Chapter 9
Spatial Audio Mixing in Virtual Reality

Anders Riddershom Bargum, Oddur Ingi Kristjánsson, Péter Babó, Rasmus Eske Waage Nielsen, Simon Rostami Mosen, and Stefania Serafin

Abstract The development of Virtual Reality (VR) systems and multimodal simulations presents possibilities in spatial-music mixing, be it in virtual spaces, for ensembles and orchestral compositions or for surround sound in film and music. Traditionally, user interfaces for mixing music have employed the channel-strip metaphor for controlling volume, panning and other audio effects that are aspects that also have grown into the culture of mixing music spatially. Simulated rooms and two-dimensional panning systems are simply implemented on computer screens to facilitate the placement of sound sources within space. In this chapter, we present design aspects for mixing in VR, investigating already existing virtual music mixing products and creating a framework from which a virtual spatial-music mixing tool can be implemented. Finally, the tool will be tested against a similar computer version to examine whether or not the sensory benefits and palpable spatial proportions of a VE can improve the process of mixing 3D sound.

9.1 Introduction

Mixing is the activity of placing and levelling sounds. When a sound source in the real world is placed spatially, the sound source's distance and direction is digitally managed in the process of mixing and thereafter perceived through different cues. This is done on a mixing console, being the primary interface for mixing 2D music; the console includes additional parameters to manipulate the general relationship between the sound sources. The parameters are amongst other things panning, volume, and equalisation, on each track. The mixing console is divided into functional sections, which constitute different metaphors [10]. As an example, each track has a channel strip, used as the main way to adjust the volume. The volume parameter

A. Riddershom Bargum · O. Ingi Kristjánsson · P. Babó · R. Eske Waage Nielsen · S. Rostami Mosen · S. Serafin (✉)
Department of Architecture, Design, and Media Technology, Aalborg University Copenhagen, Copenhagen, Denmark
e-mail: sts@create.aau.dk

M. Geronazzo and S. Serafin (eds.), *Sonic Interactions in Virtual Environments*, Human—Computer Interaction Series, https://doi.org/10.1007/978-3-031-04021-4_9

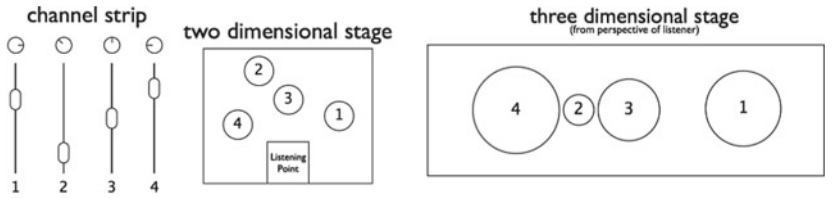

Fig. 9.1 Channel strip versus stage metaphor. *Hand Motion-Controlled Audio Mixing Interface* [23]

(also called a fader), is often built on a slide potentiometer and is seen as a universal metaphor for amplitude [3]. Simultaneously the pan potentiometer represents the track's placement and spatial position of a source in a stereo mix. Here, the pan potentiometer maps the left and right position of a knob to the left and right location of a sound. In general, the mixing console can be seen as a metaphor on its own and it has brought with it visually rich interfaces and representations of controls via graphical faders and knobs [15]. A mixing console can be divided into two main categories:

- Analog mixing console: it deploys a one-to-one mapping where every slider, knob and button has a dedicated function [10]. It is widely used and is known to be fast and intuitive.
- Digital mixing console: it reduces the design of the analog mixer by introducing sub-menus and layers, breaking the one-to-one mapping. It is still built on the channel-strip metaphor but enables a smaller interface as a user can scroll through the tracks. It thus allows for much more possibilities but demands more effort from the user, while it also might be harder to learn and control in a live situation [10].

The general channel-strip metaphor has furthermore been the standard way of implementing a mixing interface in different Digital Audio Workstations (DAWs).

Contrary to a mixer based on the channel-strip metaphor, either physical or digital, the stage metaphor/paradigm is a popular way of representing sound sources in a stereo field. In the stage metaphor, the level and stereo position (and possibly other parameters) are modified using the position of a movable icon on a 2D or 3D image of a stage [7] as seen in Fig. 9.1. This metaphor was first proposed by Gibson, who called it a 'virtual mixer' [12]. Even though the stage metaphor mostly uses one-to-one mapping in terms of volume and panning, it also incorporates one-to-many mappings, as the position of each sphere as an example can affect both the volume but also filtering and reverb in relation to the distance as mentioned earlier [7].

Both mixing metaphors come with drawbacks usability-wise. Firstly, it is obvious that when using several tracks on the original channel-strip mixing console, it can be hard to visualise and get an overview of the different tracks—all tracks that are panned to the left will as an example not necessarily be controlled by the faders and pan potentiometers on the left side of the mixer. This is easier in the stage metaphor, where each source is graphically visualised in a 3D space. However, the stage metaphor suffers from organisational consequences as all tracks are scattered

around the virtual room. With the channel-strip metaphor, each audio channel is always located in the same place, which makes it easy to find and control [10].

As mentioned by Gibson, the way humans perceive sound, besides physical sound waves and directional cues, is by imagining sounds between two speakers [12]. 'Imaging' works as a substitute of actually seeing the source that produces the sound [12]. Mixing engineers use both sound pressure as a perceptional tool when mixing, but also imagination, as it allows the engineer to create a wide range of dynamics [12], through the likes of asymmetric panning or uneven volume relationships, when visually placing sound sources 'between' the speakers [12]. Especially here the stage metaphor and the general visual representation thereof serves as a helping tool and underlines the importance and usefulness of VR in such a connection.

9.2 Audio-Visual Interaction

Vision is an important element in sound localization. As stated by Yantis and Abrams [28] and extensively discussed in Chap. 10, when conflicting information about sound localisation is given to both the auditory and visual system, the visual information dominates the perception. This effect, called *ventriloquism effect*, makes the person perceive the sound coming from the location determined by the visual system.

There are, however, three factors that can bias the visually perceived location [24]:

- The visual and auditory events must be close in time. Ideally, the visual event should happen before the auditory event.
- The events should be plausibly linked. In other words, the sound must be something that could have come from the visual source.
- The visual and auditory events must be plausibly close together in space. An example of this is if the sound is played through headphones, but the visual source is located behind the wall, you are likely to perceive the sound coming from the headphones and not the visual source.

In a study by Tabry, et al. [26] subjects were asked to localise a sound source by either pointing in the direction with their head or hand. There were two conditions: one where they were blindfolded, one where they were not. The results showed that the subjects were able to better localise sound on the horizontal axis than the vertical. Moreover, the subjects localised the sound more accurately when not blindfolded. This supports Abrams and Yantis' statement that humans, in some way, rely on both visual and auditory stimuli when it comes to sound localisation.

Furthermore, it was stated by Tabry, et al. that the results suggest a greater dependence on visual cues for orienting one's head towards a specific location in space

than for orienting one's arm [26]. It can be argued that this fact supports the use of VR when mixing spatial audio as it gives the user visual feedback of the sound localisation.

9.3 Designing Computer Music in VR

While there has been a lot of investigation in the world of designing VR for computer music, especially within the field of Virtual Music Instruments (VMI) and New Interfaces for Musical Expression (NIME) [16], little focus has been on graphically representing mixing, mastering and audio effects processing. With VR defined as an immersive artificial environment experienced through technologically simulated sensory stimuli [25], there is no doubt that this, combined with VR's inclusion of multidimensional spaces and free rotation/movement, enables the possibility of visually placing, moving and mixing sound sources in a 3D space. Concerning VMIs, Perry Cook states that copying an instrument is dumb, leveraging expert technique is smart [6]. This principle is, in particular, relevant to interfaces in VR as its visual qualities and lack of physical limitations can be used as a tool and paired with external applications like DAW or real-world speaker systems. The following section will focus on the aspects that might facilitate this and outline important guidelines for designing a VE for sound synthesis and mixing. The most important principles will include technological considerations such as latency and cybersickness, interaction types and possibilities, modelling sound in a physical space, as well as the overall graphical representation of the system.

9.3.1 Technology

There are multiple aspects to consider when designing virtual spaces and applications in VR in general, such as ensuring smooth interactivity through minimum latency as well as by preventing cybersickness. In a real-world modelling ideal it is preferable with a latency of 15 ms or less, when moving a head or object to see a new and corrected view of the scene. However, a lot of the head mounted displays (HMDs) only achieve a latency of 30 ms [21]. Getting as close as possible to the latency limits is important as synchronisation between the arrival of stimuli in different modalities is known to influence the perceptual binding that occurs in response to an event producing multimodal stimulation [17]. As individual senses in a virtual world still are not represented independently, synchronised audiovisual feedback is not only important as it serves as a response to a user's actions, it also creates a bridge between an activity and its given sound [17]. In 1998, Miner and Claudel [19] investigated the sensitivity of delay of auditory stimulus in multimedia applications. According to their analysis, requirements for sound-synthesis simulation of environmental effects like reverberation, Doppler Shift, and the generation of 3D sounds are at least 66

ms. It is thus clear that high latency when manipulating sound in a virtual 3D space will affect both the user experience and the overall perception of sound. Latency is as earlier mentioned believed to increase cybersickness, which also is affected by aspects such as display flicker and wrong calibrations. To prevent cybersickness, especially in an environment dealing with movement and placement of virtual objects, one-to-one mapping between virtual and real translations/rotations is advisable, as the vestibular system, in particular, is sensitive to such motion [25].

9.3.2 Interaction

Considering the interaction in a VR system, categories within the field of both user-orientation and user experience have to be examined. To fulfil good interaction in computer music interfaces, the musician, computer scientist, and designer Ge Wang suggests that the system amongst other things should [27]:

1. Be real-time if possible.
2. Design sound and graphics in tandem and seek salient mappings.
3. Hide technology and focus on substance.
4. Introduce arbitrary constraints.

In general, this means that interaction with sound sources should be easy, quick, streamlined and noticeable and that virtual objects need to match location and motion of auditory objects. Simultaneously, the user should not be confronted with technology or implementation, to increase excitement and interest. Another thing that will support the user's interest, but also immersion and virtuosity, is feedback. Various studies state that especially haptic and tactile feedback allows a user to develop musical skills and understanding of controls [25]. Inclusion of external controls that allow for touch or vibrational feedback thus could be beneficial. Gelineck et al. investigated this by comparing the stage metaphor (iPad App visualising a stage) to the channel-strip metaphor (normal faders and panning) when completing a stereo mix [11]. While they concluded that there was no significant difference in terms of performance, the iPad application was user experience-wise preferred for its intuitiveness, enjoyability and its ability to reveal the spaciousness of the mix [11]. They, however, outline a side effect of representing the mix visually, with the fact being that it might take away focus from listening. It thus is important to find the right balance of the graphical representation and haptic/tactile feedback, in order to keep the focus on the main aspects: mixing and listening to sound.

Using the strengths of VR will possibly improve this interaction. As mentioned by Serafin et al. it is believed that virtual reality shows the greatest potential when facilitating experiences that cannot be encountered in the real world [25]. This leads to the principle of considering natural and magical interaction in the system. The principle suggests that combining natural interaction (normal feedback to real-world movements) with magical interaction (interaction that is not limited by real-world

constraint as flying and teleporting) will open up for new and non-traditional interaction possibilities for already realised interfaces [25].

9.3.3 Sound in Space

Looking at the space in which the sound will be virtually presented, VR has several possibilities as introduced in Chaps. 1 and 6. Sound itself will in different physical spaces be shaped by the room's spectral characteristics and modified by room properties such as size, material, and shape. One can choose different methods when employing the models of spatialisation to the virtual rooms adjustment of the sound. Robert Hamilton distinguishes between two main models: the user-centric perspective and the space-centric perspective [14]. In the user-centric perspective, the sound will be manipulated from a first-person point of view, where sounds in the virtual world will correspond to a real-world-based model of hearing: they will be placed in a general aural spectrum known from the everyday, with corresponding depth cues implemented through filtering and delay. This can be done by tracking the coordinate distance between event locations and the user's in-game avatar [14]. The space-centric perspective, on the other hand, shifts the focus to the sound itself correlated between the virtual and physical world. In this model, sounds are no longer contextualised based on their proximity and relationship to a given user [14]. Instead, they are processed in relation to both the virtual and physical world, meaning the placement in each environment (as an example a spatialised speaker system) will affect it. This allows for multiple users and a communal experience [14]. In relation to the user-centric perspective, Gödde et al. [13], with a focus on the cinematic narration in VR, describe two possible 'user-centric' roles: a passive role, where the viewer is only an observer with no connection to the scene [13]—here the experience is more laid back and requires lower involvement resulting in focus on narration and the environment, and an active role where the viewer is part of the scene [13]—here the experience is involving resulting in a higher potential of presence that however might take focus away from narration and environment [13].

Besides handling the geometrical aspects of a virtual room, such as a sound source's spectral position and distance from the listener, it is also necessary to include a simulation of the acoustics of the given room. This will ensure a VR application that realistically represents the perception of sound. As stated by Falch et al. incorporating room simulation in binaural sound reproduction systems is important to improve localization capabilities as well as out of head localizations [30], which undoubtedly indicates the importance of acoustics when replicating binaural sound.

9.3.4 Graphical Interface

In relation to the graphical and visual representation of objects and environment in a 3D world, Wang proposes four aesthetic principles [27]:

1. Simplify: Identify core elements, trim the rest
2. Animate, create smoothness, imply motion: it is not just about how things look, but how they move
3. Be whimsical, organic: glow, flow, pulsate, breathe: imbue visual elements with personality
4. Aesthetic: have one; never be satisfied with 'functional'.

Since it is known that spatial audio approaches tend to facilitate interaction that is intuitive and familiar, the above principles are important as they can further enhance this. Especially the characteristics of simple and organic elements, as well as animating smooth motion of objects, will increase the user experience. This can as an example be done through the addition of shader programs as they aim to make virtual objects similar to their real counterpart, as shape, behaviour and appearance [18]. Shader programs are mainly used for the adjustment of a scene's illumination, post-processing or special effects [18], and the two most known shader types are Vertex shaders: the process that performs the transformations of vertices and texture coordinates from object space to window space, and, fragment shaders: A pixel shader that takes care of how the pixels between the vertices look [22].

Besides the principles of Wang, Gale et al. furthermore suggest that one especially should avoid visual clutter, meaning too many objects potentially overlapping and/or occluding each other on the screen [29]. As a part of object cluster and general control, Serafin et al. state that it additionally can be enhanced by visually representing the player's body [25]. People cannot see their own body in VR and this can be overcome by generating a visual substitution of a person's real body seen from first-person perspective. This will create a visual illusion and result in a 'virtual body ownership', which allows users to get the necessary presence that successful feedback requires [25]. However, different visual representations of the body will create different interaction expectations. A realistic representation of hands has as an example proven to create a more natural interaction experience than the given system allowed [2]. Thus the appearance of the virtual representation and the expectations it produces is important to consider.

9.4 Existing Mixing Interfaces

Different programs that have been created and used for mixing audio for 3D will be examined in this section. The focus will be on the implementation, design, and usability of the systems. Furthermore, features and standards of the existing programs

will be examined in order to find inspiration and reach a state-of-the-art level for this project's product.

'Auro Technologies'[1] is a company that aims to create the next generation audio standard by becoming the leader in state-of-the-art sound. They offer a product that can be used in the game and film industries as well as for mobile and automotive industries . This is made possible with AAX plugins, which allow the user to mix for a 11.1 system where the approach is to treat different elevation angles as layers, 'lower', 'height', and 'top'. Through algorithms, the audio is backward compatible with 5.1 and 7.1 systems.

The plugin offers an overview of where each speaker is located in a 3D space as well as displaying modifiable parameters, such as depth of reverb, bass and treble equaliser, and volume of the sound.

While mixing, each individual audio track in the session has a relevant plugin inserted. These plugins include 'Auro-Panner', 'Auro Bus', 'Auro-Mixing Engine', and the 'Auro-return'. The Auro-3D system is thus comprised of several plugins which, furthermore, requires a processor called 'A3DHost' to be running in the background while working with Auro plugins, as well as 'Auro-Dmix Control' in order to down-mix the bounce to a specific format.

Objectively, the approach of Auro-3D can be problematic as it may require a significant number of plugins to be running at the same time in a large project, which will affect the processing power. A powerful computer is therefore needed for it to be used with low latency. Simultaneously, one might argue that it is troublesome and counter-intuitive to individually place several plugins on each audio channel. An additional downside to this product is that it only works with Pro Tools, and only on Mac computers , which can eliminate a large number of potential users.

However, the product is still used heavily in the industry , and has won multiple awards.[2] Especially the design of it is important to have in mind when it comes to the product of this project. Even though it requires multiple plugins on each individual audio channel, the plugins have a clear user interface that highlights affordances and uses signifiers and feedback to give the user an understanding of what can be adjusted and modified. An example is the effects controls. Firstly, they are designed to look like knobs with labels above them to signify what each knob controls.[3] Additionally, there is feedback in the middle of the knobs to show which value they are set to, which furthermore is highlighted with lights around them to show where on the rotation axis they are set. This light also visualises in which range the knobs work and their boundaries in both directions. Some sliders control the volume of both the sound source and the amount of the reverb, which is a common way to control volume when working with audio. Even though the interface is not in 3D, it can be seen that Wang's aesthetic principles (Sect. 9.3.4) are relevant. Only essential settings for the volume and reverb are visible and modifiable (simplification), the sliders have value

[1] https://www.auro-3d.com.

[2] https://www.auro-3d.com/about-us/mission/.

[3] Knobs are commonly used on audio-related products and mixers.

feedback and the knobs have both value and light feedback (animation), which gives the interface and its controls a simple design that is easy to get an overview of.

Another product, although not commercialised, was made by Wakefield and Gale [9]. Their product was created in a research on how to solve perceptional problems in the 3D stage paradigm/metaphor when it comes to mixing audio. When more audio tracks are added, the visualisation can soon become cluttered, which causes problems in relation to depth perception that will be limited, leading to difficult interaction [9]. Furthermore, they wanted to minimise the risk of 'gorilla arm', which is a term referred to when users keep their arm elevated for a long period of time [5, 9]. Wakefield and Gale created an environment in VR for mixing audio. The system allows multiple options for adjusting each audio signal. There are send-effects control, filters, equaliser, volume, and pan parameters. All this is controlled with one controller. According to their studies, the VR mixing interface may have helped with depth perception of the audio. However, it did not improve clutter and object occlusion [9]. It may be reasonable to think that the UI of the program has affected those problems. Only one audio track is in the environment but the effect controls fill out almost the whole screen. So, even though the necessary parameters are present and no mentioning from participants of them being problematic, the displaying and arrangement of them might be something to keep in mind when designing the UI for the product of this project. With Wang's aesthetic principles in mind, it is clear that the interface is neither simplified nor aesthetic.

Dear Reality is a German company which specialises in creating 'ultimate tools for immersive 3D audio production'.[4] They offer multiple products under the name 'dearVR' for game engines, controllers, and DAWs . The 'dearVR Pro' product offers full 360° manipulation of sound with built-in acoustics and reflections controls

9.5 Target Group

Since this project aims to develop an aid for mixing spatial audio in VR, the user is expected to have previous experience in mixing audio but not necessarily spatially. This will allow the user to be aware of the given possibilities that a product facilitating spatial mixing gives (panning, volume change in depth, filtering as a result of elevation), but still explore the product as an entity. The product thus can be targeted at different groups ranging from game developers wanting to quickly sketch an audio-based atmosphere for their in-game environment, to music composers mixing spatialised audio for surround sound or VR applications and experimental musicians wanting to explore the use of 3D sound.

Since VR offers sensory feedback and spatial proportions differently to a desktop application, and since it is known that programs using the stage metaphor are intuitive (see Sect. 9.3.2), an everyday use of the end product could target composers and producers, that eventually might need a quick and easy assisting tool for the audio

[4] https://www.dearvr.com.

spatialisation process, be it music for film, sound design or audio for games. With this scenario in mind, the target group, therefore, covers both hobby producers, semi-professional producers as well as professional mixing engineers and composers, etc.—as long as they are familiar with mixing. To further understand the needs of a producer or composer mixing music for spatial media, an expert interview was conducted. An extensive questionnaire was sent to audio engineer Gestur Sveinsson, from the recording studio 'Studio Syrland' in Reykjavik, in order to get his opinions on necessities when mixing audio spatially. Having worked with surround sound for both cinema and music, Sveinsson notes the importance of having touchscreen mixing tools that allow him to quickly and intuitively translate his idea into reality. He furthermore adds that a visualisation tool for audio placement indeed would make sense as long as it is based on the idea of analog faders and panning knobs. In relation to his personal workflow, Sveinsson states that he visually sees the angles of the sound sources on a screen and arranges the mix without having to turn his head towards the angles that a given sound is coming from. Nonetheless, he thinks that a face tracking system would be useful and especially from a consumer point of view, it could make the experience 'hyper-realistic'. In relation to his personal preferences, he rates the aesthetics and design as well as the intuitiveness, and thus the time it takes to do a mixing task, as very important aspects in a mixing device, whereas the precision of it comes secondary.

With this knowledge in mind, the virtual mixing tool thus will be implemented using the research information above to target composers or producers that can use the mix both quickly, intuitively and precisely.

9.6 Conceptual Overview

The user is placed in a 3D VR environment where different tracks from Ableton Live are represented as spherical sound sources in space with belonging labels. A ray is cast from the user's controllers to signify which sound source one interacts with. If the ray is positioned on a sound source, the controller will respond with vibration to signify contact between the ray and a sound source. After selecting a sound source, the user can now move it in space. Data on position, distance, and angles will be passed into Max by Cycling 74 where the auditory placement in space will happen. This will result in the visual locations, as well as the auditory locations of the objects to match, and give an audiovisual experience in space as well as create a tool for users to visually place and mix different sound sources spatially.

9.7 Virtual Environment

To keep the centre of attention on the sound source's spatial proportions, the environmental setting will consist of very few props being a stage and virtual objects linked to audio tracks. It has been decided to design and model a stage-like environment

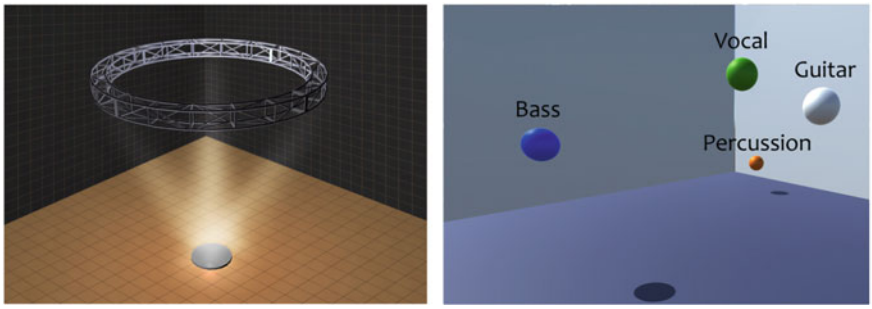

Fig. 9.2 Sketch of the design in its initial stage

as the main aspect of the surroundings, as this might elicit the stage metaphor and the virtual mixer as explained by Gibson in the analysis. This will allow the user to manually place objects in space with respect to a realistic and intuitive behaviour of the sound sources in the scene, like changing size according to distance or lighting up when pressed. Design-wise, it has been decided to develop an environment where the user is centralised in the scene on a slightly elevated cylindrical surface, to emphasis its role as a mixer/positional conductor. A circular truss will be positioned above the user (Sect. 9.7.1) to further highlight the 'stage' look. An illustration of both the initial design of the stage, truss as well as sound sources depicted as spheres is shown in Fig. 9.2.

To focus on the 'substance' principle, the scene will consist of: (i) A simple yet aesthetic environment to focus on the importance of the mixing task. (ii) The different objects' relations to the localisation of sound. (iii) Picturing spheres around a stage, to use the benefits of the stage metaphor like 'intuitiveness, enjoyability and its ability to reveal the spaciousness of the mix', as earlier stated by Gelineck et. Al (Sect. 9.3.2). To keep an aesthetically pleasing, yet simple design that guides the user's attention to the mixing task rather than the visuals, it has been decided to avoid scenes with myriad different elements such as concert halls and theatre stages, as this might take focus away from listening. To further enhance the feeling of spaciousness within the environment, it has been decided to use a 'grid-like' structure on walls, floors and ceilings. This is, as seen in the state-of-the-art section (Sect. 9.4), a widely used technique to display and give the user a sense of dimensions and will automatically create spatial constraints as it outlines the boundaries of the room, giving possible distance limits in relation to sound source placement. These constraints are furthermore supported by the truss, which represents the outer boundary of object placement—the user cannot place sound sources outside of the truss area.

An aspect that is widely used especially within cinematic VR, also called 360° videos, is the use of cues to guide the user's attention, as the user can freely rotate its head and thus choose the field-of-view (FoV) [20]. Whereas these cues normally focus on storytelling and narration, cues like implicit diegetic cues are

also applicable within this exact environment. Implicit diegetic cues are factors like objects or props within the scene that implicitly guide the user to do something [20]. This is, as an example, seen in the truss acting as an environmental constraint as mentioned above—the truss is a barrier signifying the limitation of object placement and thus forcing the user to re-orientate. The sound sources themselves are furthermore important implicit diegetic cues as they give the user a sense of their placement and space when the user has to redirect attention. The sound sources act between being onscreen and off-screen diegetic cues according to when they are present within the FoV [4]. As the spheres serve as spatial guidance, both their look, sound, and feel are of high importance when it comes to attention leading and general orientation within the environment. In relation to this, it has, as an example, been chosen to make the different sound sources light up when hovered over by the user, as this gives the user a better overview of the mix, as well as avoiding confusion by changing colour when something is 'soloed'—the act of isolating a sound.

As mentioned, these cues simply affect the FoV and thus also the 'user-centric perspective' that manipulates the sound from a first-person point of view and contextualises it to the position of the user. The placement of the camera, which also serves as the perspective of the user, therefore has to match the general viewing position. This will be done in Unity using the camera as the viewpoint. For the vertices and fragments of objects and shaders, the 'user-centric perspective' will be handled using the 'object to clip' node, which transforms a position in object/local space to the camera's clip space.

9.7.1 Rendering and Lighting

The 'Lightweight Rendering Pipeline' (LWRP) in Unity will be applied to render the scene and its light. Since the Oculus Quest used is dependent on its hardware, the LWRP will be optimal as it targets a broad range of mobile platforms, VR and games with limited real-time light capabilities.[5] By making a few trade-offs in relation to lighting and shadows, like fewer draw calls, the LWRP optimises the real-time performance of the system thus allowing for uncomplicated real-time processing and salient functional mappings, which was mentioned as important design requirements.

In relation to the lighting within the scene, general directional lighting is used to illuminate the environment. The lighting was chosen to be coloured to add to the atmosphere within the scene. Coloured lights were simultaneously used as decoration within the scene, where bars in red and blue represent LED strips. The emission of white rings on the walls and in the surface additionally adds light to the scene and through global illumination, surface reflections were simulated. To enhance the 'stage aesthetics' even further, coloured fog was included, as fog is usually experienced within concert experiences.

[5] https://docs.unity3d.com/Packages/com.unity.render-pipelines.lightweight@4.0/manual/index.html.

9.7.2 Interaction

The VR system contains three main interaction types and their respective feedback, including visual and auditory feedback, which is a combination that reinforces a user's given action, meaning that the user both sees and hears the results of the actions made. The different interactions and their feedback, as pictured in the conceptual overview above can be explained as:

- Touch/haptic interaction: the selection and manipulation of the different sound objects will constitute the haptic interaction. Here the user is allowed to touch/ select, by pressing a button and move, by moving its arm, the different objects.
- Visual feedback: an object will light up corresponding to it being clicked/ selected, and move corresponding to the user's force and arm movement. The visuals are thus designed with a focus on natural mapping as the sound sources, with respect to their auditory perception, are placed where they would be placed in a real-world situation.
- Auditory feedback: the panning and the volume of the chosen sound source will change accordingly to the placement of the object both on the horizontal axis (azimuth) and depth (distance).
- Tactile feedback: vibration will happen when the user hovers over an object to signify its allowance of being selected. This is to help the user find and aim at the desired object.

To sum up, it can be said that the visual feedback of the haptic interaction facilitates the stage metaphor and virtual mixer analogy, as sound sources are positioned in space relative to the user, whereas the auditory feedback facilitates the binaural synthesis and combine the interplay of visual and auditory cues used in human perception. The tactile feedback, on the other hand, constitutes increased usability of the product and the potential of the user to, within the environment, gain skills and understanding of the different controls. Additionally, it acts as a substitute of the mixing console, which as earlier mentioned also is a tangible controller. How the user scrolls through the audio tracks, and visually as well as auditory pans and levels them, is now an integrated part of the VE, rather then the mixing console.

9.7.3 Shaders and Visual Appearance

In Fig. 9.4, the visual appearance of the final environment is shown. This design was reached from aesthetic and stylistic ideas received from different scenarios seen in the mood-board below. Inspired by the 'stage metaphor', spheres were used as sound sources, instead of objects picturing the actual instrument/object the sound source is coming from. This was done to avoid unrealistic representations of the sound sources, which potentially could create user aversion and additionally introduce latency problems for the Oculus Quest. The spheres were furthermore chosen as

Fig. 9.3 Inspirational mood-board for visual appearance and colours

the main audio representation as they could constitute an abstract feeling to the very artistic subject that music and mixing is, as well as being used in products such as the dearVR (Fig. 9.3).

As seen in the mood-board, the colours blue and red, as well as the effect of lasers serve as a big inspiration for the look of both the environment and the shaders.

9.7.4 Audio Design

For the audio, head related impulse responses (HRIR) from MIT were used.[6] The pack includes IRs ranging from –40° to +90° on the vertical axis where each elevation had their own IR for the azimuth (5 degrees between each IR). Each IR was measured at a distance of 1.4 m. As this pack consists of 710 different IRs, the computation would both be heavy and complicated and, therefore, it was decided to evaluate whether it was needed to implement the IRs for elevation, as humans have perceptual difficulties placing audio on the vertical axis.

9.7.4.1 Can We Remove Auditory Elevation Cues?

A total of 14 participants were gathered for the evaluation, which was set up at Aalborg University in Copenhagen. The participants were informed of the research question *'Do you feel like the sound is matching the position of the object?'* before the test started and asked to answer either 'yes' or 'no', with the option to hear the

[6] https://sound.media.mit.edu/resources/KEMAR.html.

sound again, if needed. Additionally, they were encouraged to focus on a sphere centred in the middle of the screen, but they were allowed to look around.

The threshold of accuracy was set to 80% for scenes with no elevation, and 50% for scenes with elevation. There was no audio manipulation (volume change, filtering) for the elevation. The results showed that 91.4% of the participants felt that the audio matched the position of the object with 0° elevation and 95.2% felt the audio matched the object's position when it was elevated.

The results from this evaluation show that having visual cues for the audio source made the audience interpret the sound to originate from the visible object. Even though no audio manipulation took place regarding elevation, overwhelming majority perceived the audio to be elevated. Therefore, it was decided to not implement the HRIRs for the elevation, and instead, sound design-wise only relies on azimuth and distance cues.

Even though no audio manipulation was implemented for this particular test, it is important to implement acoustics manipulation relevant to the environment. This relates both to volume change based on distance (Inverse-square law) and reverberation based on the dimensions of the room. Both reverb and low-pass filtering will, therefore, be added to change timbral properties and give the sound an applicable aspect of room acoustics. The properties will be applied based on subjective sound-aesthetics and not by using physical models.

9.7.5 Summarising Design

Conclusively, to keep the attention of the user focused on the sound sources location, the VR will consist of very few props consisting of a stage and virtual sound sources. The stage metaphor, as well as a grid shader for walls and floor, is used to reveal the spaciousness of the mix as well as the VR. Other shaders, such as a fresnel effect, are used to signify user action as well as provide aesthetic value. Using theory from cinematic VR, sound sources can be considered on- and off-screen diegetic cues, guiding the user's attention, while the view facilitates an active user-centric perspective. To optimise the performance of the system the rendering pipeline 'LWRP' is used, while it also was decided to keep the lighting of the scene relatively simple. A combination of haptic, visuals, auditory and tactile feedback is used to enhance usability. Spheres were chosen to represent sound sources to constitute an abstract aspect of audio mixing and imaging as explained by Gibson. Based on the conducted evaluation of perception of elevation, showing that an overwhelming majority perceived auditory elevation based on visual feedback only, it was decided to only implement HRIRs for azimuth. Simple acoustics manipulation will furthermore be applied to simulate distance of sound sources. An illustration of the final environment, including colours, shaders and lighting can be seen in Fig. 9.4.

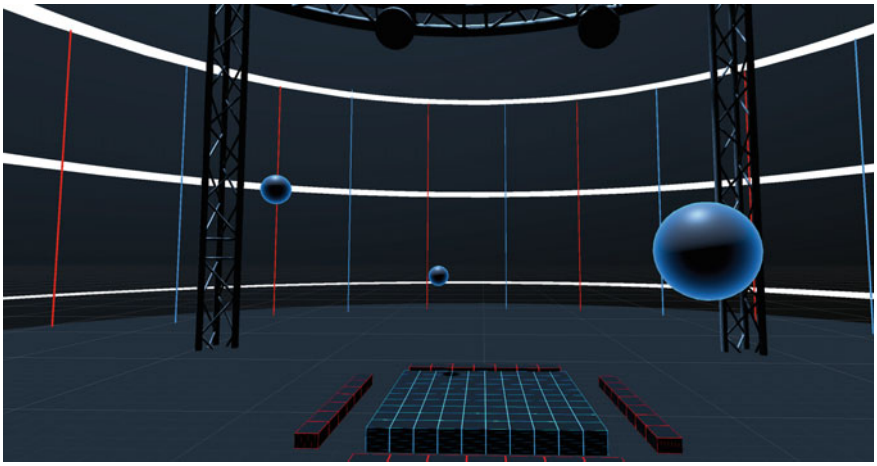

Fig. 9.4 Final design of the environment

9.8 Implementation

The implementation of the interactive VR environment and its inclusion of dynamic binaural synthesis consists of different steps and programs:

1. Firstly, a combination of the Oculus Quest system and the game engine Unity will be used to create a 3D environment that allows the user to manipulate and position objects within a virtual space. Support for VR in Unity will be imported through asset store items, in this case, Oculus Integration is used.[7]
2. Secondly, object coordinates, angles and user head rotation, will be implemented and retrieved based on different scripts. This will be sent through Open Sound Control (OSC) to Max via User Datagram Protocol (UDP) connection. An additional Unity asset store item called 'OSC simpl' is here used.[8]
3. Finally, Max and its live integration with Ableton Live will execute real-time sound rendering and binaural synthesis, through a convolution process of different HRIRs related to the respective sound object angles.
4. The communication between Unity and Max will furthermore be emphasised, as the track/audio names from Ableton Live will be displayed as part of the sound sources in the VR environment. This will additionally be implemented through OSC communication.

UDP is a connectionless communication protocol used across the Internet, especially for time-sensitive transmissions and is considered a quick communication protocol, as it allows data transfers before the receiving party agrees to the commu-

[7] https://assetstore.unity.com/packages/tools/integration/oculus-integration-82022.

[8] https://assetstore.unity.com/packages/tools/input-management/osc-simpl-53710.

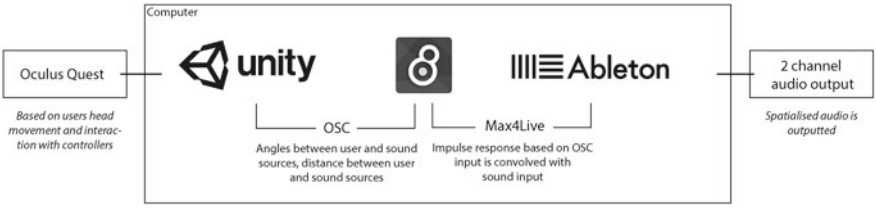

Fig. 9.5 Illustration of the system

nication.[9] OSC is, likewise, a protocol especially for networking sound synthesisers and computers. It uses UDP to transfer data within local subnets and is thus an obvious protocol for UDP communication. An overview of the different stages, systems and software used can be seen in Fig. 9.5:
The process of representing real-time audio spatialisation is done using a range of scripts developed to ultimately pass the necessary information from Unity to Max to create a spatialised mix.

Convolution

The convolution of the incoming signal and the signal of the different HRIRs is done in the frequency domain using the `pfft ~` object. The `pfft~` object essentially is a processing manager that splits the FFT process into smaller tasks, each taking care of their own FFT process.

Distance Simulation

While it earlier was confirmed that humans have a hard time distinguishing between elevations of sound, especially being accompanied by a visual object, the simulation of distance is easy to perceive and important both in relation to the display and localisation of sounds. In this project, it has been decided to use the Inverse-square law to simulate distance. In relation to the difference of sound in each ear, due to the acoustic shadow of the head, this project only takes the interaural time difference (ITD) and sound intensity in space into consideration. The frequency dissimilarity and the ITD at longer distances (the ITD is covered through the HRIRs at closer distances) are thus not considered. This has been decided due to the fact that spectral cues at shorter distances (10 m) are insignificant and that sound has to travel more than 100 m for frequencies around 4kHz to be attenuated 7 dB [1].

[9] https://www.howtogeek.com/190014/htg-explains-what-is-the-difference-between-tcp-and-udp/.

Integration With Ableton

The Max patch created is added to each audio track in Ableton as an audio effect. The azimuth and distance for left and right ear are shown to the user. Additional information is visualised for the user such as the *wet* amount of the reverb and the cutoff of the filter used. From a drop-down menu, the user can choose which port this track should listen to. This ensures that the same patch can be used for all the tracks and only the port has to be changed. Additionally, the IRs are read in and when loaded, they are automatically applied to all active patches in the Ableton project.

9.9 Creating the Computer Version

As earlier mentioned, a computer version was created for comparison purposes with the VR version. It has been decided to compare the VR version to its 'computer screen' counterpart, as this will put the possible advantages of the VR version in focus and thus not be influenced by aspects such as look or controls. The computer version of the final product was created to be as similar to the VR version as possible, however, there are some key differences when it comes to how the interaction is carried out. A ray is cast to wherever the mouse cursor is placed in the scene. When holding right-click, camera rotation is enabled, and the camera rotates based on mouse movement to imitate the camera interaction of the VR version. The 'track' object is 'grabbed' by hovering the mouse cursor over an object and holding left-click, whereafter it is possible to move and place the grabbed object by using the keyboard keys 'W-A-S-D'. A combination of left- and right-click makes the grabbed object follow the camera rotation.

The environment in the scene uses the same exact objects, coordinates and other visual aspects to eliminate any bias against or towards any of the two versions in this aspect. However, there are some small differences, because of different hardware, like display colours and refresh rate.

9.10 Evaluation

The following evaluation presents the setting, procedure and results of both the final focus group interview and mixing task evaluation. The mixing task evaluation uses a t-test to investigate the relationship of precision and time used between a VR version of the program and computer screen version. This is done to examine whether or not the benefits of VR and its sensory inclusion, can be seen as an overall improvement when mixing spatial audio. The computer screen version thus acts as the control condition representing similar interaction, affordances and sensory stimuli that can be found when placing audio spatially on a computer screen. However, it is important to note, that it is not a specific resemblance of already existing products.

For the t-tests, the following two hypotheses are further evaluated:

- **H0**—The mixes made by the participants in the VR version have no difference or less relative precision to the reference mix, than the mixes made in the computer version.
- **H1**—The participants in the VR version, used less time on recreating the reference mix, compared to the computer version.

9.11 Setting and Procedure

Both the focus group interview and part of the mixing task evaluation took place at Aalborg University Copenhagen. The other part of the mixing test took place in an apartment in Nørrebro due to convenience of test participants. Twelve students and a professor from the 'Music Production Bachelor' of the 'Rhythmic Conservatory of Copenhagen' attended the focus group interview, whereas 24 participants with different musical backgrounds and mixing experience, took part in the mixing task evaluation.

The *focus group interview* was carried out after the 'Danish Sound Network' event 'Behind The Scenes' on the 2nd of December 2019, where each participant had tried the VR mixing program for at least 5 min, including instructions. During the trial, the participants were instructed to try out and notice different aspects such as free rotation, elevation, auditory feedback and visual appearance. After the possibility of giving individual oral feedback, a focus group interview was conducted. Here each participant had the possibility of discussing and evaluating different topics selected by the conductor, with each other. The focus group interview was carried out in informal surroundings and lasted for 55 min.

The *mixing task evaluation* was carried out over three days, from the 9th to 11th of December 2019. The evaluation took place in a separated room, where each participant tried both the VR version and the computer screen version of the program, in randomised order to avoid an experience bias. Both time and precision for each participant were computed. It is important to note that the precision of the participant mix in relation to the reference mix, is measured in units (in unity called metres). While the amplitude of sound normally decays logarithmically, the distance between sound sources close to the participant is of higher significance than sound sources farther away. Half a unit is, as an example, visually experienced as a bigger distance near the participant than for objects in the distance. The measure of precision should thus be seen as a relative unit of measurement and not a counterpart to objects in the real world.

The participants were clearly instructed that their only task was to recreate the given audio mix in each version and that they had an unlimited amount of time until they felt satisfied with the mix. In both cases, clear instructions and a poster showing the different controls were offered. The participants had time to familiarise

themselves in each environment before they proceeded to the mixing task. The Oculus Quest VR-headset and a Macbook Pro were used for the VR and computer screen versions, respectively. The audio was routed from a separate computer through a pair of wireless AIAIAI TMA-2 headphones. The mixing task sessions lasted between 25–45 min for each participant, depending on the amount of time used on each mix.

9.12 Participants

Before the participants of the mixing task evaluation started, they were asked to answer a few questions regarding their musical/production/mixing background. These questions were used to ensure that the participants, and therefore the data gathered, matched the pre-defined target group; hobby producers, semi-professional producers, professional mixing engineers and composers (Sect. 9.5).

It could be seen that the majority of the participants had a lot of mixing experience, where 54.2% had 3+ years of experience, and of them, 41.7% had 5+ years of experience. The rest, 45.8% had 1–2 years of experience in mixing music. The same percentage was seen regarding their experience in mixing spatial audio (binaural, surround, ambisonic). The majority (54.2%) did not have any experience, whereas 45.8% had experience. Lastly, a question regarding whether or not the participants had tried VR before, was asked, where the majority had tried it before (83.3%).

9.13 Results

This section will focus on presenting the results from the evaluation. The section will be divided into three different sub-sections: Focus group interview (conducted with students from Rhythmic Conservatory of Copenhagen) and the mixing task evaluation including a post-evaluation survey.

9.13.1 Focus Group Interview

As mentioned, the first part of the evaluation consisted of a focus group interview. Below different quotes, opinions and summaries divided into themes and main topics are outlined, as discussed in the focus group interview. See appendix C for full interview transcription.

Efficiency

- ‘I consider time. If it takes more or less time to do in VR. I don’t know’.

- 'I could imaging that you would get done faster with some things. It seems very effective'.

The participants were asked about their initial thoughts regarding VR as a mixing tool and were all concerned of how efficient it potentially could be. Some participants felt that VR might be used as a quick sketching tool and compared it to a big brush painting on a canvas. Besides the concept of this project, it could be used as a more creative tool, rather than something one would use for precision.

Environment

- 'It could potentially set an atmosphere'.
- '... when I mix it is definitely something visual happening in front of me, I see the elements in front of me. It is not necessarily that I see the orchestra in front of me—it is much more abstract. A sprinkle over here, the sub-frequencies being another shape'.

It was stated that the virtual room could set a mood for the production by having different abstract elements. One participant pointed out that if the room should set a mood, it should be in a visually abstract way and not by looking realistic as this was the way he mixed mentally. It was furthermore stated that the decision of keeping the environment relatively neutral made sense in order not to influence the mix in an undesired direction and that the visuals used were pleasant and made sense in a mixing situation.

Spatial Sound Algorithm

- 'I found it slightly under-dimensioned so when you panned things to the side it was not as much as you would imagine. Front and back made sense well. When I panned something this much to the right I would also expect to hear it more to the right'.

One participant described the panning as being 'under-dimensioned', but in general the participants found the spatial algorithm satisfying. A few participants noticed the exclusion of elevation, whereas most participants felt the match between sound and source movement realistic.

Features

- 'It could also be used to do automation in a mix. ... You would have a much bigger area to draw on. I think that would be extremely useful'.
- '... they (objects) could have different shapes depending on if it is a string instrument or a wind instrument, or drums, or vocals'.
- 'Shouldn't the other button be mute?'
- 'I think it is necessary to know that there is activity on the track'.
- 'Put a number on the dB, like a gain volume'.

Multiple participants described how they could imagine using this tool to create automation. Having different visual representations of the different sound sources was also confirmed by several participants as well as having a visual representation of sound activity on each track as a VU-metre on mixers.

Precision

- 'In a DAW [it would be most precise]'.
- 'But also hard to say, when we only tried this a bit'.

There was a general consensus that a DAW was expected to be more precise than the implemented VR program. It was pointed out by one of the participants that it is hard to say when they have not spent more time using the VR program, but in general, the program, together with the binaural algorithm was experienced as something quick rather than precise. All participants agreed that the program gave enough information to make a judicious mix, but the concept made it hard to fiddle and go into small detail.

Concept

- '... I still think it suddenly is more about my experience mixing it instead of making an experience for others. And those two things can of course play well together, but I think I can install myself in a studio environment which helps me to get a good experience without a VR-headset'.
- '... it feels a bit more like you are going to play a game. In some parts of the process it might be a positive thing, I mean also more for composition'.
- 'I cannot accept that I have not decided what it is this movement does. ... I do not have any emotional connection to this'.
- 'I think as the program is right now it might work even better for people who do not have experience making music and have to learn to visualise music in an extremely intuitive way'.

A participant pointed out that he found the prototype to be useless for him since it was designed for spatial mixing and not directly suitable for exporting a stereo mix. He pointed out that the prototype seemed to be designed for the producer to have a good experience instead of the final listener to have a good experience. It was mentioned by one participant that the application would be more relevant to use if it at least included the functions of a large format console channel strip, for each audio track.

Comfort in VR

- 'I felt a bit sick. When I took of the glasses I felt really dizzy, but I think it is something you maybe have to get used to'.
- '(In the environment I could spend) 5–10 min or something like that'

One participant explained how he felt dizzy after using the prototype, while another participant imagined that he could not spend more than 5–10 min in VR. It was furthermore discussed by several participants whether switching between headset and screen was better than staying in VR. Both ideas were supported by other participants.

Table 9.1 Mean and Std Deviation of Precision and Time across the two experimental conditions

	Precision (Unity units)		Time (s)	
	Screen	VR	Screen	VR
Mean	35.74	35.60	558.38	448.04
Std. dev.	12.64	12.48	325.21	248.17

9.13.2 Mixing Task

Besides the focus group interview, the final evaluation consisted of a mixing task in two different versions. The means, standard deviations as well as histograms of the collected data for both time (in seconds) and sums of relative precision compared to the reference mix (in Unity units), are shown (Table 9.1).

The level of measurement of this evaluation is ratio data and since the study is using a within-group design, homogeneity of variance can be assumed without testing this. To test if the data is normally distributed an Anderson-Darling test is used on each data set for both precision and time.

For precision data, the results of the Anderson-Darling tests confirmed that the null hypothesis 'the data are normally distributed' should not be rejected, outputting p-values for screen version and VR version, respectively, $p = 0.4381$ and $p = 0.0693$.

For time data, the results of the Anderson-Darling tests showed that the null hypothesis 'the data are normally distributed' should be rejected, outputting p-values for screen version and VR version respectively, $p = 0.0005$ and $p = 0.0422$, meaning that the data is ***not*** normally distributed. The differences of the samples were also seen in the standard error, which is a measure of the sample reflects the population, in this case, 66.3 and 50.5, respectively. Additionally, the correlation coefficients describing the relationship between accuracy and time spent with the mixing task indicate no significant correlation, neither in VR ($r = -0.3$) nor in the PC version ($r = -0.2$).

Since only the data from the precision test is normally distributed and thereby fulfils all assumptions for parametric tests, t-tests will be used only on the precision data. The t-test does not reject the null hypothesis 'The mixes made by the participants in the VR version has no difference or less relative precision to the reference mix, compared to the computer version' with 95% confidence. A high p-value, ($p = 0.9531$) simultaneously shows that no significant differences between means are found.

It was mentioned by multiple test participants, that they had a significantly harder time spatially positioning the 'choir' track (a track in the pre-made mix) in the VR version, compared to the other tracks (discussed in Sect. 9.14). If the position of the 'choir' track is left out in the data sets, the t-test rejects the null hypothesis 'The mixes made by the participants in the VR version has no difference or less relative precision to the reference mix, compared to the computer version' with a 95% confidence level. It furthermore has a low p-value ($p = 0.0015$), meaning that

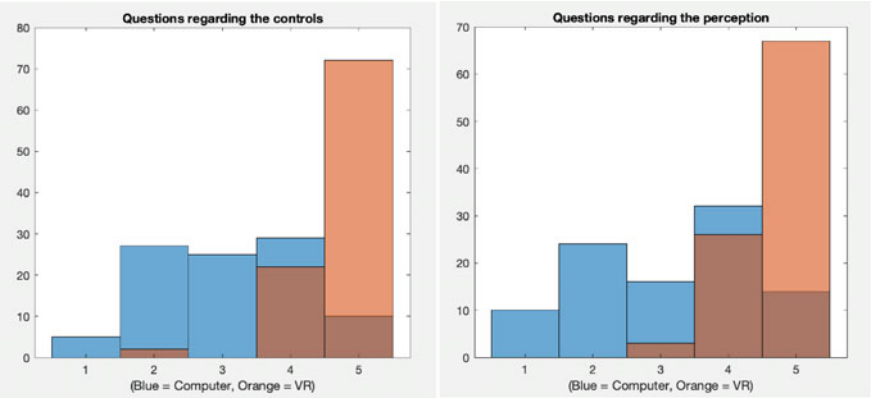

Fig. 9.6 Results from the two different categories of questions. Left: Questions for the controls. Right: Questions for the perception

Table 9.2 The means of the answers for each category and medium

	Controls	Perception
Overall	3.917	3.917
VR	4.708	4.667
Computer	3.125	3.167

a significant difference between means is found and that there is a small possibility that the difference between the groups happened by chance alone. Moreover, the effect size of the t-test result was calculated to be 0.16, which is a small effect.

As mentioned at the start of this section, a post-evaluation survey was carried out where the participants were asked to answer questions related to the two evaluated platforms. The results can be seen in Fig. 9.6, where all items addressing the same aspect have been added together. This was possible as the categories showed a Cronbach's alpha value of 0.81, meaning there is a good internal consistency, and thus reliability, within the answers [8]. Additionally, the mean of the answers, both overall for each category and separately for each medium is shown in Table 9.2.

9.14 General Discussion

The following section will discuss the results outlined in the result chapter above. It will examine the outcome of the t-tests and debate the opinions of the focus group interview.

Precision and Effect of 'Choir' Track

As mentioned in Sect. 9.13.2, the results from the evaluation fail to reject the null hypothesis that the VR version has no difference or less precision compared to the computer version. Even though the mean of accuracy is **0.14** units lower in the VR version, the difference is so little that no conclusion can be made. However, it was seen that especially one audio track seemed to cause problems in the VR version, the 'choir' track. The mean of the accuracy of the 'choir' track in the computer version was **1.86** while being **9.95** in the VR version, a difference of **8.09**, even though the audio track being identical. There are a few things to consider why that may be the case.

In the computer version, the position of the 'choir' track in the reference mix was very close to the initial position of the visible 'choir' track, while in the VR version the 'choir' track was positioned behind the user and with greater distance. However, that can be said about some of the other audio tracks as well. Another explanation could be that the 'choir' track, even though being converted to mono, had different amplitude changes for each voice and the recording included a natural reverb. This could have made it difficult to perceive at what distance as well as angle the source was placed in the scene. Some participants mentioned this, saying they felt the choir track was in stereo which hindered them from correctly placing the audio track. They simultaneously added that it was hard to navigate as the track faded in and out, making it difficult to know whether or not it was playing. By removing the choir track from both scenes the mean accuracy of the computer version goes from **35.74** to **33.88** while the VR average goes from **35.60** to **25.63**. Furthermore, the removal of the choir from both versions results in the rejection of the null hypothesis with a 95% confidence as mentioned in Sect. 9.13.2. One might, therefore, consider the fact that the VR version statistically can be seen as being significantly more precise than the computer version.

Efficiency

Another evaluated topic was whether the VR version was more efficient than the computer version. As mentioned in Sect. 9.13.2, the average time spent in VR was just under 2 min less than in the computer version. For the computer version furthermore,two outliers both spent just under 25 min. The time these outliners spent in the computer version does, however, not result in more accuracy as both were above the mean, **44.92** and **37.80**, respectively. However, even though it seemed that the participants on average were quicker in the VR version, there was no possibility of proving this statistically, as the data did not meet all of the parametric assumptions.

Looking at Fig. 9.6, its content might support the indication that VR, in this case, can be considered a quicker tool. Here it can be seen that the VR version scores significantly higher in regards to the controls and thus seemed more intuitive. Specifically, two questions stand out. Firstly, the question *'I felt comfortable using the controls in the computer version'* resulted in a completely split opinion as can be seen in Fig. 9.7.

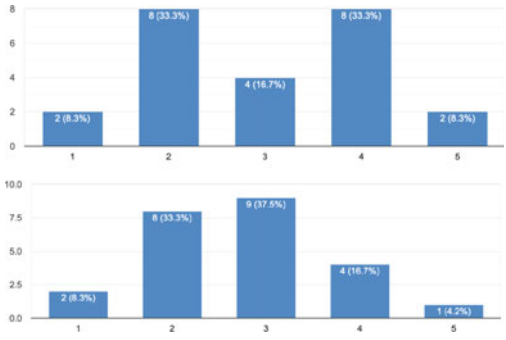

Fig. 9.7 Top: Question from post questionnaire regarding the comfortableness of the controllers in the computer version. Bottom: Question regarding the ease of placing sound sources in the computer version

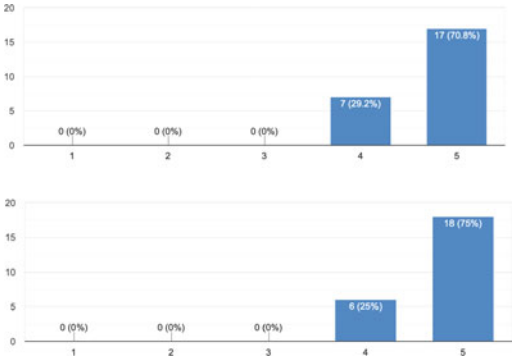

Fig. 9.8 Top: Question regarding comfortableness of the VR controllers. Bottom: Question regarding the ease of placing sound sources in VR

Secondly, when asked *'In the computer version, the controls made it easy to place the visual sound sources (spheres) where I wanted them'* the majority were either neutral or felt the controls made it difficult, as shown on Fig. 9.7.

Since the participants struggled with placing the sound sources at the desired position, combined with finding the controls uncomfortable, it might have resulted in them taking longer time replicating the reference mix. Therefore, it cannot be concluded that VR is more efficient, however, the concept of the VR version hints towards a more instinctive way of working. The same questions for the VR versions resulted in *slightly agree* or *agree*, as shown in the figure (Fig. 9.8).

Another important aspect to consider for the time data is why it did not meet the parametric assumption of normality. The data of the time spent in the different versions by the participants simply was not symmetrically distributed around the mean in each version. In both scenarios, outliers were experienced and especially the VR version included widely spread data. These conclusions are supported by the standard deviation of each data set being **325.21** and **248.17** s respectively. The mean in each data set is thus not a good representative measure of the participants and looking at the standard error being **66.3** and **50.5** s for each distribution, it additionally is seen that the sample mean does not represent the population specifically well. Therefore, there could have been a missing correlation in the time spent by the participants and one might discuss whether the participants were enough alike to be

considered a unified sample group/population. The experience in VR does not show any warning signs as almost everyone (83.3%) had tried it before. The questions then lie in whether or not mixing experience, as well as experience with and without spatial audio, could have affected the efficiency of the participants, creating a gap between their results. However, nothing supports this as 3 of the 5 outliers of time spent in VR had 5+ years of experience mixing music, which also was seen in 1 out of the 2 outliers in the computer version. Therefore, the whole test setup and the way that time is measured have to be reconsidered.

Focus Group Interview

Looking at the qualitative data from the focus group interview two main findings are clear: 1. The participants saw the product as a quick sketching tool to test ideas and outline mixes, rather than a tool to control precision and finer spatial details within the sound, and 2. The participants were overall positive about the interaction with the product and its visual appearance, sensory benefits and intuitive controls, which were generally well-received.

In relation to the first finding, several participants stated, just like the expert in the target group section, that time and intuitiveness, in general, was an important aspect for them in a mixing tool/device, and that the program indeed seemed to facilitate this. A participant expressed, amongst other things, that he 'could imagine that you would get done faster with some things', and that the program seemed 'very effective'. Furthermore, participants agreed that the VR program definitely could be used as a quick sketching tool for swift ideas and testing of audio placement in a given space. This could, amongst other things, have been a result of the intuitive way of placing sound sources as well as the quick dynamic sound feedback and the possibility to link it up with Ableton Live. However, it could also have been due to the simplicity of the environment and the fact that only fundamental controls, pre-made audio effects and interaction possibilities were included, giving it a 'to the bone' concept. Besides allowing positioning of sound sources (panning and volume), the program simply did not offer state-of-the-art possibilities such as the potential to manipulate the sound in finer detail, thus forcing the participant to use more time in the environment. This was moreover seen in the 'features' discussion of the interview, where participants emphasised a need for interactive dB scales, mute buttons and the possibility to 'do automation in a mix'. It is, therefore, clear that the opinion of the participants in the focus group interview, is somehow contradictory to the actual findings of the mixing task, which failed to show a significant time difference between the two conditions/programs. To understand this missing relationship, the post-evaluation survey as discussed above has to be taken into consideration. Besides the fact that the participants were overall satisfied with the interaction, it was clear in the opinion on the mixing task, that both the controls and the sensory feedback of the VR version made the experience easier, as well as more comfortable compared to the experience of spatially mixing audio on a computer screen. An aspect that, therefore, might have influenced the focus group participants' opinion on efficiency, could have been the general user experience—even though no statistical evidence showed that

the participants were quicker in the VR program compared to the computer version. Using VR simply seemed like an intuitive and instinctive way of placing sound sources spatially, as it utilises both visual and auditory perceptional topics known from the everyday.

Contrary to the time data, there actually was a significant difference in the precision of the mixes in the two environments, which as mentioned to some extent conflicts with the thoughts of the focus group. It is important to note that the difference is statistically evident without the 'Choir' track, but there is no doubt that it does not match the expectations of the focus group, who did not experience the program as something specifically precise. When compared to the framework of a DAW, there was a general consensus that the environment made it hard to be exact, and few participants even saw this as a result of the under-dimensioned mappings distance- and angle-wise. This was, as an example, expressed in the quote 'I found it slightly under-dimensioned so when you panned things to the side it was not as much as you would imagine'. This may be the result of having 5° interval between IRs on the azimuth but could also just be the effect of the binaural HRIRs as they represent how sounds are naturally perceived instead of allowing for absolute panning.

The focus group had different opinions on the visual appearance of the environment, though the majority agreed that the neutrality of it was beneficial. Besides some participants mentioning that having the room aesthetics change relative to the mood of the music being mixed, the visualisation of the audio in the process of mixing was experienced to be subjective. Therefore, having pre-defined visuals and scenery might be more disadvantageous than beneficial. However, few participants mentioned that if the appearance should change, it should be unrealistic, abstract visuals since if it was too realistic it would be distracting.

As mentioned above, there were a few features lacking in the program. Firstly, colour coding tracks according to instruments, as well as making different shapes depending on instruments lacked. Secondly, participants uttered that having feedback or audio-reactive shaders on the spheres representing the tracks would help the user to understand when a track has sound on it. Suggestions for this were either a VU-metre for each audio track (as mentioned above) or having the spheres change shape relative to the sounds they represent, using frequency or timbre to manipulate vertices. Additionally, the focus group agreed that adding more tracks in the VR environment would introduce clutter problems. As they only worked with five tracks, having more potentially could eliminate the benefits of VR compared to PC. Related to this, amongst other things it was stated: 'When we tried it here it was very manageable with five tracks, but if you have 67 tracks as we talked about, it might hinder you more than it helps'. A suggestion for this was being able to group tracks, 'Things could also be grouped visually, then you could press "show all" or see this group or this part [like in layers]. In that way, you could visually mute something'—as found in digital mixing consoles mentioned in the analysis.

In relation to the concept, some participants struggled to grasp the core idea behind the product, spatial mixing. The fact that the mix changed relative to head movement confused many participants and hindered them in understanding the possibilities and functionality of it. As one participant said, 'I often ended up looking one way and

then imagining that I mixed in stereo [...] this just made me feel that everything was in-precise'. Other mentioned that due to the lack of exporting it so the user could perceive it the same way, the product would be of no use, '...also the thing where the sound picture changes when you move. It is fun by itself but if it has to work in real life you have a dimension included that you do not have in the end'. The fact that they, in some way, did not seem to understand the thought process behind the product, could have resulted in them preferring to work in DAWs with traditional controls, which they were more comfortable with and, therefore, believe VR to be less accurate.

While it is apparent that recreating a mix in the VR version in some instances and for some instruments, was more precise than doing it in the computer version, the results of the evaluation show no statistical evidence that it was more efficient for the participants to work in the VR version compared to the computer version. These findings contradict the opinions of the focus group interview of experts within the field of music production, who experienced the VR program as a quick sketching tool for ideas and a spatial overview. The following discussion will debate these results, and look further into the design of the test and the fulfilment of the requirements made for this product and project.

Starting off by looking at the time data and the fact that the VR version was not more efficient than the computer version, it is important to consider the test setup and the way the researchers used time as a measurable variable. In the different test conditions, time was used as a dependent variable influenced by the independent variable being the two versions of the program. However, time simultaneously was of secondary importance, as the participants were told to recreate the mix until their satisfactory (until they thought they were precise enough). The amount of time used, and thus efficiency, was therefore not an explicit part of the mixing task and its validity as a measure, therefore, is debatable. Time was omitted from the task introduction as it was seen as a potential confounding variable of the precision measure, forcing the participants to slack on the mix recreation in order to get a quick time. Thus, it was hoped for time to represent the natural efficiency, however, the researchers did not reflect on the potential bias that could be within this, while designing the test—participants could as an example have had different visions about time, maybe they were busy or, contradictory, immersed using more time. Thus, in order for time to have been a valid measure, it might be argued that two tasks for each version should have been carried out: one with precision as seen in this evaluation, and one with time where the participants were asked to recreate the mix to their satisfaction as quickly as possible. This would have given efficiency a more prominent role and possibly made it a valid and streamlined measure.

Another aspect that could have been changed about the test setup was the methodology used. It initially was decided to use a one-way Repeated Measures ANOVA, with three conditions allowing the researchers to exclude the role of the controls in the test. Whereas the evaluation now only has the possibility to give indications about the role of the controls, the ANOVA test could have completely ruled this out, since a middle variable combing the two version was used for extra guarantee. Conclusions cannot, in fact, be made of the Likert scale in this post-evaluation, only indications

can be drawn, and therefore the scale's relevance to the test as well as the connection to the opinions of the target group interview should be considered carefully.

The opinions of the focus group and whether or not these, in general, are representative, is additionally an aspect that should be discussed. It was, as a starting point, decided to only use participants in each evaluation process that were experienced with mixing sound in one way or another. It was seen that both participants in the focus group interview (counting music production students), as well as people in the mixing task (including 'Sound and Music Computing' students, tonmeister and sound engineer students as well as professional and hobby producers), fulfilled these requirements. Even though the two sample groups might have had different visions, they represent the target group and it is thus assumed that results and opinions can be compared. However, the target group definition might in the first place have been wrong. As mentioned by the focus group participants, who only saw the product as an easy and quick sketching tool, its intuitiveness could be beneficial for beginners, who might not care about fine detail nor the lack stereo features. It could thus have been interesting to see the feedback and mixing task results of potential music production newcomers with less experience, to see if the sensory benefits and intuitive interaction of VR, would make even more sense in a beginner situation.

Additional aspect mentioned by the focus group was difficulties grasping the idea of mixing binaural audio. Multiple participants discussed how the fact that the audio mix changed relative to the head movement made it difficult to understand how the final mix would sound like. Especially the fact that if the mix was exported, they could not ensure that the end-user would hear it in the same way the mixing engineer intended. This was largely due to the fact that many of the focus group participants were locked on the idea of stereo mixing. The lack of a clear 'centre' position was new to them, as they are used to mixing in a fixed position in front of two speakers. With that said, the focus group agreed with the fact that the product enabled the user to very quickly and efficiently place sound at its correct position in the 3D space, possibly more quickly than with a computer. A possible application of the product was, as mentioned by one participant, placing audio sources at a correct position when working with movie sound. However, the lack of tools available, such as EQ, the possibility of choosing their own reverb, etc. would hinder them with mixing and, therefore, the product would be more suitable as an audio placement tool rather than mixing tool. Therefore, it is worth considering whether mixing music binaurally is suitable for current platforms and perhaps the focus of the test should have been on placing audio sources in a 'correct' position relative to the visuals when investigating accuracy. Additionally, an aspect that was not taken into consideration during the implementation of this project was the FoV. It is plausible that having a different FoV, be it bigger or smaller, could have affected the impression of the IRs. In other words, had the FoV been bigger, angles such as 90° or 270° may have mismatched the audio and visuals. The same can be said with a smaller FoV. This may have been a reason a member of the focus group felt the spatial algorithm to be under-dimensioned as he stated 'When you panned things to the side it was not as much as you would imagine [...] When I panned something this much to the right I would also expect to hear it more to the right'. Even though no other comments were made regarding the

'spatialness' of the sound, this is worth keeping in mind and it, looking back, could have been beneficial to make initial tests investigating the right relationship between the FoV and the sound.

With regard to the final evaluation, some technical problems were experienced. The main problem happening when the user soloed an audio track. Almost every participant (some more than once during the test) encountered a problem where the audio track's shader indicated that the track was soloed when in fact it was not. It is believed that this was caused by interference in the OSC (UDP) transmission between the computer running Unity and the one running Ableton Live. The problem took place when a track was soloed in Unity, which triggered the corresponding track to be soloed in Ableton Live. When the track was then un-soloed in Unity which updated the shader, the track stayed soloed in Ableton which caused confusion for the participant. This problem was something that was experienced during initial testing of the program and a specific trigger was created in order for the evaluator to quickly change/repair the state of the audio track in Ableton Live. This problem was almost non-existent in the VR version and could, therefore, have been a thing that affected the time measure. No initial test of controls and interaction was conducted before the final evaluation, whereas the VR version and controls had been tested with the focus group. The controls for the computer were thus purely decided by the project group, which may have resulted in non-intuitive controls and interaction, as was backed up by the post-evaluation survey. An initial evaluation or pilot test of the computer controls should have been conducted to ensure a more fair comparison.

In relation to the technical aspects, it also is worth discussing whether or not the HMD used for this project was the correct choice. The Oculus Quest was the chosen HMD due to it being wireless and thus consumer-relevant, providing the highest screen resolution, as well as having a satisfactory refresh rate. However, since it has the hardware built into the headset, the computational power is limited. Therefore, the whole Unity project had to be specifically optimised for the Quest and the complexity of the scene reduced. This may have come at a price of limited features, reactive shaders and objects. Having VU metres for each audio track or having the shape of the spheres change relative to the sounds they represent, may have made it easier for the user to distinguish which tracks were active. More or improved lighting may as well have improved the aesthetics of the environment as a whole, which could have been possible with another HMD such as the HTC Vive that solely relies on the power of a computer running the program. Using a different HMD could thus have optimised the VR program as well as the overall appeal and desired features of the focus group participants might have been fulfilled. However, wireless capabilities, ease of use and accessibility would, in this case, have been lost.

9.15 Conclusion

Based on an investigation of VR and spatial audio, it has been concluded that VR both has the sensory and interactive benefits to potentially enhance the experience of visually positioning sound sources in space. Its intuitive controls, sense of depth and user-including advantages are from research shown as instinctive behaviours and, from that standing point, this project examined whether or not aforementioned values could improve the process of spatial audio mixing, which nowadays mostly is carried out using 2D plugins on a computer screen. On this basis, a design framework for VR covering both interaction, binaural audio, perceptional cues, and graphical principles was built and an application was implemented to allow a user to visually mix real-time audio, retrieved from the DAW Ableton Live, using dynamic binaural synthesis. Each component and control of both the visual and the auditory system was carefully chosen based on requirements targeting both the benefits and necessities of pairing visuals and sound and in order to answer the final problem statement of the project, the VE was evaluated against its 'computer screen' counterpart.

Defined as measurable improvements by the projects researchers, the two main aspects 'efficiency' and 'precision', were evaluated in two different conditions: one evaluation using a focus group interview with experts examining the opinions, perceptions and feelings about the VE, and one evaluation consisting of a mixing task that quantitatively measured time and precision differences between the VR and the computer version. The evaluations were carried out over four days at Aalborg University Copenhagen, and 13 and 24 participants took part in each evaluation respectively.

The results of the evaluations showed ambiguous tendencies. The focus group participants were in general positively minded towards the program and saw it as a quick sketching tool due to its intuitiveness, controls and apparent sensory feedback, rather than a tool for finer detail and precision manoeuvring. These opinions were, however, not possible to prove in the mixing task test. Firstly, the data of time measurements did not meet the parametric assumption of normality, therefore it could not be tested through a t-test. Secondly, when comparing the means of the precision scores (average distance from reference mix) the t-test proved, that when removing the 'Choir' track—a track that caused problems for all participants—the VR version was more precise than the computer version with a 95% confidence. The evaluations thus indicate that even though the VR was not perceived as a precise tool, its sensory benefits and interaction possibilities, whose qualities both sample groups were in agreement about, are general improvements to the ones found on a computer. However, all conclusions should be taken with care, as especially the setup of the mixing task evaluation should have been reconsidered. A whole new test should, as an example, have been carried out to get efficiency as a valid measure and it simultaneously is important to note that the findings from the mixing task evaluation regarding the VE, can only be seen in the light of its 'computer screen' counterpart.

References

1. Adler, D.: Virtual audio-three-dimensional audio in virtual environments (Swedish Institute of Computer Science, 1996).
2. Argelaguet, F., Hoyet, L., Trico, N., Lécuyer, A.: The Role of Interaction in Virtual Embodiment: Effects of the Virtual Hand Representation in (Mar. 2016).
3. Bongers, B.: Physical Interfaces in the Electronic Arts Interaction Theory and Interfacing Techniques for Real-time Performance. Trends in Gestural Control of Music **2000**, 41–70 (2000).
4. Bordwell, D., Thompson, K.: Film Art an Introduction (1986).
5. Boring, S., Jurmu, M., Butz, A.: Scroll, tilt or move it: using mobile phones to continuously control pointers on large public displays in Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7 (2009), 161–168.
6. Cook, P. in A NIME Reader: Fifteen Years of New Interfaces for Musical Expression (eds Jensenius, A. R., Lyons, M. J.) 1–13 (Springer International Publishing, Cham, 2017).
7. De Man, B., Jillings, N., Stables, R.: Comparing Stage Metaphor Interfaces As A Controller For Stereo Position and Level in (Sept. 2018).
8. Field, A., Hole, G.: How to Design and Report Experiments (SAGE Publications, 2002).
9. Gale, W., Wakefield, J.: Investigating the use of virtual reality to solve the underlying problems with the 3D stage paradigm.
10. Gelineck, S., Büchert, M., Andersen, J.: Towards a more flexible and creative music mixing interface in (Apr. 2013), 733–738.
11. Gelineck, S., Korsgaard, D., Büchert, M.: Stage- vs. Channel-strip Metaphor: Comparing Performance when Adjusting Volume and Panning of a Single Channel in a Stereo Mix English. in Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2015) (ed Berdahl, E.) (Louisiana State University, June 2015), 343–346.
12. Gibson, D., Petersen, G.: The Art of Mixing: A Visual Guide to Recording, Engineering, and Production (MixBooks, 1997).
13. Gödde, M., Gabler, F., Siegmund, D., Braun, A.: Cinematic Narration in VR - Rethinking Film Conventions for 360 Degrees in (June 2018), 184–201.
14. Hamilton, R.: Building Interactive Networked Musical Environments Using q3osc in (Feb. 2009).
15. Holland, S., Mudd, T., Wilkie-Mckenna, K., McPherson, A., Wanderley, M. M.: New Directions in Music and Human-Computer Interaction 1st (Springer Publishing Company, Incorporated, 2019).
16. Jensenius, A. R., Lyons, M. J.: A NIME Reader: Fifteen Years of New Interfaces for Musical Expression (Springer, 2017).
17. Kohlrausch, A., van de Par, S.: Auditory-visual interaction: from fundamental research in cognitive psychology to (possible) applications in Human Vision and Electronic Imaging IV (eds Rogowitz, B. E., Pappas, T. N.) 3644 (SPIE, 1999), 34–44.
18. Machado, L., Moura, I.: Shader Integration in a Virtual Reality Framework in (May 2013).
19. Miner, N., Caudell, T.: Computational Requirements and Synchronization Issues for Virtual Acoustic Displays. Presence: Teleoper. Virtual Environ. **7**, 396–409 (Aug. 1998).
20. Nielsen, L. T. et al.: Missing the Point: An Exploration of How to Guide Users' Attention During Cinematic Virtual Reality in Proceedings of the 22Nd ACM Conference on Virtual Reality Software and Technology (ACM, Munich, Germany, 2016), 229–232.
21. Orland, K.: How fast does "virtual reality" have to be to look like "actual reality"? en. 2013.
22. Pursel, E.: Synthetic Vision: Visual Perception for Computer Generated Forces Using the Programmable Graphics Pipeline (Jan. 2004).
23. Ratcliffe, J.: Hand and Finger Motion-Controlled Audio Mixing Interface in NIME (2014).

24. Recanzone, G. H., Sutter, M. L.: The biological basis of audition. Annu. Rev. Psychol. **59**, 119–142 (2008).
25. Serafin, S., Erkut, C., Kojs, J., Nilsson, N., Nordahl, R.: Virtual Reality Musical Instruments: State of the Art, Design Principles, and Future Directions. English. Computer Music Journal **40**, 22–40 (2016).
26. Tabry,V., Zatorre, R. J.,Voss, P.: The influence of vision on sound localization abilities in both the horizontal and vertical planes. Frontiers in psychology **4**, 932 (2013).
27. Wang, X., Tokarchuk, L., Cuadrado, F., Poslad, S.: Adaptive Identification of Hashtags for Real-time Event Data Collection, 1–23.
28. Yantis, S., Richard, A. A.: Sensation and Perception (Worth Publishers, New York, NY, 2014).
29. Ye, J., Campbell, R., Page, T., Badni, K.: An investigation into the implementation of virtual reality technologies in support of conceptual design. Design Studies **27**, 77–97 (2006).
30. Zölzer, U.: DAFX: Digital Audio Effects (Wiley, 2011).