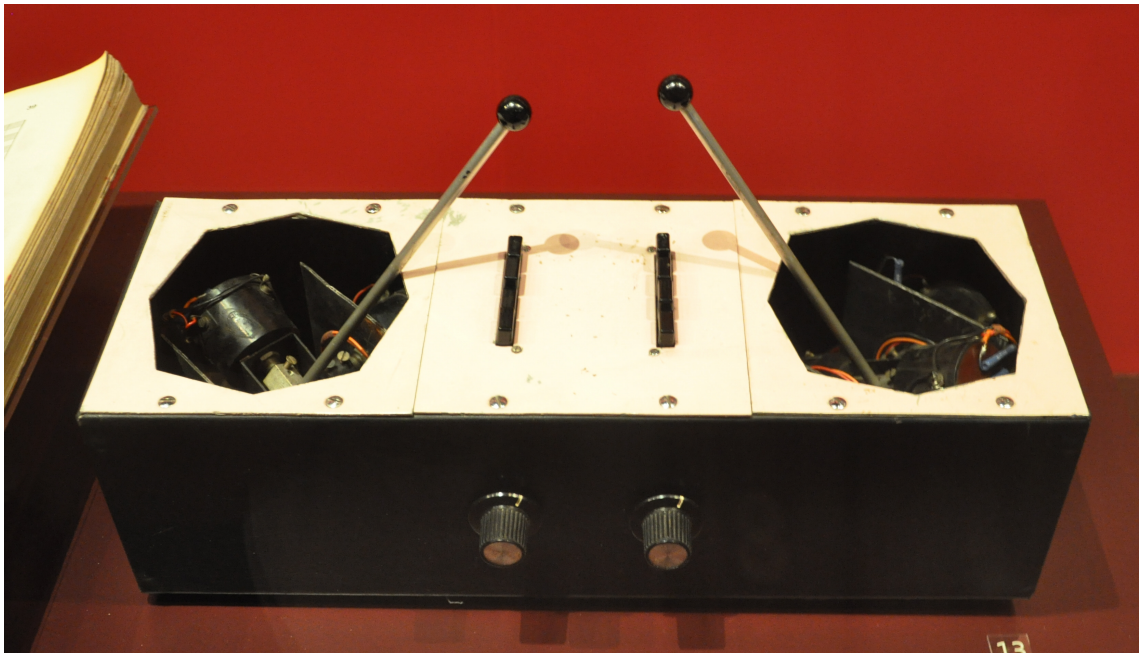

Figure 2: Azimuth Co-ordinator, a 4.0 surround panning interface from 1969.

2 Spatial audio

The first surround sound live music performance was constructed in 1967 by the band Pink Floyd. The setup was based on quadraphonics, or in more modern terms 4.0 surround audio. Surround sound reproduction was controlled with a joystick-based device called Azimuth Co-ordinator (Figure 2) operated by the band keyboardist, enabling the use of surround effects as part of the live music experience [15]. Since then the variety of surround sound systems and use cases has increased greatly.

2.1 Loudspeaker arrays, arts and exhibitions

There's an increasing amount of high density loudspeaker setups used in arts and exhibitions. Fulldome video experiences in places like planetariums calls for spatially mixed audio to increase immersion [16]. In this case the listener is located in a physical space, listening to sounds coming from a multi-channel loudspeaker setup around them. Different installations have different loudspeaker layouts which poses challenges quite different from an assumed stereo listening setup or headphones. Instead of having a loudspeaker array surrounding the listeners, it is also possible to create spatial soundscapes where the loudspeakers are placed on the stage near to where the sounds are wanted to originate from.

2.2 Games, spatial computing and binaural reproduction

In virtual 3D environments, such as games, surround sound may be an important part of an enjoyable, immersive and interesting experience. With binaural headphone reproduction it is possible to mimic a 3D soundscape at a modest price point. In a gaming environment the player is often moving within a virtual environment with 3 degrees of freedom. Here the listening spot is not fixed, which is counter to many user interfaces provided in DAW-based mixing environments. The tooling used in video game production is currently often part of the game engine and traditional DAWs are not at the core of the sound design of games [12].

Spatial computing headsets, such as Meta's Oculus and Apple's Vision have brought upon a need for new spatial audio production tools as the current solution landscape is limited [17]. Spatial headsets have head-tracking to create an experience of presence in a virtual space. To bring about an immersive experience, not only the visual world, but also the audial one needs to react to head movements. In other words, 6 degrees of freedom need to be accounted for. Interactions can be had with either separate controllers, or more recently, the gaze & pinch control scheme described earlier. In the case of spatial headsets, binaural audio rendering is often used to create more immersion. Entertainment usage such as spatial video recordings and headset-based gaming need tools supporting this kind of production targets [16].

Whenever audio is targeted at spatial reproduction, one needs to take into account the internalised expectations humans have about how spaces and soundscapes interlink. The moving of an object causes changes in the timing, the volume and the spectral content of the sound arriving at the listener. Whenever the virtual environment aims for immersion, there's a need to simulate the spatial characteristics of the virtual environment surrounding the listener. Many spatial production plugins and tools provide ways of simulating the acoustic characteristics of environments, providing smart tools where one doesn't have to match the effect with traditional adjustments like reverb and delay. Most stage metaphor based user interfaces also adjust several parameters for the user when moving the sound object, also counting as smart implementations.

3 Spatial audio production

3.1 Challenges in spatial audio production

As stated before, the reproduction setups used in real-world loudspeaker arrays may differ greatly. This means pre-rendering either cannot be fully materialised or has to be done per setup. Solutions like SpatialSound Wave from Fraunhofer offer an ability to render several mono audio tracks to a spatial composition. Alternatively, a scene based approach can be used that utilizes Ambisonics to allow for distributing the sound in spatial-native format. That can then be decoded to the desired setup

on demand [16]. While SpatialSound Wave is only meant for loudspeaker arrays, Ambisonics can be more flexible and be rendered to binaural headphone setups as well, utilizing either generic HRTFs or individual ones. Proprietary formats are also commonly used, especially in the movie and music industries, with licenses from the likes of Dolby.

Spatial audio mixing needs are different from stereo. Visualizing the location of the sound sources mixed in the space is more challenging than between just two sides. In an experiment conducted for amateur audio engineers, it was found that users mixing multi-channel audio may desire visualizations for EQ and compression, more than the traditional level and pan visualizations commonly available in existing user interfaces [1]. In addition to placing sounds in the 3D environment around the listener, it may also be desirable to be able to mix audio sources inside the listeners head, in a more traditional stereo panning style [17]. This is to account for e.g. narration or other artistic choices in storytelling.

The amount of audio channels or audio sources may be high in spatial audio production. Research suggests that it is important to provide a good overview that can be viewed with a glance, without the need to scroll or otherwise explore further [18]. The stage metaphor can be useful in discerning information about the whole group of channels when it comes to level and placement. However, the channel strip overview is the most convenient when one already knows which channel's state to look for [18]. This is echoed in Gelineck et al. concluding that, for experienced users, the traditional channel strip paradigm allows for quicker location of basic adjustments as opposed to the stage paradigm [10].

3.2 Designing mixers for spatial audio

Figure 3: Sound sources that are inactive may be de-emphasized by dimming them. Image from Gelineck and Uhrenholt (2016) [13].

Gelineck and Uhrenholt [13] explored stage metaphor based user interface concepts with different derived metrics, such as activity, level and frequency being visualized by size, shape and brightness. They found that dimming the objects representing inactive sound sources is intuitive and useful. They also conclude that mapping

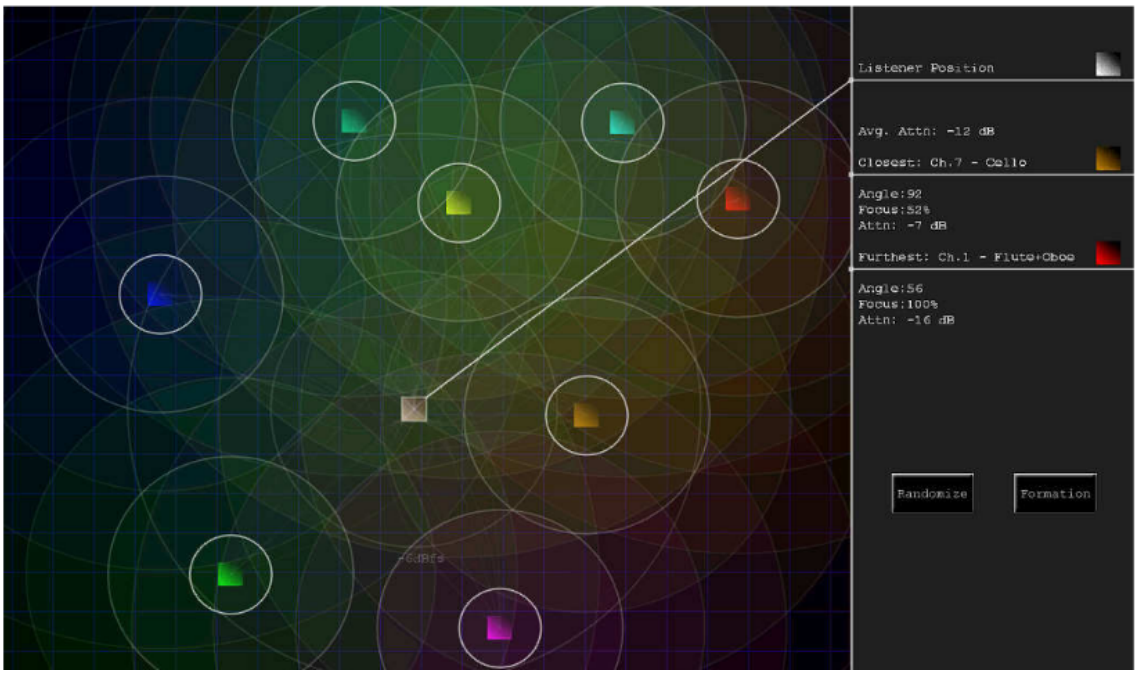

Figure 4: AWOL, an experimental touch-screen interface for surround sound mixing. Listener position is adjustable and the sound sources (colored rectangles) have attenuation radius as an attribute. Image from Diamante (2007) [12].

the spectral centroid to object brightness is more distraction than use. Similarly, mapping level to object size was found to be a distraction.

Diamante [12] presents a touch-screen based interface for mixing audio. It presents a 2D view of the listener and sound sources. Moving a sound source adjusts its width, level and high frequency filtering to bring a sense of distance. Sound source can also be assigned an individual attenuation radius. Moving the listener position then automatically adjusts all necessary parameters to provide a sense of any given listening position in spatial arrangement. The tool is not user tested in any meaningful way, but it provides an example of how one may manipulate a surround mix as something of a coreography of sound sources.

Riddershom Bargum et al. [19] demonstrate a simple spatial headset-utilizing sound mixing approach. Using Oculus Quest as a headset and bridging the controls to Ableton Live (Figure 5) the user may manipulate the locations of spheres in a virtual 3D environment. Binaural reproduction was used and the control scheme was Oculus Quest controllers. A test group used the tool and answered a questionnaire. Based on results this approach could be useful for quick sketching, or given improvements in the accuracy of controls, for more. There were also differences on how well the test group found the visual and auditory experiences to match when panning the sound sources in the virtual environment. An important area for furter study.

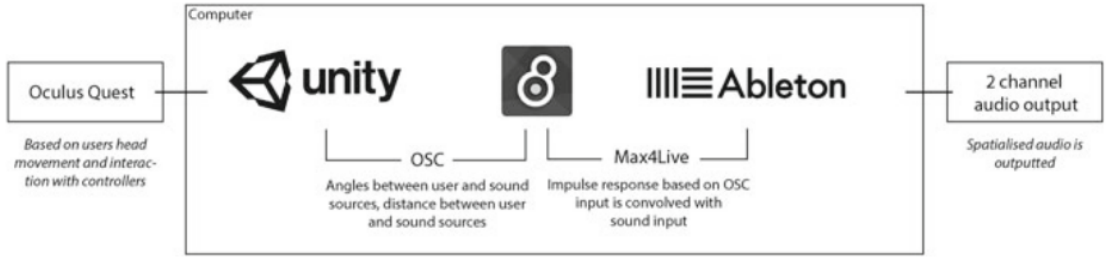

Figure 5: Spatial headset, a game engine and a DAW can be used together to produce audio experiences in virtual spaces. Example architecture image from Riddershom Bargum (2022) [19].

3.3 Current spatial audio user interfaces

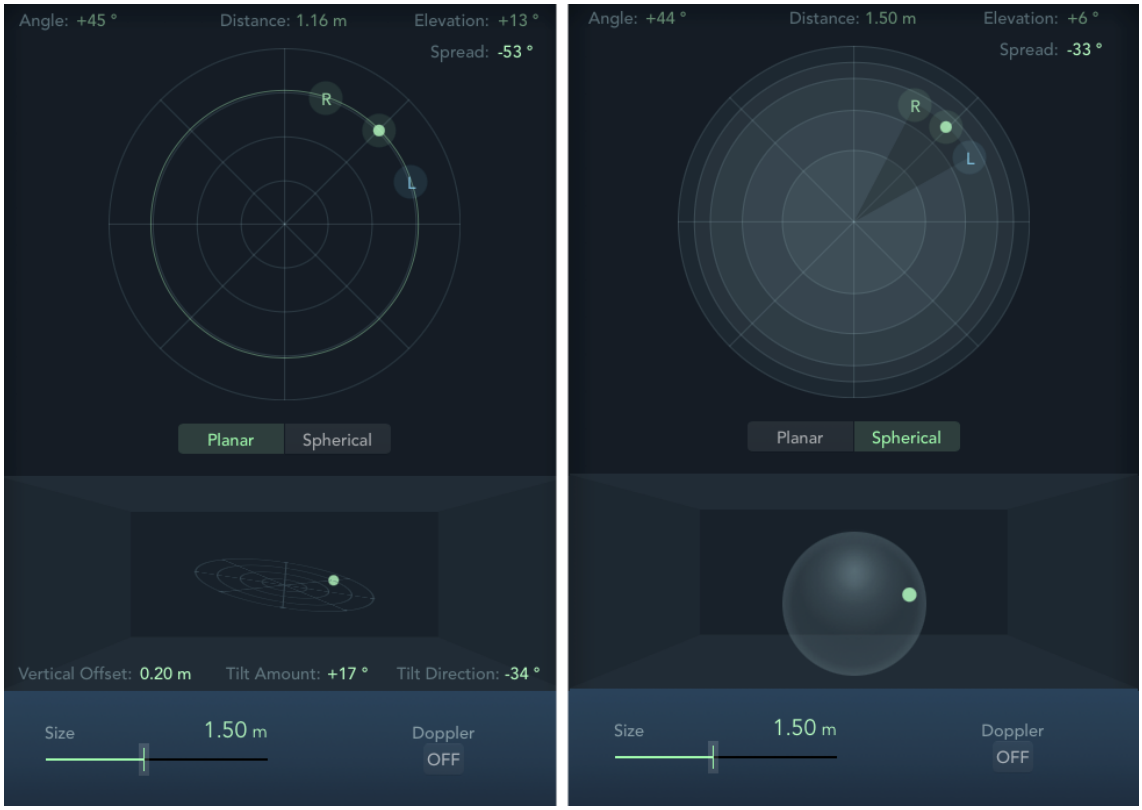

Figure 6: Logic Pro X has a built in binaural panning tool, here depicted in both planar (left side) and spherical modes (right side).

At least some digital audio workstations have built in spatial panning tooling, for example Logic Pro X's Binaural Panner tool (Figure 6). Through it the user assigns a spatial location relative to the listener for each channel strip. Stereo sound width can be adjusted for stereo sources and two choices of interface usage can be selected: planar mode and spherical. In planar mode one can adjust the location of the sound

source on a radar plate and the tilt of that plate. In spherical mode all adjustment is done on the radar plate. The former mode somewhat reminds of the analog origins of surround panning but the modern revision further extends the concept via the plate tilt control.

Figure 7: dearVR is an add-on for DAWs providing spatial mixing interfaces and tooling, including support for head-tracking. Image from Dear Reality [20].

A plugin suite called dearVR Pro is a one add-on choice for major digital audio workstations. It provides for instance a 3D panner, virtual room acoustic effects and supports up to 3rd order Ambisonics formats. It also allows for connecting to a head-tracking device and thus allow for listening the spatial production more immersively straight from a DAW. The panning tool (Figure 7) is again a digital iteration of the analog joystick panner with additional sliders allowing for also elevation adjustment.

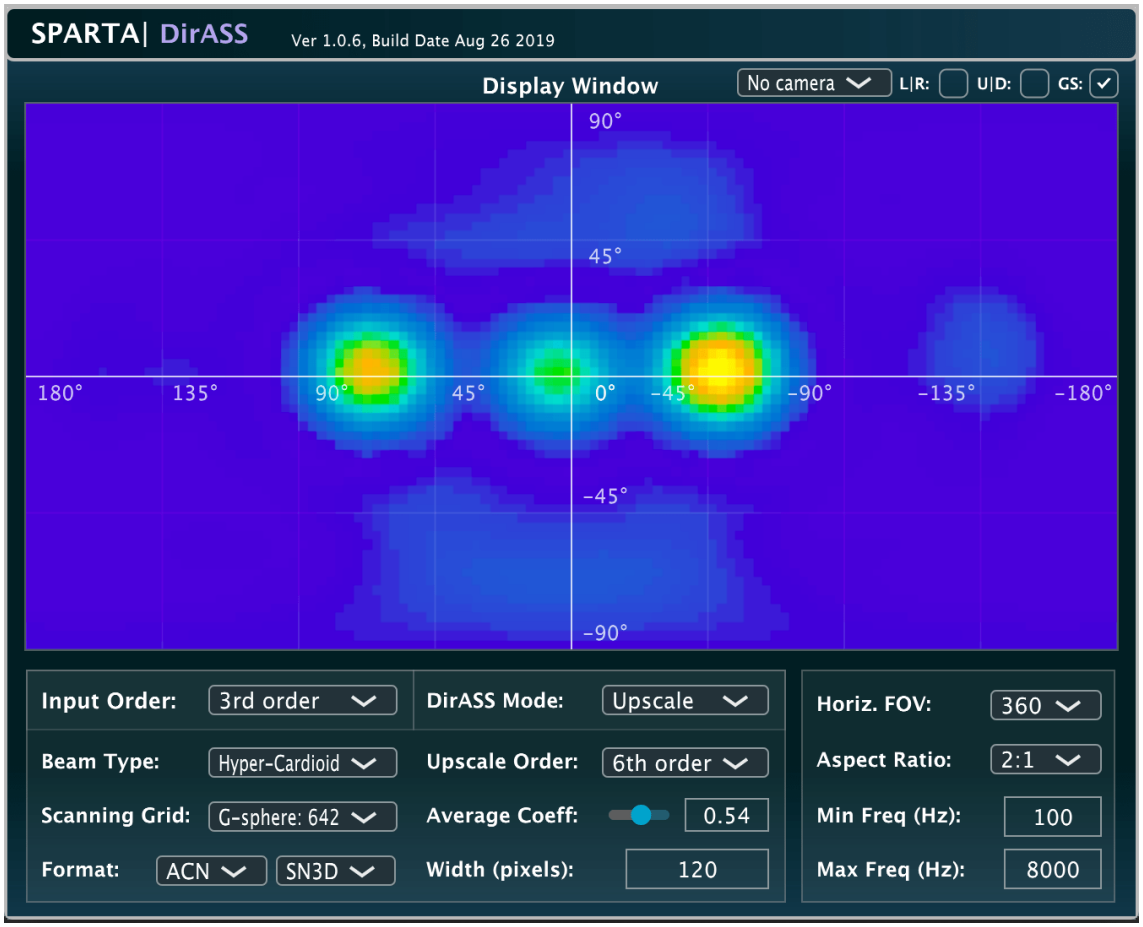

Figure 8: DirASS is a sound-field energy visualizer included in the SPARTA suite. Image from SPARTA website [21].

SPARTA is an open-source plugin suite for spatial audio. It includes various tools for e.g. panning, adding virtual room acoustic effects and analyzing ambisonic recordings. Support for head-tracking is included and various ways of manipulating ambisonic audio are provided. Being open-source it is especially useful for research and study scenarios and could be utilized for faster prototyping of innovative user interfaces. Simulation of room acoustics is available based on either a shoebox room image-source simulation or by using user-provided Ambisonic room impulse responses.

Unity and Unreal are commonly used "engines" in producing virtual spatial environments and games and they have first class support for the most used spatial computing headsets. They both have a simplistic spatial audio engine built in, but more advanced plug-ins such as Audiokinetic Wwise, Steam Audio and Microsoft Project Acoustics exist. All of these solutions integrate to the visual objects one manipulates in the engine and provide a physics-based soundscape matching the virtual environment, with either loudspeakers or binaurally using HRTFs. As such they rely almost completely upon setting the physical and artistic attributes and

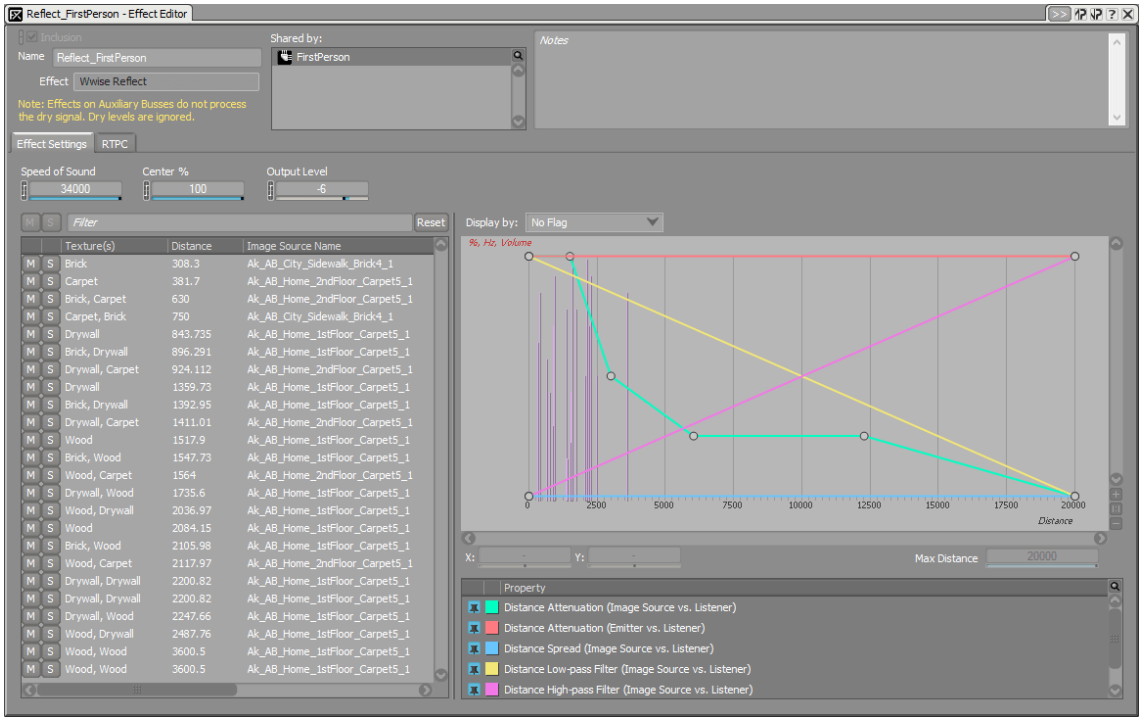

Figure 9: When physics-based simulations are used for spatial audio production, the user interface may contain mostly adjustments of the acoustic parameters of the objects in the virtual environment. Image of Reflect plugin from Audiokinetic website [22].

have little of the same user interface paradigms we commonly see in DAWs (Figure 9).

4 Conclusions

Spatial audio production needs differ from the needs of the times when many of the user interface paradigms for audio production were solidified. For reproduction and live performance scenarios the tools used are often DAW based, while in spatial computing platforms they're built around the engines used for creating the virtual experience. Some approaches on combining these two have also been presented. As the workflows of audio professionals moved from the constraints of physical hardware to more abstract workflows, the production chains became widely customisable. Many solutions can already be built around widely available tools and technologies and even some open-source solutions exist. In practice much of the work revolves around building plug-ins and physical controllers for either the DAWs or for the major game / virtual environment engines.

Some of the recent developments in spatial headset, like use of gaze & pinch interaction is currently missing in research publications. Many of the existing research also does not sufficiently define the target user group and thorough user evaluations

with several target groups is rarely performed. More efficient tools could likely be built based on the research understanding, at least for somewhat specific use cases. Stage paradigm has been explored in various forms for musical and live production and a lot about design choices have already been found. Expert users are already accustomed to the classic channel-strip paradigm, so perhaps use cases for more amateur users would find more significant use for the conducted research.

Spatial audio production is also a broad topic, including the needs of movies, music, performance spaces of all kinds, as well as complex spatial headset experiences. These areas have dissimilar needs and the existing solutions display a degree of specialization towards different, specific use scenarios, while the plug-in nature allows for combining various tools according to the needs and resources of the production. No major convergence to specific paradigms and metaphors can yet be declared and the field keeps evolving.

References

[1] Christopher Dewey and Jonathan Wakefield. Elicitation and quantitative analysis of user requirements for audio mixing interface. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[2] Adam Bell, Ethan Hein, and Jarrod Ratcliffe. Beyond skeuomorphism: The evolution of music production software user interface metaphors. *Journal on the Art of Record Production*, 9, 2015.

[3] Christopher Dewey and Jonathan Wakefield. A guide to the design and evaluation of new user interfaces for the audio industry. In *Audio Engineering Society Convention 136*. Audio Engineering Society, 2014.

[4] Donald A. Norman. *The design of everyday things*. MIT Press, London, 2001. ISBN 978-0-262-64037-4.

[5] Juan Pablo Carrascal and Sergi Jordà. Multitouch Interface for Audio Mixing. In *NIME*, pages 100–103, 2011.

[6] Jarrod Ratcliffe. Hand motion-controlled audio mixing interface. *Proc. of New Interfaces for Musical Expression (NIME) 2014*, pages 136–139, 2014.

[7] Jonathan Wakefield, Christopher Dewey, and William Gale. LAMI: A gesturally controlled three-dimensional stage Leap (Motion-based) Audio Mixing Interface. In *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[8] Jarrod Ratcliffe. Motionmix: A gestural audio mixing controller. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[9] Ken Pfeuffer. Design Principles & Issues for Gaze and Pinch Interaction. *ArXiv*, abs/2401.10948, 2024. URL https://api.semanticscholar.org/CorpusID:267069101.

[10] Steven Gelineck, Morten Büchert, and Jesper Andersen. Towards a more flexible and creative music mixing interface. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, pages 733–738, Paris France, April 2013. ACM. ISBN 978-1-4503-1952-2. doi: 10.1145/2468356.2468487. URL https://dl.acm.org/doi/10.1145/2468356.2468487.

[11] David Moffat and Mark B. Sandler. Approaches in Intelligent Music Production. *Arts*, 8(4):125, September 2019. ISSN 2076-0752. doi: 10.3390/arts8040125. URL https://www.mdpi.com/2076-0752/8/4/125.

[12] Vincent Diamante. Awol: Control surfaces and visualization for surround creation. *University of Southern California, Interactive Media Division, Tech. Rep*, 2007.

[13] Steven Gelineck and Anders Kirk Uhrenholt. Exploring Visualisation of Channel Activity, Levels and EQ for User Interfaces Implementing the Stage Metaphor for Music Mixing. In *Proceedings of the 2nd AES Workshop on Intelligent Music Production*, volume 13. Audio Engineering Society, September 2016.

[14] Steven Gelineck, Dannie Korsgaard, and Morten Büchert. Stage- vs. Channel-strip Metaphor - Comparing Performance when Adjusting Volume and Panning of a Single Channel in a Stereo Mix. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, NIME 2015, pages 343–346, Baton Rouge, Louisiana, USA, 2015. The School of Music and the Center for Computation and Technology (CCT), Louisiana State University. ISBN 978-0-692-49547-6. event-place: Baton Rouge, Louisiana, USA.

[15] Michael Calore. May 12, 1967: Pink Floyd Astounds With 'Sound in the Round'. *WIRED*, December 2009. URL https://www.wired.com/2009/05/dayintech-0512/.

[16] Johannes Ott, Anca-Stefania Tutescu, Niklas Wienböker, Jan Rosenbauer, and Thomas Görne. Spatial audio production for immersive fulldome projections. *Audio for Virtual*, Augmented and Mixed Realities: Proceedings of ICSA 2019 ; 5th International Conference on Spatial Audio ; September 26th to 28th:p. 179, November 2019. doi: 10.22032/DBT.39974. URL https://www.db-thueringen.de/receive/dbt_mods_00039974. Publisher: [object Object].

[17] Chris Pike, Richard Taylor, Tom Parnell, and Frank Melchior. Object-based 3D audio production for virtual reality using the audio definition model. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2016.

[18] Josh Mycroft, Tony Stockman, and Joshua D Reiss. Audio mixing displays: The influence of overviews on information search and critical listening. In *International Symposium on Computer Music Modelling and Retrieval (CMMR)*, 2015.

[19] Anders Riddershom Bargum, Oddur Ingi Kristjánsson, Péter Babó, Rasmus Eske Waage Nielsen, Simon Rostami Mosen, and Stefania Serafin. Spatial Audio Mixing in Virtual Reality. In *Sonic Interactions in Virtual Environments*, pages 269–302. Springer International Publishing Cham, 2022.

[20] dearVR PRO 2, . URL https://www.dear-reality.com/products/dearvr-pro-2.

[21] SPARTA, . URL https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/.

[22] Audiokinetic Wwise Reflect, . URL https://www.audiokinetic.com/en/wwise/plugins/reflect/.