Audio for extended realities: a case-study informed exposition

Special Issue: Immersive Storytelling

CONVERGENCE

Audio for extended realities: A case study informed exposition

Convergence
2023, Vol. 0(0) 1–38

DOI: 10.1177/13548565231169723
journals.sagepub.com/home/con
Sage

Justin Paterson
University of West London, London, UK

Oliver Kadel
1.618 Digital Ltd, London, UK

Abstract

An area of immersive storytelling in rapid evolution is that of extended reality. This emergent mode of experience employs spatial mapping, and both plane and object detection to superimpose computer-generated images in the volumetric context of a physical space via a head-mounted display. This in turn produces a unique set of challenges and opportunities for associated audio implementation and aesthetics. Creative development of this audio is often a function of evolving toolsets, and the associated workflow is far from standardized. This paper forms a context for such audio workflow, one that draws from precursor technologies such as audio for games and virtual reality and develops this into an outline taxonomy that is both representative of the state of the art, and forward facing towards evolution of the technology stack. The context is framed through a series of case studies. Between 2019 and 2021, BBC and Oculus TV commissioned Alchemy Immersive and Atlantic Productions to produce virtual reality and mixed reality experiences of several classic documentary series by Sir David Attenborough: Museum Alive, Micro Monsters, First Life VR, Museum Alive AR and Kingdom of Plants. This portfolio received numerous award nominations and prizes, including from the Raindance Festival and a double Emmy. The sound design, audio postproduction and spatial audio for these experiences were implemented by the company 1.618 Digital, and drawing from first-hand-creator involvement, the workflows are deconstructed and explored with reference to tools, technologies, techniques and perception. Such an exposition forms the basis for an analysis of both this and broader creative practice in the field of audio for extended reality, and this is subsequently used to present a speculative vision of audio in the future of immersive storytelling.

Keywords

Immersive audio, virtual reality, mixed reality, workflow, immersive storytelling, binaural, sound design, Sir David Attenborough

Corresponding author:
Justin Paterson, University of West London, St Marys Rd, London W5 5RF, UK.
Email: justin.paterson@uwl.ac.uk

Introduction

The concept of immersive storytelling is far from new. A good book most certainly falls into that category. Whilst the word immersive has gained popularity alongside virtual reality (VR) and 3D audio, it is yet to be universally understood and applied. Lee (2020) conducted a literature review of numerous deployments and backdrops of the word and developed a framework for an 'immersive experience', broadly represented by the degree of intersection between physical presence, involvement and social/self-presence. The concept of storytelling perhaps needs less analysis, but it is commonly understood to involve narrative, imagery, evocation, imagination and engagement – which Lewis (2011: 505) encapsulated as 'cognitive processes and products of cognition'.

One of the great storytellers of the last half-century is the UK naturalist Sir David Attenborough – who has conveyed countless hours of insight into the natural world to a global TV audience.[1] Self-declared 'immersive-storytelling' company Alchemy Immersive (Alchemy Immersive, 2019, 2020, 2021a, 2021b, 2022) recreated a number of these classic TV series for various extended reality (XR)[2] formats between 2019 and 2022. These projects were:

- Micro Monsters with David Attenborough
- The Kingdom of Plants with David Attenborough
- David Attenborough's First Life VR
- Museum Alive with David Attenborough
- Museum Alive AR with David Attenborough

These projects were aimed at home use for the general public, although they typically required specific hardware, for instance a head-mounted display (HMD) or an iPhone. The intention was to amplify the experiential and educational aspects of the original TV series by immersing the viewer in a compelling virtual environment.

The audio – spatial sound, postproduction, and sound design – for these experiences was (re) created by the company 1.618 Digital, of which the Head of Audio is a co-author of this paper. Accordingly, by using these five projects as case studies, this text offers first-hand insight into the audio-production and authoring workflow and reveals some of the key elements of the creative process that helped to tell the stories. This will be framed by first highlighting some key psychoacoustical, technological and conceptual context that influenced the practical approach. Case studies one and four will introduce many key concepts in some detail, whereas the others will be more succinct since they draw from these concepts and exhibit production pipeline overlap, and so instead will offer a focus on sound design. Following this, an outline taxonomy of the workflow will be extrapolated from the case studies. Finally, a vision of the future of audio in XR immersive storytelling will be presented.

Related work and concepts

Psychoacoustical and technological

Underpinning technologies. Ambisonics is a technology that pervades the case studies. It "is an expandable, mathematics-based approach to spatial audio reproduction. It encompasses the encoding, storage, and rendering of directional auditory data by formulating the spatial sampling of an infinitesimal sound field such that it may be resynthesized by a finite number of point sources" (Armstrong and Kearney, 2021: 99). Unlike channel-*based* audio,[3] Ambisonics is agnostic to the

number of speakers required for playback, but there is a strict paradigm of configuration/quality considerations. It still comprises a number of constituent monoaural-audio channels, but these are encoded, and must be decoded for playback on any given system. 'The number of Ambisonic channels depends upon the *order* of Ambisonics being used. Higher orders require a greater number of channels and generally provide an increased spatial resolution at the cost of data storage, computation and increased complexity' (Armstrong and Kearney, 2021: 99). Another comprehensive discussion of Ambisonics is offered by Zotter and Frank (2019). Ambisonics can be decoded for headphone playback, and this mode lends itself to head-tracking whereby a sound source can be locked to a visual location (emitter) and maintain this relationship independently of head movements.

Another pervading technology is binaural playback. This is generally played via headphones and it endeavours to present spatial sound in the manner that humans hear rather than the way that 'a microphone hears'. 'Binaural signals can be typically obtained either by recording the sound scene at the eardrum or the ear canal entrance of a listener/dummy head or by synthesizing the virtual audio using HRTF filters' (Sunder, 2021: 137). A 'dummy head' is a recording device in the shape of a human head (like a mannequin), that hosts two microphones in the ears. This morphology causes the audio magnitude and phase that reaches each of the ear-mics to be filtered by its shape. Alternatively, this filtration can be emulated – synthesized – through head-related transfer functions (HRTFs) applied to monoaural signals; this is a dummy-head-equivalent 'transfer function that describes the acoustic propagation between the same source and the listener's ears' (Sunder, 2021: 131). Much more information can be found in Blauert and Braasch (2020) and Litovsky et al. (2021).

Interactive audio: From games to extended reality. Goodwin (2019a) summarizes the contemporary video-game approaches to synthesizing audio environments. In game engines, 3D geometry is commonly modelled for both graphics and physics, the latter being computationally much cheaper, and its simplified representations of objects lends itself to also being deployed for audio in order to replicate the reflections and occlusions associated with an object or surface, although differences between an object's collision and acoustic properties can create difficulties. To emulate an acoustic profile, a given object intended to interact with audio needs to convey two parameters for the degree of reflectivity and the resulting timbre. For volumetric purposes, a third class of geometry describes sound-area meshes and effect-area meshes to define sound emitter zones and reverberation parameters, respectively. Fırat et al. (2022) discuss specific production methods for deploying 3D audio and Goodwin (2019b) presents further aspects of such geometry-oriented workflow.

VR endeavours to operate in a similar matter, but with 'presence' – a sense of being there – being a key objective (Nordahl and Nilsson, 2014).[4] The general conventions of sound design and implementation practices in both are largely the same as games, and a useful discussion can be found in Schütze and Irwin-Schütze (2018). Silzle (2008) identified aspects of user-perceived immersive-audio 'quality' in virtual environments, including timbre, both location and dynamic accuracy, loudness balance, auditory spaciousness, reverberation, and artefacts. However, with the emergence of mixed reality (MR), new opportunities and challenges are emerging. In this medium, the virtual content is presented within one's immediate real-world surroundings which sets a different level of expectation from the user's point of view, requiring acoustic cues more representative of the specific real world around the user (Llewellyn and Paterson, 2021). As Devereaux (2020) observes, 'XR may be like the Wild West in the sheer combinatorial vastness of its novelty, dangers and opportunities'.

Virtual-acoustic considerations. It is well understood that aural perception of the surrounding environment is a function of its geometry, scale and absorption. The acoustic fingerprint of a space (including the objects within it) is often represented by its room impulse response (RIR), but it is complicated by the observer's six-degrees-of-freedom (6DoF) motion within it, further by objects of acoustically significant scale moving within it, and further again by sounds that they might emanate. Traditionally, the real-time synthesis of such an acoustic environment – auralization – has been technically challenging and computationally expensive, but the field of 'virtual room acoustics' has been developing in the last 20 years or so (Vorländer et al., 2013). More computationally efficient rendering of area-volumetric sources (Schissler et al., 2016) is developing, and Jot and Lee (2016) proposed the "reverberation fingerprint" model as characterizing an acoustic space for the overlaying of virtual sound sources in XR. More recently, simulating more acoustically complex "multiple enclosures coupled via an aperture" via scattering delay networks (Atalay et al., 2022) has become possible. However, it is still acknowledged that there is insufficient data on auralization tools and practices (Thery et al., 2019), but the 'kansei' effect – 'presence, verisimilitude, realism and naturalness' is understood to enhance the quality of audio communication (Suzuki et al., 2012) and hence is a goal for immersive storytelling.

Audio-visual perception. The spectral cues for localization in the horizontal plane are superior to those in the vertical plane (Blauert, 1997), and Lee (2021) further discusses the psychoacoustics of height perception. Lee and Lee (2017): (59) stress the importance of optimizing immersion with the correct balance of ambience volume; 'providing auditory stimulation as a harmonized and complementary means of information transfer'. Visual information can help to improve the accuracy of sound-localization estimates (Ahrens et al., 2019), and audio and visual synchrony can improve cognition when paired stimuli co-occur within a limited time; the 'temporal binding window' (TBW) (Zhou et al., 2020). 'It is likely that a spatial mismatch between visual gaze and auditory attention leads to increased task demands, as reflected in slowed [reaction times] during the incoherent condition' (Pomper and Chait, 2017) and if a spectator looks toward a sound object with only their eyes whilst keeping their head position static, binaural cue anomalies can be introduced leading to a divergence of aural and visual localization (Maddox et al., 2014).

Better externalization during binaural sound reproduction (BSR) was observed in virtual rooms that offered visual congruence with the audio localization (Werner et al., 2016). Brinkmann et al. (2016) exposed the participants to a VR wasp with the headphone-audio rendered in different modes. The binaural audio was shown to be capable of significantly increasing self-reported anxiety. In an internal study, Sennheiser demonstrated that BSR externalization can reduce auditory fatigue amongst call-centre operatives (Immersive Audio Podcast, 2022). The relatively narrow field of view offered by current MR headsets mitigates this (to a degree) by 'forcing' the head to turn to bring emitters into the line of sight, and eye tracking within an HMD is becoming more commonplace too, although a physical HMD has been shown to degrade audio localization (Genovese et al., 2018; Ahrens et al., 2019). However, deployment of spatial audio provides information about events beyond the field of vision (Hermann and Ritter, 2004) and in virtual environments, spatialized audio significantly increases a listener's sense of presence (Blauert, 1997).

Audio cognition. Bregman (1994) sets out the challenges of auditory-scene analysis (ASA), amongst which are understanding the human assimilation of sequential and concurrent sonic elements in order to form a perceptual stream, and then segregate actual conflicting simultaneous sources or temporal sound patterns – in order to actually identify them.[5] Yost et al. (1996) then demonstrated that binaural listening mitigated the 'cocktail-party effect' – source determination amongst

numerous competing audio stimuli – compared to conventional playback using a single headphone. Using over-ear headphones, speech intelligibility over noise has been shown to be as effective in binaurally decoded first-order Ambisonics (FOA) as in higher orders (Dagan et al., 2019) inferring that the higher channel count of the latter is unnecessary for intelligibility, although it might still be preferable if higher spatial resolution is required. Dynamic placement of the audio resulted in increased spatial release from masking (SRM) for listeners, offering enhanced speech intelligibility when spatialized against noise. Vazquez-Alvarez and Brewster (2010): (256) investigated cognitive load whilst comparing simultaneous versus sequential presentation of multiple audio streams, 3D-audio to spatially localize audio sources versus a single point of presentation, and dynamic movement versus fixed locations of sources. They determined 'a trend for participants preferring the spatially fixed' yet also noted that use of spatial audio improved recall against single-point presentation with simultaneous streams.

Dynamic binaural synthesis (with head movement)[6] can compromise the perceived 'authenticity' of audio,[7] although this was most noticeable for pink noise – in preference to music – and mitigated by reverberation (Brinkmann et al., 2017). VR users typically underestimate distance and increasing the reverberation time associated with an audio object can help to mitigate this (Huang et al., 2021) although in the near field, this can cause audio-visual sensory segregation. However, due to its pass-through nature, in MR, users have a precise visual perspective and expect audio cues to match, and techniques are being developed to facilitate this (Audfray and Jot, 2019; Gupta et al., 2022). Further, if the reverberation introduces interaural differences beyond those of the source object, externalization is improved (Leclère et al., 2019). Sunder (2021: 130) details personalization of HRTFs and how they relate to localization resolution, stating that they 'are a critical component of binaural audio, and they determine much of the quality of any binaural 3D sound that we experience. Human pinnae are as unique as fingerprints and therefore, HRTFs are highly idiosyncratic as well'.

Mixed-reality acoustic modelling. Spatial mapping – whereby a system has the ability to analyse the physical space around it – has been evolving for quite some time; Behringer et al. (1999) provide a useful early perspective. Magic Leap headsets offer a contemporary implementation, first modelling static elements of the surroundings with a point cloud – 'world features', and then with 'world models' – increasing detail with meshes and also detecting planes (eg tabletops) that are useful for augmenting with virtual objects (Magic Leap, n.d.). The current implementation cannot associate audio reflections with individual planes,[8] although the application programming interface is evolving (Audfray et al., 2018) and an environmental reverberation model is being parameterized, for example, to characterize 'the reverb_gain', 'the total energy of the reverb when the listener and the source are collocated' (Audfray and Jot, 2019: 2). This has been extended by the addition of clustered reflections to provide distance simulation for individual audio objects, allowing 6DoF navigation in the Metaverse (Jot et al., 2021).[9]

The concept of sound design

Contextualization. A further aspect worthy of brief contextualization is the creative implementation of sound design, which has a powerful effect on the narrative. The term was first coined by Walter Murch in 1979 (Whittington, 2009) and is often given slightly different definitions. It generally refers to the creation and deployment of all audio elements associated with a moving image, although often excluding music. These might typically include dialogue, Foley sounds, special effects, and 'atmos' tracks.

- Dialogue is any spoken word; it is commonly recreated in an audio studio following filming in order to control and optimize sound quality, a process termed automated dialogue replacement (ADR).
- Foley sounds are those created by everyday movement – these must also be recreated since the original recording will have had them stripped away to make way for ADR, and they are essential for a sense of realism. They are often divided into three categories: walking, props and cloth.
- Special effects (SFX) are all the incidental sounds beyond Foley, for example, engine noise, a gunshot etc. In the 1930s (the era of the radio play), the BBC categorized SFX as being realistic (confirmatory or evocative), symbolic, conventionalized, impressionistic or music-as-an-effect (Moffat et al., 2019). It is common if recording the actual authentic source for it to sound underwhelming, and so these sounds are often contrived from other (often multiple) sources. In recent decades, drawing parallels from visual arts, a sound design trend referred to as 'hyperrealism' has become popular, with the objective of making the SFX larger than life for dramatic effect. It 'questions the existence of objective reality by taking into account […] artefacts and treating them as objects' (Puronas, 2014). One example might be the sound of a loudly beating heart when representing panic.
- Atmos tracks are environmental sounds that typify and enhance the setting of the visuals, for example, cicada chirps in a night scene or wind noise on a mountain top. They can be complex collages to represent a dynamic environment, for instance to reflect multiple camera angles or movement between different spaces.
- A sound that originates from within the 'story world' is termed as being diegetic; its source can either be on or off screen. If a sound is only apparent to the audience, for instance narration or soundtrack music then the sound is non-diegetic (FilmSound.org, n.d.).

Approaches. Sound designers typically have an adaptable palette of approaches that are repeatedly deployed where there might not be need to reinvent the nature of a sound, for instance a Foley footstep on concrete (although there might). Beyond this, they take pride and pleasure from devising novel and creative sonic solutions, for instance Ben Burtt's lightsabre sound in the first (1977) Star Wars movie, which was derived from the motors in a projector superimposed upon microphone interference from a cathode-ray-tube TV to generate a static hum. The motion of the lightsabres was added to by playing back that composite from a speaker and re-recording the sound whilst swinging the microphone to emulate the actors' fight movements, which induced a motion-synchronized Doppler-effect pitch sweep (*Ben Burtt Interview: The Sound of Lightsabers*, 2014). Beyond XR, as spatial audio increasingly becomes a part of different media such as movies and gaming, such novel approaches are pursued where time and budget allow, and the 3D sound stage offers fresh opportunities. In the following case studies, sonic inspiration was drawn from each of the original Attenborough TV series, Burtt's spirit of ingenuity was a pervading influence, and Murch's attitude and work constantly set the benchmark.

Further reading

The interested reader can find further detail of spatial audio phenomena and technologies in sources like Paterson and Lee (2021), and Roginska and Geluso (2017). The chapters in these multi-author editions also offer numerous extensive reference lists. Sonnenschein (2001) and Avarese (2017) offer a detailed discussion of various conceptual and practical aspects of sound design, their interaction and hybridized emotional effect.

Case Studies

Case study 1: Micro monsters with David Attenborough – contextual analysis

"Micro Monsters" originally aired on TV in 2013 – the title slide is shown in Figure 1. A VR version of the series was commissioned by Oculus TV (2020). It consisted of 5 five-minute episodes produced for the launch of the 'Meta Quest 2' HMD. This project employed novel high-resolution 3D camera capture and stitching to deliver 180° video, using the original narration by Sir David Attenborough. Various arthropods were filmed and upscaled to appear enormous in order to reveal detail to users. The brief offered a creative licence for 3D sound design that blended conventional sonic paradigms for documentary and gaming into a hybrid approach. Traditionally, audio post-production conventions for TV documentaries were based on channel-based surround formats such as stereo or 5.1. Since 2017, Next Generation Audio (NGA) has facilitated more hybrid approaches (EBU, 2017; Füg et al., 2016; Olivieri et al., 2019). Here, combining BSR audio-object 3D panning with three-degrees-of-freedom (3DoF) video offered a multimodal immersive experience.

Alchemy Immersive worked closely with the team at Oculus, whose parent company Meta offered a technological increment by further developing the audio-playback pipeline – to enable two

Figure 1. Micro Monsters project title poster.

independent audio streams to be 'muxed' (multiplexed) with the video stream for the final content distribution – the first time that this was implemented on the Oculus content platform, although it only brought it in line with extant FB360-type rendering functionality. However, this increment superseded the prior (YouTube) audio-delivery protocol for 360° video, where all elements of the program needed to be embedded in a single FOA AmbiX file (YouTube, n.d.) – which precludes certain flexibility, for instance when wishing to commit components of a mix to either a head-locked or head-tracked-stream.[10]

Sound design and content acquisition. The ambient tracks from the original live-action scenes were evaluated for repurposing, but they were mostly unusable due to having been archived only as stems in a manner that was unsuitable for the desired spatialization workflow; some individual sound effects had been rendered with the atmos tracks,[11] and so could not be isolated as monaural sources (as is optimal) for 3D spatialization in the game engine. Instead, numerous stereo or mid-side (Rumsey and McCormick, 2009: 505) replacement tracks were sourced from existing libraries, then layered and upmixed to third-order Ambisonics (3OA)[12] to create enveloping sonic environments. Despite the 180° visuals, the audio was spatialized to 360° (using FB360 in Pro Tools Ultimate) which provided a greater sense of presence (Hendrix and Barfield, 1996) and envelopment, and mitigated perceived discontinuities at the boundaries of the 180o-field during a user's head rotation. Whilst the audience would concentrate on forward-facing visuals, the omnidirectional ambient backdrop provided multimodal continuity and immersion (Hermann and Ritter, 2004). The 'risk' of taking this 360/180 approach was precipitating the 'exit-door effect', as discussed by Elvemo (2013) and others. However, Elvemo attributes resolution of such spatial conflicts (in a 2D-cinema-visuals-plus-surround-sound scenario) to come from a combination of perceptual cognition and memory (Vazquez-Alvarez and Brewster, 2009; Vazquez Alvarez and Brewster, 2010). In comparison to that scenario, a dynamic head-tracked sound field – with a height component – amplifies both cognitive modes, and so careful and sympathetic sound design can act to principally dissolve front/rear transition boundaries without introducing distraction.

Implementation and playback optimization. The positions of the dynamic BSR of the moving creatures (sound emitters) and ambiences need to be synchronous with the visuals for perceived authenticity and to lessen potential psychological tensions (Pomper and Chait, 2017; Zhou et al., 2020) as discussed in the 'Related Work and Concepts' section. With head-motion, the HRTF-induced attenuation and filtration enhance the perception of emitter motion and hence authenticity. However, once audio undergoes any such spatialization via binaural rendering, it can also lose spectral definition and intensity of dynamic range as a function of head position (Brinkmann et al., 2017). Binaural spatialization can improve utterance detection in a cacophonous audio environment (Yost et al., 1996) which suggests that a spatialized voice-over narration (VO) could be more intelligible even against more dramatic volume increases in the 3OA content (hence permitting them) and the use of spatial separation (eg VO versus music) can mitigate spectral-overlap masking (Ihlefeld and Shinn-Cunningham, 2008) in voices – improving clarity.

This is related to the timbral-fidelity versus spatial-fidelity 'conflict', as investigated by Rumsey et al. (2005), which can problematic in music production. However, anecdotal discussions within the Micro Monsters team suggested – perhaps surprisingly – that it was more so with more 'highly-produced' musical genres where there were fewer original acoustic properties from the recordings preserved in the final mix. Brinkmann et al. (2017) found fewer such problems with acoustic music than with pink noise, but did not investigate other genres.

Following the upgraded delivery format from Meta, the full audio mix could consist of 1) a dynamically spatialized BSR 3OA mix-stem containing the ambiences and individual diegetic audio-objects – the creatures, and 2) a two-channel stem that was head-locked and might carry non-diegetic elements.[13] In consideration of all the above trade-offs plus Vazquez-Alvarez and Brewster's (2010) user preference, it was decided that it would be distracting if VO and music changed position as the user looked around in a 3DoF environment, and so these were placed in the second stem The VO was placed as a centrally panned monaural source, hence perceived to be originating 'in the head' whilst the music still offered stereo width, as its mix intended, together maximizing clarity in line with the findings of Yost et al. (1996), as also discussed in the 'asset preperation and implementation' section of Case Study 1. This stem also carried synthetically produced low-frequency cinematic SFX. Such content was generally treated as music, providing impact without distracting from the narrative.

The storytelling and gameplay (user-interaction) aspects were also affected by the 'fidelity-conflict' issue (above), and so the two-stem playback proved important to ensure that the holistic audio ecology could be delivered without compromise. The educational aspects of the experience might also be improved since spatial staging slightly improves recall (Vazquez-Alvarez and Brewster, 2009). However, it is interesting to note that Rumsey et al. (2005) also determined (of horizontal surround) that naive listeners had little awareness of spatial-audio fidelity in the front 60° – which is interesting to compare with Ahrens et al. (2019) in an audio-visual (AV) context, where visual information contributed less to frontal sound-localization estimates. Accordingly, through discussions with the team, such decisions were made on which approach (head-locked or head-tracked) better delivered cognition and information retention in each specific context – always balanced against aesthetic concerns. For instance, occasionally, some SFX were situated or duplicated into the spatializer to highlight a certain directionality; these were given additional high-frequency content to aid localization.

Video editing can be seen in Figure 2. Meta's two-stream playback also allowed for the management of the overall dynamic range. This was particularly important when portraying softness and intimacy in the quietest scenes, perhaps followed by a very loud and climactic passage of music and SFX. However, since the HMD headphone preamp was not overly powerful, it was deemed necessary to take steps to also optimize overall program loudness. Although specialist tools for 3OA signal-processing can be deployed to manipulate dynamic range in the BSR domain, the perception of localization is extremely sensitive to any inconsistencies or unnatural anomalies in the spectral domain (Gutierrez-Parera and Lopez, 2016), and the music was of course head-locked. In Paterson and Llewellyn (2019: 173), Martin states 'forget compression – it's the opposite of what you want', claiming that whilst in stereo, compression is commonly used to create punch and fatness, in 3D, it can instead render a "canned" sound. Dynamic-range compression can also distort the spatial image of the sound field (Wiggins and Seeber, 2011),[14] therefore inducing the type of problems alluded to in the 'audio-visual perception' section of 'Related Work and Concepts'. Accordingly, the spatialized stem underwent minimal dynamic processing to both maintain optimal localization and of course, preserve the naturalness of the sound.

Conversely, had the score, narration and cinematic SFX been integrated within the spatialized 3D stream, they would need to have been much quieter in order to avoid clipping. An Ambisonic file rendered out with insufficient headroom could clip the binaural decoder on a streaming platform and distort upon playback. As a precautionary measure during production, a multichannel limiter would be typically be applied and set to – 6dBFS according to the manufacturer's recommendation (Noise Makers, n.d.)[15]. However, it was found that in practice, setting between −2 to –3dBLUFS below maximum peak level (see Figure 3) produced no audible artefacts on this material and this target was

Figure 2. Video editing in Adobe Premiere – note the immersive-simulation view at the top of the picture.

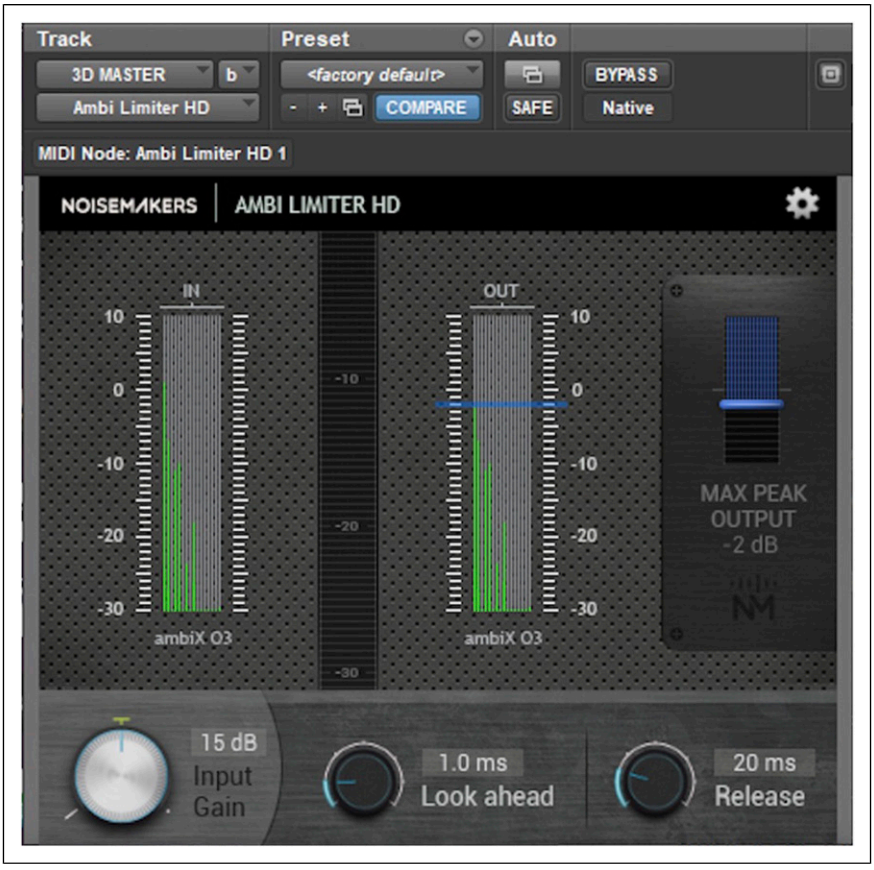

Figure 3. A multichannel limiter.

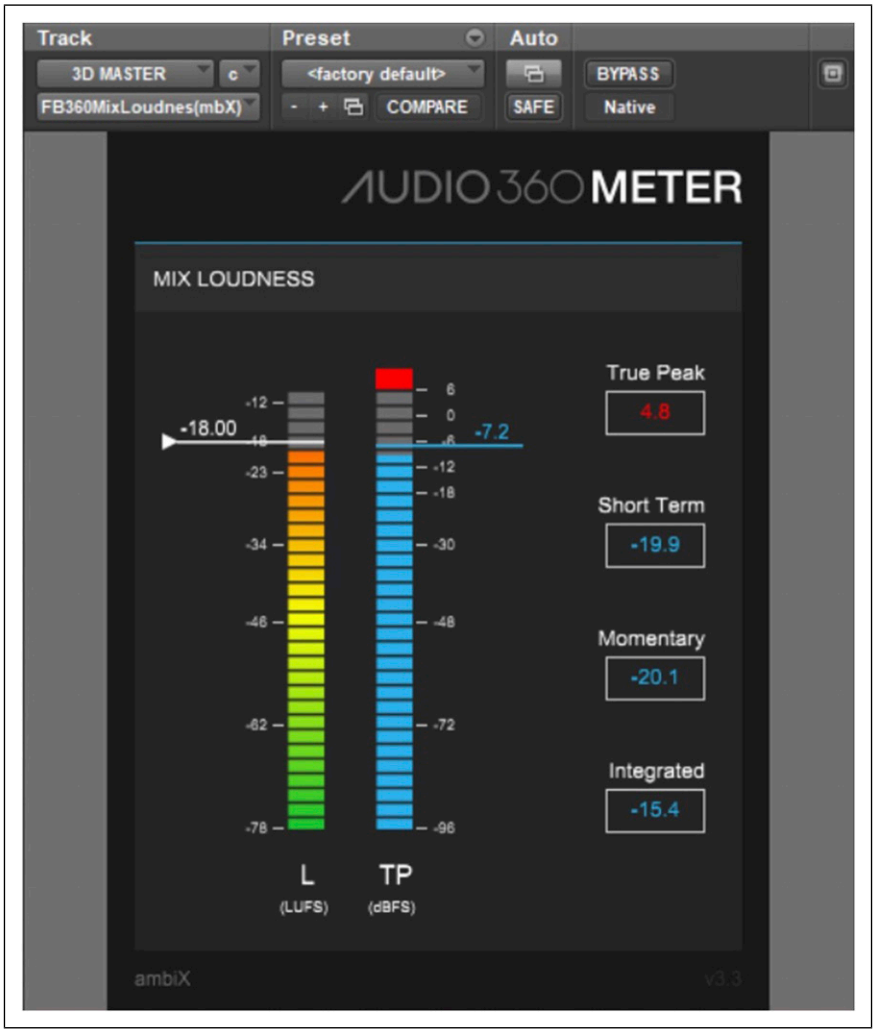

Figure 4. FB360 audio 360 meter.

therefore used to optimize loudness. Furthermore, it was imperative to also utilize the FB360 Audio 360 Meter shown in Figure 4, to monitor loudness-peak variation due to rotational HRTF attenuation, which might be by several decibels (frequency-dependent). Despite this, being able to use compression to 'master' the head-locked stem separately allowed it to be much louder in comparison to spatialized diegetic content, thus delivering an overall mix with the optimal methodologies across two mixing paradigms. A film of 'behind the scenes' is available at Alchemy Immersive (2020).

Case study 2: The kingdom of plants with David Attenborough – sound-design analysis

"The Kingdom of Plants" series originally aired on TV in 2012. An immersive VR series consisting of 3 five-minute episodes was commissioned by Meta and Oculus TV in 2021. The series featured time-lapsed scenes of exotic flowers blooming, insectivorous sundew and Venus flytrap plants, fungi and a microscopic view of tiny seeds. It was shot over several weeks both at Kew Gardens and various other locations around the world and again narrated by Sir David Attenborough – the title slide can be seen in Figure 5.

As discussed in Case Study 1, Micro Monsters had necessitated the development of a novel authoring pipeline, and its release and subsequent critical acclaim (see 'Awarding Organizations (wins and nominations)' section) suggested that the approach was successful, and so the production blueprint was redeployed for the Kingdom of Plants project, which had broadly similar objectives. However, there were some different aspects; it employed a superior high-resolution (8k 3D, 60fps) 3D camera that captured various arthropods and plants, which were again upscaled to appear enormous, and stitched to deliver live-action 180° video combined with bespoke computer-generated imagery (CGI) in VR. It largely presented a similar set of challenges and opportunities for audio, and as such, much of its technical workflow mirrored Micro Monsters with the two-stream approach and so does not need to be reiterated here, but the sound-design brief offered significant scope for a creative licence to auralize footage derived from macro cinematography. The core objective was again to support making the key protagonists appear larger than life, and this time, to also project anthropomorphic characterization onto organisms that often functioned in a manner imperceptible to human senses – perhaps too slow for the eyes or too quiet for the ears. This was implemented by assigning sonic motifs to each character with a hyper-realistic sound design.[16] (Figures 6 and 7).

Sound design. Various impressionistic techniques were employed to auralize dynamic footage of plants, for example, the sound of a Foley artist throwing a lasso to emulate the plant's movement when attaching itself onto taller plants over time-lapse footage (see Figure 8), or a processed sample

Figure 5. Title poster for Kingdom of Plants.

Figure 6. Stills from the Kingdom of Plants films.

Figure 7. Filming set in Kew Gardens.

of a bomb explosion accompanying the opening of a majestic bloom. Inspired by the sound-design work of Diego Stocco (2022), electrical signals were sourced from plant biodata and converted into MIDI notes and control changes, and then fed into a synthesizer.[17] The nature of this data gave a chaotic sonic palette that required judicious editing, but this approach preserved 'conceptual authenticity'[18] and led to the creation of a suitable high-frequency-rich texture that could intensify the visuals of a swarming aphid colony gradually eating a pine tree. In a different study, cross-modal stimulus has been demonstrated whereby participants were presented with square blocks in a virtual

environment, and by binaurally playing back the sound of particular tactual textures (eg sandpaper) as users interacted with them using bare hands, illusory pseudo-tactile responses were elicited. 'The pseudo-tactile sensations were also found to be robust, with 11 out of 18 participants who described a sense of touch also discussing the impossibility of this experience (without being prompted to do so)' (Bosman et al., 2021).

Playing on the wasp experiment of Brinkmann et al. (2016), when offering such visual proximity to aphids, an additional idea for this high-frequency chaos was to attempt to elicit a cross-modal 'spine-tingling' sensation through the autonomous sensory meridian response (ASMR) phenomenon; 'typically characterized by electrostatic-like tingling across the scalp, following the line of the spine downwards, extending to the arms and further depending on the intensity of the response' (Barratt et al., 2017). Whilst the actual precipitation of an ASMR response in an audience is open to considerable chance, it was thought to be useful metaphorical approach regardless.

Artistic licence was also taken both for footage that represented plant movement for which the natural sounds were too quiet to record, and also for time-lapse montages where use of the real-time sound was not feasible. So, for these, various organic and synthetic samples were spotted simply to enhance the narrative (Figure 9). Gestalt psychology has its origins in visual perception and might be summarized by the common expression 'the whole is greater than the sum of its parts'. Sonnenschein (2001) described an equivalent *sonic* gestalt concept of sound design which supported this artistic licence; a pervading subtext that always endeavoured to hybridize constituent elements in order to contribute emotive stimuli beyond the ostensible holistic experience.

In the episodes that were recreated, Sir David Attenborough only features as a narrator and does not appear on camera. In one scene, he discussed a once-a-year spectacle in which an 'epiphyllum oxypetalum' cactus blooms for only a single night. To make this scene intimate and bring the viewers closer to the moment, the relevant portion of the narration was externalized by placing the dialogue in 3D space just out of frame to immerse the viewer and give the impression of the host

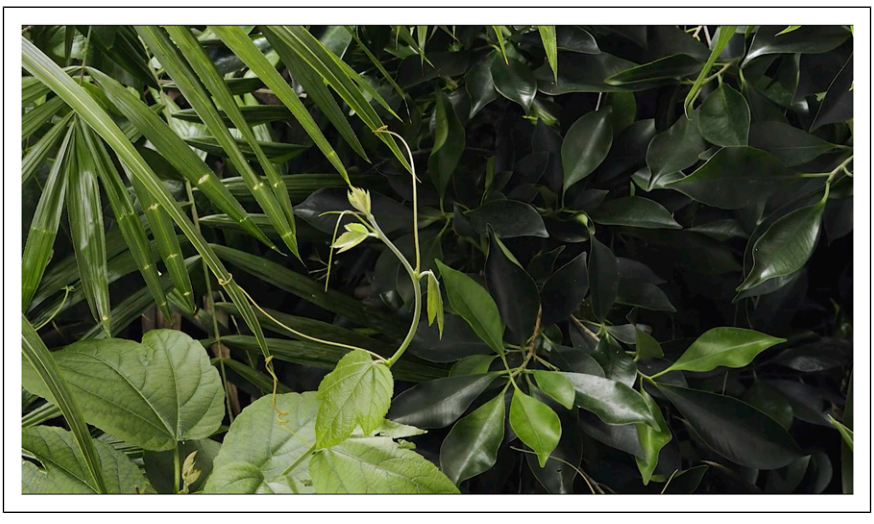

Figure 8. The tendril that inspired the lasso characterization.

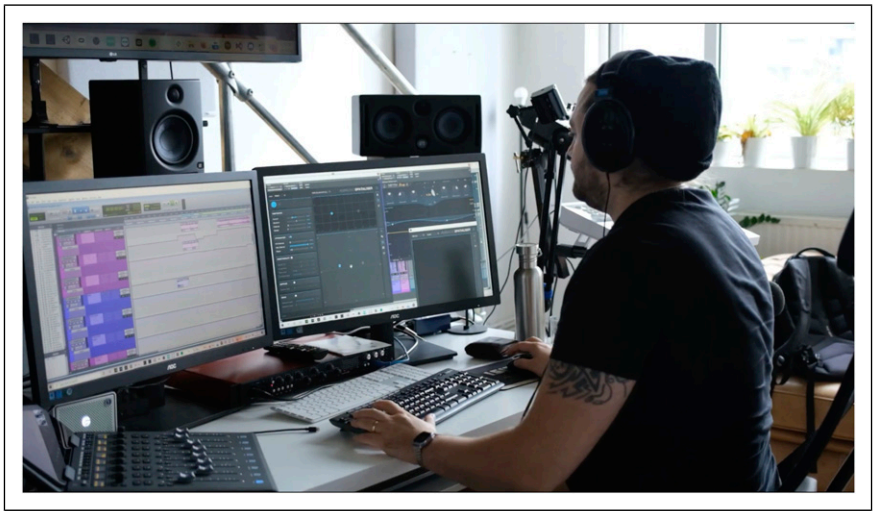

Figure 9. The audio-editing/sound-design workstation; the FB360 spatializer in Pro Tools Ultimate.

standing beside them in the room.[19] As stated in the 'implementation and playback optimization' section of Case Study 1, a head-locked narration is generally preferable; however, this 'off-screen' approach can be effective for short periods of non-diegetic narration and precipitates a sense of proxemic experience (Hall, 1963) that can cultivate a deeper sense of presence for the audience – to support the kansei effect (Suzuki et al., 2012).

Overall, the approach to sound design again combined externalized-3D with internalized audio to deliver complex immersive experiences. Although this had previously been possible in game engine workflow, this was the first time it had been done with 180° video.

Case study 3: David Attenborough's first life VR – sound-design analysis

The 2010 Sky TV documentary series 'David Attenborough's First Life' depicted the origins of life, through the earliest multicellular organisms from the Ediacaran biota to the first complex creatures of the Cambrian Explosion, some tens of million years later. Again, Alchemy Immersive produced a VR version of this with Meta Quest, also in 2021 (Figure 10). In contrast to the preceding titles, this one used computer-generated models with a focus on scientific accuracy (collaborating with Zoo VFX) in pre-rendered stereoscopic 8k at 60fps – the increment being that this time it was in full 360° video. According to Ghandi in 'David Attenborough's First Life VR | Official Behind The Scenes 2021' (Alchemy Immersive, 2021a), this was the first time that this resolution/format had been successfully rendered commercially.

The goal of First Life VR was to immerse viewers in the Earth's ancient oceans. From an audio perspective, the animations of numerous primitive sea creatures such as the five-eyed Opabinia, the shrimp-like Anomalocaris, and the spiny Hallucigenia needed to be given sonic profiles that could relate to their morphology and motion in order to 'bring them to life'.

Although unlike the earlier projects, this one offered full 3DoF visuals, the prior audio had always been in 360°. As such, there was no need to adjust the audio-production pipeline and this

project stood to just consolidate the established workflow. However, there were further aspects of sound design that are potentially of interest.

Sound design. The video concept was for the camera to continuously move forward not only in space but also in time – sometimes travelling millions of years between scenes. Whilst spending the entire duration of the experience underwater, the audience needed to appreciate the multidimensionality of time and space – including related aspects such as the chemical composition, light and temperature of those environments. Sound design for these underwater worlds again presented both challenges and creative opportunities (Figure 11). One strategic decision was how to auralize an aquatic environment that no one had ever heard before. However, although primaeval oceans would likely have slightly different hydroacoustic behaviour than today's due to their lower salinity, this would not likely make a significant difference to their perceived frequencies which are independent of salinity or temperature-induced velocity variations. Further, most users will carry some preconceptions of what sound is like underwater, and so recreating-then-manipulating such effects provided firstly familiarity, and then suitable 'otherworldliness' for the visuals under dramatic license for storytelling.

To deliver this, a library of underwater recordings was captured in various locations around the UK using a professional-grade marine hydrophone (Aquarian H2A), and – submerged – a FOA microphone (Sennheiser Ambeo), a contact microphone (Crank Sturgeon Classic), a Zoom H2n in double M/S mode, to record the underwater atmos. Recording even ostensibly similar material at different locations offered a large palette of natural underwater ambiences and in the end, this mitigated the need for extensive manipulation in postproduction to create a sufficient variety of sound textures (Figure 12).

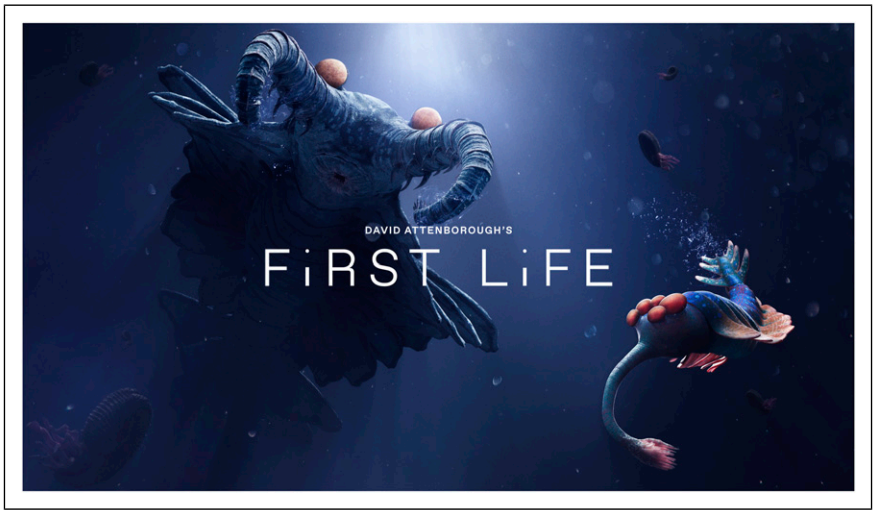

Figure 10. First Life VR project title poster.

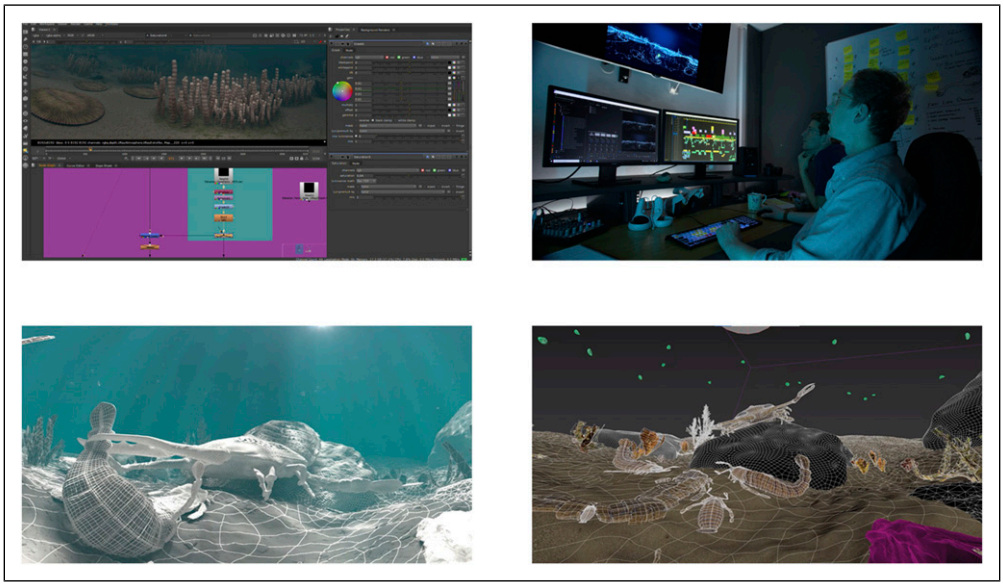

Figure 11. Video effects creation and rendering.

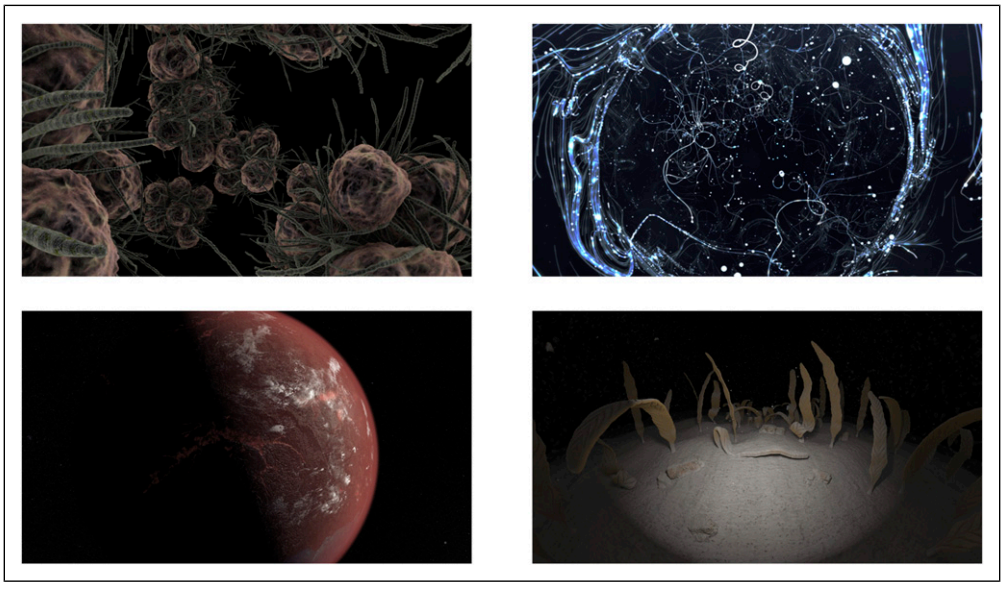

Figure 12. A selection of First Life stills.

To create these variations, some scenes were layered with multiple recordings, others were supplemented with additional monaural or stereo underwater and overwater recordings, and recordings of rain were low-pass filtered and spatialized. These methods proved to be effective when creating scenes with more spectral and temporal complexity, for example, the sense of horizontal depth could be varied, the implied density of terrain on the ocean bed controlled, and underwater currents were spatialized in multiple directions. Further, a multichannel convolution reverberation – Noise Makers Ambi Verb – was sometimes added to the ambience bus in Pro Tools to augment the sonic 'colour' or expand the apparent scale of the soundscapes (Huang et al., 2021). These techniques helped to create an ever-changing and evolving sequence of underwater scenes that represented a number of epochs and provided a compelling sonic backdrop for the featured organisms. The spectrally rich sounds of the water tended to obfuscate the emitters on the 3D SFX bus, and so in line the balance recommended by Lee and Lee (2017), high-mid frequencies between 1 and 5 kHz were boosted in order to enhance the directionality and cut through the thickness, and the reverse was done to the ambient water bus where a similar band was moderately attenuated. It is well understood that localization is an acute function of spectral profile (Schärer and Lindau, 2009), but mindful of this, subjective judgement indicated that in this case, it was not adversely compromised.

The hydrophone and contact microphone were both moved underwater and dragged across the seabed to emulate the seabed-dwelling creatures' movement. A range of sound effects was recorded at both shallow immersion and surface level to emulate movement with a variety of trajectories, intensities and speed variations. A further bank of sounds was created with objects such as shells, sand, stones (and even some foodstuffs) to build up a palette of textures that would add an expected familiarity to the creatures' interaction with their surroundings. As before, Sonnenschein's (2001) gestalt concept was always in mind, and as Chion (2015): (150) states 'the image is the conscious focus of attention, but the sound at every moment brings about a number of effects, sensations, and significations, that […] are credited to the framed image and appear to emanate naturally from the latter'. An evocative still is shown in Figure 13.

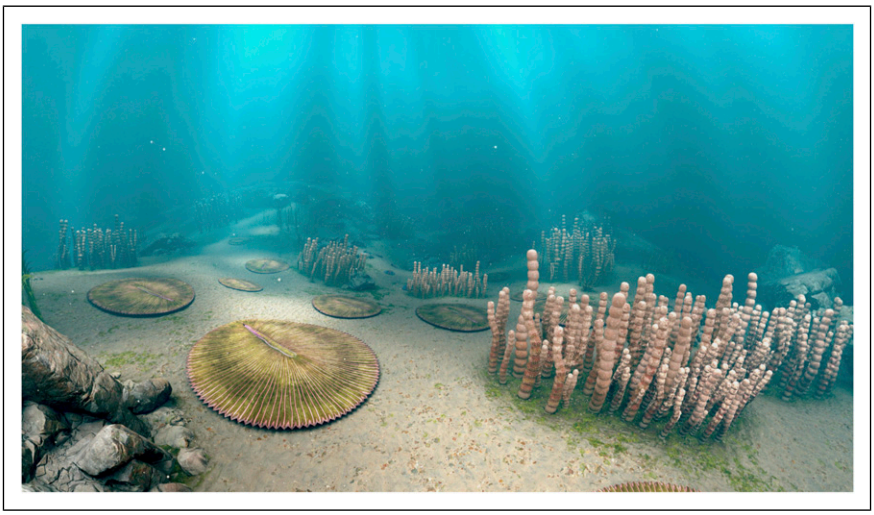

Figure 13. An underwater still from First Life.

Case study 4: Museum alive with David Attenborough – contextual analysis

The 2014 television documentary series, 'David Attenborough's Natural History Museum Alive' was originally commissioned by the BBC and produced by Atlantic Productions (Figure 14). In 2019, elements of the original series were recreated as a five-minute 6DoF experience in MR.[20] The project used spatial computing, plane detection and hand tracking, and superimposed scientifically accurate animations of extinct animals onto the user's natural field of view. Immersive-audio deployment required aspects such as the design of audio assets, audio implementation in a game engine and mix optimization for both linear and interactive-audio components of the experience.

Three prehistoric creatures were selected to provide contrasting user experiences. These were: the smilodon, a sabre-toothed tiger from the Pleistocene epoch; the dimorphodon, an early Jurassic flying animal of about 1 m in length; and the opabinia, a 5 cm, Cambrian-era sea creature. This selection entailed three different behaviour patterns and surrounding environments: running in the Savannah, flight in a dense jungle and swimming underwater. These each required characteristically different audio responses, and the implementation required BSR and elements of interactivity using approaches including Ambisonics, dynamic 3D sound-object spatialization, and near-field binaural rendering. The production pipeline for sound design and audio implementation was based around Pro Tools Ultimate, FB360 Spatial Workstation, Unity 3D, Lumin SDK, the Magic Leap One headset with its plugin, the Magic Soundfield Audio (MSA) for object-based panning (Magic Leap, 2020).

Figure 14. Museum Alive project title poster showing Magic Leap one HMDs.

Figure 15. Life-scale dimorphodon situated in the room 'as it appeared' to the HMD user.

Asset preparation and implementation. The sound-design process had a brief of making the creatures and their environments feel alive – which required complexity, depth and movement. To stay as faithful to the original documentary as possible, utilization of its existing content was preferred. However, as in the 'sound design and content acquisition' section of Case Study 1, since some of the original sounds had been grouped into stems, new additional sound content was designed for elements that required dynamic panning. Beyond a purely functional role, user-interface sound effects (UI-SFX) were highly stylized to act as an extension of the storytelling. The environmental sounds were mixed and then rendered as a single first-order Ambisonic (FOA)[21] stem in AmbiX format, using the FB360 plugin suite in Pro Tools; the 360° sound-field enhancing the perceived presence (Hermann and Ritter, 2004).

The original Sir David Attenborough narration and music score were reused after linear editing to fit into a new timeline and event sequence. Audio elements representing animals were spatialized using the MSA Unity plugin with acoustic modelling that matched the (impressionistic) reverberation characteristics of the creatures' habitats again, once again, to better exploit the kansei effect (Suzuki et al., 2012) and engage with the audience. Once all audio assets were mixed and rendered out, they were implemented in several different ways in Unity. The ambient audio tracks were binaurally decoded from FOA AmbiX using the Oculus Decoder script and set to play back with head-tracking. As previously described in the 'implementation and playback optimization' section of Case Study 1, non-diegetic narration and UI-SFX, music and hyper-realistic cinematic SFX were rendered as a stereo stem and set to play natively head-locked to aid cognitive focus (Vazquez-Alvarez and Brewster, 2009) and in line with the workflow recommendations of Susal et al. (2016).

Spatialization and optimization. 'Traditional' TV production might combine sounds of an animal's paws, body movement and its growling and breathing into a single 'object', but here, this would reduce impact and immersion since an essential aspect of the MR experience is the 'life-scale' representation of species – where the user can observe the animals from some distance, closely approach or walk around

(Figure 15). Such sounds were therefore assigned to appropriate diegetic emitters related to the smilodon's morphology and motion, as were diegetic UI-SFX that could interact with the hand gestures. This achieved superior synchresis whilst sounding more 'natural'. In discussions during development, it was again anecdotally observed by the team that the sounds appeared to be more externalized – which would appear to suggest that the room congruence explored by Werner et al. (2016) might be a more generalized phenomenon. The object-staging principle appeared to also apply in reverse; individual components could sound disjointed, and so the selection and treatment of individual sounds were heuristically tested in a 6DoF environment to subjectively identify and remedy any inconsistencies.

Combining both multiple background sound objects and general background sound with the principal sound emitters of the dimorphodon provided an important cognitive cue to the establishment of a convincing holistic scene (Roginska, 2017). This is particularly useful in the context of MR which often has limited visual information; here, the viewer can both see and hear an animal with a small area of nearby vegetation – but must rely on only hearing the surrounding habitat. The auditory scene therefore fills the missing visual gaps and helps a user to imagine what can't be seen (Hermann and Ritter, 2004), although balance must be sought between the diegetic and the environmental (Lee and Lee, 2017) – regardless of user position. Without dynamic 3D audio spatialization, the sounds would appear flat with limited perspective and suffer from a lack of envelopment; this audio modality enhances interface-transparency and hence immersion.

In the example of the dimorphodon's environment, the ambience layers included an atmosphere of a Colombian rainforest, various birds and insects in the trees – and other insects on the ground, as well as other dimorphodons in the distance. Associated reverberation staging (also see the 'implementation and playback optimization' section of Case Study 1) enabled the creation of a sonic world with multiple spatial perspectives to further aid immersion and the suspension of disbelief.[22] Birds did not exist until the later cretaceous era, although since today, we might typically associate such sounds with a forest, artistic storytelling licence permitted creative pitch shifting to create 'unfamiliar' bird calls and insect noises. The approach of staging sounds with higher frequency content above the listener allowed some exploitation of the pitch-height effect to maximize the perception of elevation (Roffler and Butler, 1968).

Several 'experience-design' considerations influenced the approach. Since the principal sound sources were originally recorded with close-mic techniques, 'track-layed' or designed from scratch (eg with a sampler), their spatial perspective is very different from if the entire scene had been recorded with a spatial microphone array and then encoded and reproduced as either an Ambisonic sound-field or some type of channel-based surround format that would have presented a naturalistic perspective. Accordingly, sounds mapped to emitters at different distances were gain-staged and filtered to create an illusion of the desired perceived depth and create compelling results to meet the greater storytelling objectives. As is common in computer-game sound, the MSA contains a set of automatically generated object-based distance properties that simulate the natural attenuation and energy absorption over distance according to the inverse-square law (Rumsey and McCormick, 2009: 18). Discrete sound sources interacted with the designated space's building component meshes along with the head-tracking and positional metadata. The combined information was fed into the audio engine for binaural rendering. The resultant audio was then played back either via the integral 'ear' speakers in the headset or headphones. In order to maintain the prevalent hyper-realistic approach, the object rendering could deliver heightened proximity by setting the spatialization to the MSA's near-field position.

Reverberation. In the Magic Leap device, spatial mapping is implemented by a feature that the company calls 'world understanding' (Magic Leap, n.d.) where both a room and objects within it are encoded as dense-mesh data. This creates a 3D perspective that allows objects in the room to implement video occlusion, but unfortunately, occlusion is not supported for audio (as noted by Goodwin (2019a)

regarding game-engine deployment), although it can determine the optimal reverberation in the MSA plugin, via the 'reverberation fingerprint' and 'per-object control and rendering of clustered reflections' described by Jot et al. (2021). The effective deployment of this feature was content-dependent – in some instances using only the MSA worked, but in others either pre-baking reverberation into audio assets or spatializing with third-party tools such as Google's 'Resonance Audio' were more effective. The interactive reverberation can be derived from two approaches; the convolution reverb engine acting on the audio engine's master bus or the environmental-geometry data being directly fed into the reverb engine. The latter method is more suitable for a larger scale experience, but to offer further enhancement requires sufficient data to be provided by the physics modelling of that environment. In any scenarios that depict lifeforms, reverberation functions more effectively when interactively applied to sounds dependent upon the relative positions of the multiple emitters and listeners simultaneously. Some pre-baked reverberation was applied to discrete audio objects but done so in conjunction with ambiences. This approach optimized the perception of creatures as existing *within* their environments and best ensured cohesion and congruence, although the technique is optimized for 3DoF (a la 360° video experiences) – not 6DoF (Susal et al., 2016). When setting this reverberation up – alongside 'naturalness' – precipitation of interaural differences was a consideration to maximize externalization (Leclère et al., 2019).

Source directivity. Using audio emitters that track the corresponding visuals can enhance realism (Ahrens et al., 2019; Zhou et al., 2020), both for the simulation of physical sound-source behaviour and also for interaction with them from the user's perspective, although Nykänen et al. (2013) point out that it gives lesser enhancement of binaural localization compared to reverberation. However, in practice, a virtual object that is automatically attenuated and filtered according to physical parameters such as distance and elevation – when the user is looking away – simply sounds more natural.

Attenuation and filtering were applied – mindful of Silzle's (2008) immersive-audio 'quality' – according to the physical positioning of objects within Unity, with the MSA facilitating an integrated workflow. It uses three subtended zones (inner angle, outer angle and interpolated transition in between) that define gain and frequency-dependent radiation relative to the 'forward' direction of the source. Minor adjustments were applied to the attenuation curves to mitigate the effect of sudden loudness and intensity changes from *user* movement. The movement-related gain-modulation lacked 'real-life' smoothness, but fine-tuning ensured that the user did not perceive any sudden changes that might break the immersion, presence and ultimately disengagement from the storytelling (Lee and Lee, 2017). A simple act of being able to both approach a large growling creature from in front or behind – and perceive its audio differently – and differently again with head rotation – can make a difference to the kansei verisimilitude prioritized by Suzuki et al. (2012).

Case study 5: Museum alive AR with David Attenborough – sound-design analysis

Museum Alive AR with David Attenborough was a further iteration of the original programme (see Case Study 1). It was released in 2021 – adapted to a new medium – augmented reality (AR) – see Figure 16. Furthermore, unlike the aforementioned case studies where the content was produced for specific HMDs, in this instance, the experience was designed to work on Apple iOS using the ARKit development platform, which includes LiDAR plane detection and visual inertial odometry to determine the position of the host device in a room. Despite following the same theme as the MR Museum Alive, the AR mobile version concentrated on educational aspects with a significant amount of background information about the featured species and natural history more broadly (see Figure 17). The entertainment factor maintained its prevalence since users could still see the realistic 3D representations of the creatures, but it also offered more developed and expanded environments.

Figure 16. Museum Alive AR project title poster.

Of course, when represented on a phone screen, users see a miniature version of those habitats, as opposed to the previous life-scale version. The scenes were more akin to small ecosystems that can appear to fit on the floor or a desk, as illustrated in Figures 18 and 19.

This shift in scale and rendering method subsequently introduced some issues for the audio. Since the sound design and audio implementation had already been created for the MR version, this adaptation required the development of a correct approach to preserve the levels of realism and immersion. It was apparent that any manipulation of the audio to reflect the smaller scale of the experience would be detrimental since distance and attenuation parameters in 3D panners offered insufficient resolution, and keeping the previous life-scale volumetric qualities of the audio translated only adequately. This is analogous to when large objects appear on a TV screen where their renditions are smaller than their actual size.

Largely relying on the previously produced assets, the head-tracked Ambisonic beds and emitters were rendered and implemented with BSR.[23] The music and VO remained in a head-locked mode. The level of music had to be considered differently and attenuated to avoid too much frequency masking from the music's phantom centre channel versus the centre-panned VO – this arrangement was unable to benefit from SRM (Dagan et al., 2019) as the head-tracked stem might. The reason for

Figure 17. Screenshots of the application.

this adjustment is that unlike HMD playback, a lot of mobile-app users are likely to use their device without headphones, and all elements had to come through with a degree of clarity during such playback. Since the introduction of the iPhone 7, iThings have offered stereo playback over speakers, so despite the very narrow sound-field produced by the device, a user could still appreciate some movement of the creature-emitters across their trajectories. However, if the user chose to wear headphones, they would receive much more detailed BSR of externalized diegetic components of the mix. BSR does not faithfully translate to speaker playback, but the induced artefacts are unlikely to be noticed given the fidelity of phone-type speakers.

If the audio was playing via Bluetooth speakers, then the issues might be more apparent, and in addition the AV collocation would be disrupted, but this was accepted as a shortcoming of the format. Although not a consideration at the time, with the advent of the head-tracking introduced in iOS 15 on accelerometer-equipped headphones,[24] users could better appreciate the 3DoF version of this app. However, Bluetooth latency is well beyond the TBW (Zhou et al., 2020) and video playback would need to compensate for this. Further, audio-visual spatial mismatches are likely with a handheld display, and this induces increased task demand (Pomper and Chait, 2017). Further, if a user does not hold the device directly in front of their face (although regardless of head orientation), divergence of aural and visual localization might be experienced (Maddox et al., 2014).

In summary, although the required suspension of disbelief might not be optimally maintained, an engaging and flexible user experience was still created.

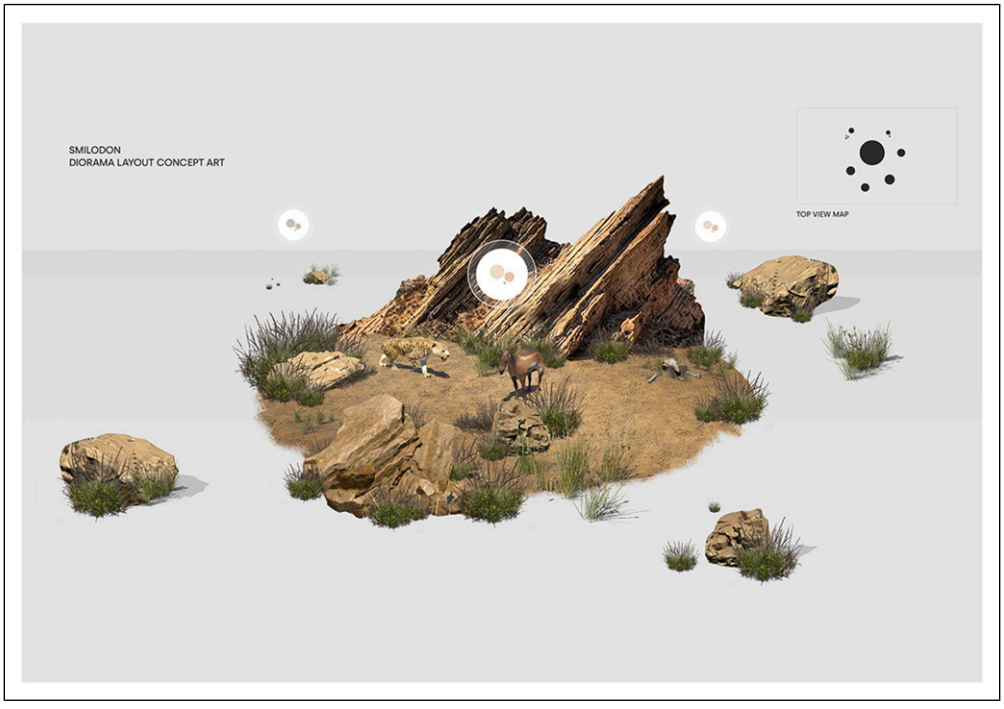

Figure 18. Render of a 'miniature world'.

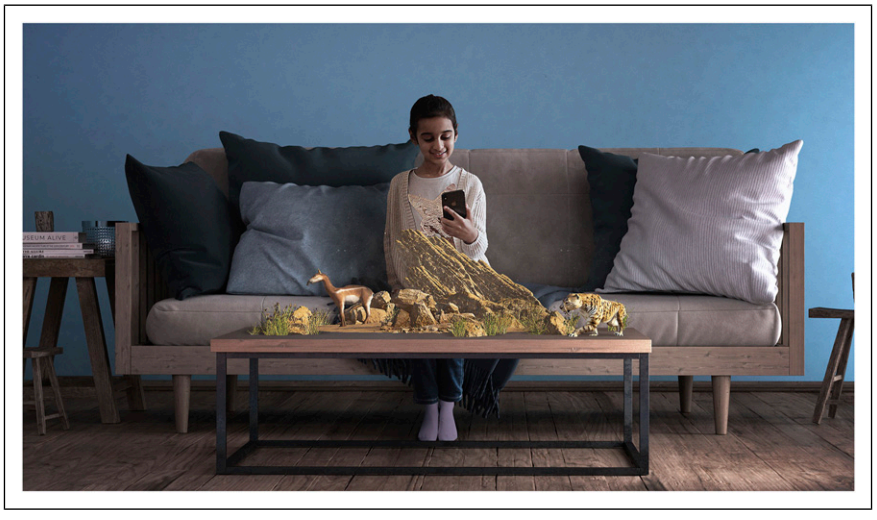

Figure 19. The miniature world presented on a tabletop illustrating how the user would see it on their phone.

Discussion

The case studies repeatedly refer to the core audio-authoring areas of technologies, tools, techniques and perception (3TP). These areas are interlinked and profoundly influence each other in an ecology that could be described by various sociological approaches, but that is beyond the scope of this text.[25] Instead, 3TP topics will be considered specifically with regard to their deployment in the above work and collated by area.

Technologies

Ambisonics is a very popular spatialization system largely because of the computational ease with which it can be decoded to head-tracked binaural playback. Binaural synthesis is required to stage emitters in 3D space, and optimal rendering is context-dependent and can be performed in real-time or in advance. Reverberation is a key component in generating convincing emitter externalization, and in MR, auralization of rooms is important for the sonic collocation of virtual objects within them; spatial mapping and plane detection are improving and will increasingly contribute to this. Extant (generic) audio creation/processing capabilities are also relevant.

Tools

To deploy these technologies in an authoring pipeline for XR immersive-audio experiences, the various software tools are highly specialized, and range from recording-studio and game-development staples (Pro Tools Ultimate, Unity) through to dedicated 3D audio spatializers (FB360). The delivery hardware – typically an HMD – is often idiosyncratic and might offer a unique feature set, but may also require bespoke software (eg MSA) to configure experiences. Different projects would quite possibly require further tools, although the pipeline is broadly similar (even for VR versus MR). Rapid obsolescence is an ongoing issue, yet new tools constantly arise with extended capabilities.[26] At the time of writing, there are many panners and manipulators available; a useful collation of free spatialization software is given by Baltic Immersive Audio Network (2021). There are of course 'attendant' tools such as microphones that contribute to the ecosystem. The application of all these tools is influenced by the associated visual modes, including film versus CGI and presentation in 3DoF or 6DoF.

Techniques

Conventional audio-asset creation/preparation and sound-design approaches yield the basic sonic palette, and postproduction techniques (eg atmos and spot effects) are a key backdrop, although these might be spatialized in 3D. Tying specific sounds to emitters is important in dynamic head-tracked conditions to ensure optimal multimodal localization, but the addition of a parallel audio stream to offer head-locked elements is advantageous. The sound stage can be used to draw the user's attention beyond their current gaze, extend the aural environment beyond the visual boundaries and add functional or emotive emphasis to visual objects. Expanding the audio-visual space with sonic emanation from beyond the boundary of the visuals has sometimes been seen as contentious, for instance in surround sound for traditional cinema. However, deployment of 360° audio with 180° video demonstrated experiential advantages by dissolving the boundary of the head-tracked audio field during head rotation. UI-SFX provide valuable interaction feedback and might be a function of either hand-controller-driven or hand-tracked engagement; these can

contribute beyond pure functionality contribute to 'game-play', immersion and entertainment. All sonic elements can be metaphorical, literal or hyper-real, but 'edutainment' experiences offer significant creative scope (beyond for instance, training applications); aesthetic and functional experience design must always be a consideration.

Perception

The important kansei qualities are all functions of perception. Audio localization and envelopment are key to immersion, and these are functions of numerous perceptual aspects, most notably HRTF compatibility when listening over headphones. Emitter reverberation should be a function of intended virtual distance, and in VR, acoustical cues should be aligned to the environment that they represent, and in MR/AR, the acoustics of the surrounding real environment needs to match those of virtual aspects. AV synchrony must be ensured and kept within the TBW. Presenting audio in 3D (perhaps particularly when externalized) appears to offer opportunities to improve dialogue clarity, cognition and information retention, and also to mitigate auditory fatigue.

Outline taxonomy of 3TP

An interactive graphical realization of the above interrelationships can be accessed in the 'supplemental materials' accompanying this online text (PowerPoint required). This gives the reader an opportunity to navigate between key audio-authoring areas to further explore and consider the ecosystem. For simplicity, only relationships to the target area are shown, and it should be understood that sometimes the target's 'satellites' also share interrelationships. A passive thumbnail of this taxonomy is shown in Figure 20.

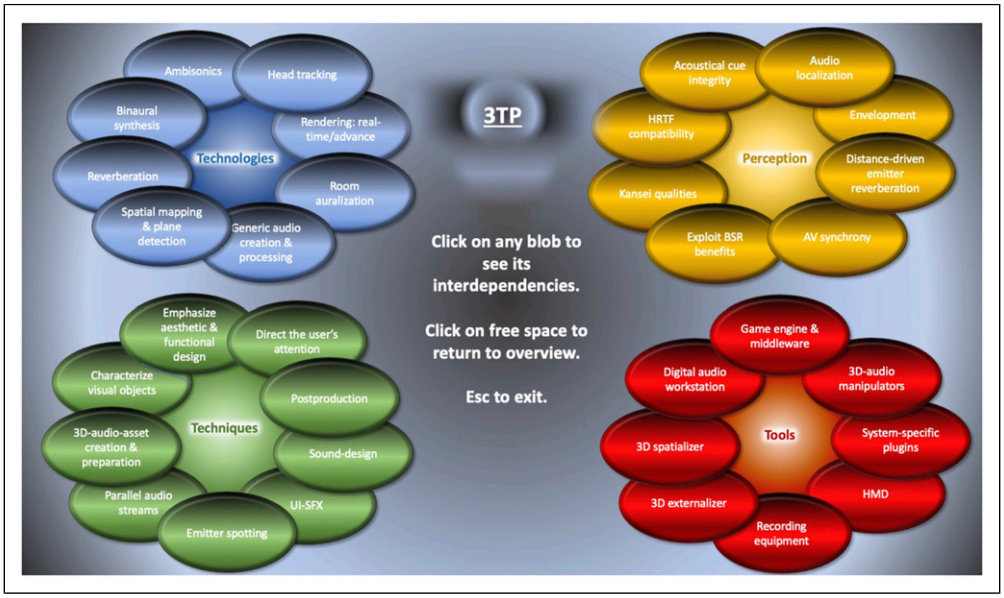

Figure 20. Thumbnail of the outline taxonomy of 3TP.

The future of audio in immersive storytelling

Thus far, a perspective of the state of the art around audio in immersive storytelling has been presented, and the reader might wonder where this trajectory will lead in the near future. Accordingly, some opinions (based upon industry movements) are now offered that consider how work to date might extrapolate.

"First there was mono, then stereo, and now there's Spatial Audio" (Lowe, 2021) – implying that some believe that immersive audio is likely to become the dominant mode of consumer listening.[27] Dolby Atmos® appears prevalent and is being deployed in digital audio workstations (for content creation), and increasingly, streaming services are offering spatial audio, even with head tracking. Apple are soon to enter the MR marketplace (MacRumors, 2022), and if they are true to form, this will be disruptive and new market standards will form (Coccia, 2018). So, it is quite possible that the complex and hybrid authoring solutions presented in the case studies herein will soon become replaced by an integrated and streamlined pipeline.

One current challenge in delivering immersive AV experiences is processing power. Whilst some HMDs are tethered to a computer to get over this, there have been other attempts to prioritize tether-free portability, but such systems can struggle. The Magic Leap/AT&T partnership demonstrates that the pervasion of 5G will come to offer sufficiently low latency and coverage to harness edge computing and lessen such bottlenecks (AT&T, 2018). Apple have also been said to be intending to use an iPhone to distribute the processing in future XR devices (Hayden, 2019), but more recent leaks have suggested native processing from a pair of M2 chips (MacRumors, 2022).

When technologies converge to facilitate the Metaverse, significant opportunities for the development of immersive storytelling will evolve (Sigaloff, 2022). With visual aspects being prioritized, photorealistic worlds will abound with multimodality likely drawing from the precedent of gaming. However, 'games typically only allocate 10% of their memory and processing resources to audio' (Goodwin, 2019a: 43), but the difference will be that server-based processing will still likely have overhead for substantial audio improvements, way beyond those currently available natively. This will facilitate more evolved real-time room modelling, including position/material-dependent reverberation resynthesis, coupled spaces and object (sonic) occlusion for more realistic acoustics in 6DoF scenarios. Real-time environmental sound synthesis is evolving rapidly (Moffat et al., 2019), and that will bring highly manipulable virtual sounds of the real world to augment the room models – computer-generated sound worlds will offer enormous narrative possibilities. Higher parallelism in audio-playback streams will increase presentation options.

Personalization is a thread that is evolving through many aspects of the digital world, and with ever-more effective noise-cancelling headphones with head-tracking, it will be possible to exist in the dynamic sound world of our choice, with or without associated visuals, and this might be geo-responsive (Zammit and Kenna, 2016). Of course, one current limitation of presenting 3D audio over headphones is the HRTF mismatch (Sunder, 2021). Apple are already moving towards LiDAR scanning of users' pinnae to offset this and personalize playback from iOS 16 (Lang, 2022), but perhaps the technology will evolve to match audiology pro-scanners used for made-to-measure hearing aids and earbuds to add ear canal morphology to make this even more efficient (Copithorne, 2019). Near-ear HMD speaker playback bypasses these issues to some degree, largely at expense of frequency response, but nanotechnology could yet offer beamforming at a comparable scale to mitigate this.[28]

Multimodality will also extend. Currently, immersive experiences are predicated largely upon AV, but haptics is rapidly evolving (Bermejo and Hui, 2021) and work is ongoing to digitally

Figure 21. Sir David Attenborough.

harness olfactory stimuli (Priscill and Anandhavalli, 2018). Eye tracking and facial-expression recognition are already viable control mechanisms, and this is likely to develop (MacRumors, 2022). Immersive storytelling will draw from all these developments – and more – to offer creative audio responses to the increments in 3TP.

Conclusion

The case studies illustrate that there is a well-defined audio-production pipeline for XR, but it has many device-specific idiosyncrasies and these can necessitate variations across projects. Projects with similar delivery media can consistently share a particular pipeline with variations primarily in the assets. The relationship between the various software actors can be complex and requires a diverse skill set. Spatial audio presents many opportunities for immersive storytelling, and these range through enhanced entertainment and immersion through to slight cognitive benefits.

The 3TP framework represents a flexible and adaptable model. The fact that it was derived from an incremental series of related projects gives it a tangible foundation and offers a degree of credence. Many of the subcategories have open-ended titles and can absorb areas not explicit in this text, and will also be able to accommodate internal evolution yet still hold relevance in terms of their interrelationships. Professional practitioners working in the field might typically take an approach driven by project goals and guided by tool-imposed workflow; however, the taxonomical aspect of the 3TP visualization referred to in the 'Discussion' section facilitates a strategic oversight of work packages that encapsulate a range of projects, and allows a vision of where to deploy or develop resources. It also serves as a pedagogical aid for the newcomer – to represent the ecology and its key dependencies and underpin an understanding of the immersive-audio workflow.

This outline taxonomy would benefit from more detailed classification of the interrelationships. This might entail both new layers in a hierarchy (eg further relationships within individual 3TP areas), and also the establishment of more detailed descriptors that define the relationships – with an emphasis on a generalized framework to ensure currency beyond these perhaps still inchoate case studies. This represents future work.

A further project is planned with a new Meta commission, 'David Attenborough's Conquest of the Skies'. This project will build upon the above case studies and will be cognizant of the 3TP framework with a view to retrospectively evaluating its relevance over the work packages of a fresh context. It will add some increments, for example by utilizing the equal-segment microphone array 3D (ESMA-3D) for enhanced spatial recording (Lee, 2019).

Audio for immersive storytelling in XR is nascent, yet it is here. The field will grow to acquire the immersive credence of a good book, and just as the tale within is boundless, tomorrow's sound world will follow (Figure 21).

Acknowledgements

All project images are courtesy of Alchemy Immersive and Atlantic Productions.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: One of the authors conducted paid work in the creation of the featured case studies. The other declares that there were no conflicts of interest.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Credits

Full credits for each individual project can be found at the following URLs:

- https://alchemyimmersive.com/productions/museum-alive-mr/
- https://alchemyimmersive.com/productions/micro-monsters-with-david-attenborough/
- https://alchemyimmersive.com/productions/kingdomofplants/
- https://alchemyimmersive.com/productions/david-attenboroughs-first-life-2/
- https://alchemyimmersive.com/productions/museumalive/

Awarding Organizations (wins and nominations)

Thanks go to the following organizations –

Museum Alive with David Attenborough:

- Raindance x 2 (2019)
- Micro Monsters with David Attenborough:
- 360VR Festival (2020)
- Emmy x 2 (2020)
- The Real World XR Awards (2020)
- Auggie Awards (2021)
- Cinequest VR (2021)
- Doc Edge (2021)
- VR Awards (2021)

Museum Alive AR With David Attenborough
• Apple Design Awards (2021)
David Attenborough's First Life
• International Sound Award (2022)
• Emmy (2022)/
• VR Awards (2022)
Kingdom of Plants
• Venice Film Festival (2022)
• Emmy (2022)
• VR Awards (2022)
Dissemination
Oculus TV viewers:
• Micro Monsters – 324,000
• First Life – 1,300,000
• Kingdom of Plants – 278,000

ORCID iD

Justin Paterson https://orcid.org/0000-0001-7822-319X

Supplemental Material

Supplemental material for this article is available online.

Notes

1. (Britannica, The Editors of Encyclopaedia Britannica, n.d.)
2. A term originally associated with Qualcomm, but now gaining general popularity to encompass virtual, augmented and mixed realities.
3. e.g., stereo, where an audio stream must be directed towards one of two speakers.
4. The US military offered an early definition of presence in virtual environments as being "the degree to which users are involved in and feel part of the […] environment" (Witmer and Singer, 1998). The efficacy of an XR system is directly related to the presence (Wiederhold and Wiederhold, 2005).
5. The interested reader might also refer to Szabó et al. (2016) for more contemporary computational approaches to solving these challenges.
6. Binaural synthesis is the mechanism by which headphone-audio can be staged in 3D – sounds can be placed behind and above the user. A useful review of this and associated psychoacoustic perceptual issues is given by Hacihabiboglu et al. (2017): "Binaural synthesis is based on the knowledge of the acoustic transfer paths between the source and the two ears. These paths are characterized by their impulse responses, referred to as the head-related impulse response (HRIR) and head-related transfer function (HRTF) in the frequency domain. For each source position, there will be two of them, one for the left ear and one for the right. When HRIRs are convolved with dry source signals, the resulting signals will incorporate the necessary binaural cues for the given source position."
7. "Authenticity, in this context, means that the subjects at the receiving end do not sense a difference between the actual auditory events and those which they would have had at the recording position when the recording was made." (Blauert, 1997: 373).
8. At the time of writing, first generation.
9. Although the term originated in science fiction (Stephenson, 1992), the Metaverse has more recently been defined as "a massively scaled and interoperable network of real-time rendered 3D virtual worlds that can

be experienced synchronously and persistently by an effectively unlimited number of users with an individual sense of presence, and with continuity of data, such as identity, history, entitlements, objects, communications, and payments" (Ball, 2022: 29). Despite appropriation of the term by certain parties, 'Metaverse' seems in sufficiently ubiquitous usage to adopt here. The interested reader is referred to Ball – he extrapolates this definition in great detail and offers many aspects of its significance.

10. Sometimes audio needs to be delivered as a single Ambisonic file whilst still containing non-spatial elements such as narration and music. This requires less-than-ideal workarounds e.g., embedding the voice into the Ambisonic B-format W-component (which is omnidirectional), or placing music into a pseudo-quad configuration within the AmbiX file. This prevents voice and music from being panned and attenuated in response to head movements.
11. These would typically have been recorded in stereo for the original TV series. Monaural sound effects were needed for repurposing into spatial audio for Quest. These effects had originally likely been monaural, but had been rendered into the atmos tracks in an original TV mix with a pre-ordained pan position and they could not be redeployed in isolation.
12. In this mode, each audio track (e.g., per emitter) requires 16 audio channels.
13. Akin to regular headphone playback.
14. By altering interaural level differences, a key localization cue.
15. Here, a Noise Makers Ambi Limiter HD.
16. A motif is a short readily identifiable piece of audio or music that typically repeats over a narrative in order to identify particular associated qualities or characterization when they require emphasis.
17. MIDI: Musical Instrument Digital Interface; a common device-agnostic music-performance/control instruction set.
18. In this setting, the term authenticity can (in preference to the earlier binaural usage) be taken at its literal meaning regarding the concept of maintain the mapping of biodata to MIDI without mitigation for aesthetic reasons.
19. The process of externalization involves a combination of binaural synthesis and the addition of early reflections and possibly more diffuse reverberation.
20. Also see Sec. 3.5 for an augmented reality (AR) version of this series.
21. At the time of production, there was no access to higher order tools.
22. A term first coined by Samuel Taylor Coleridge in 1871; the willingness of an audience to accept fantasy as reality.
23. In preference to the more usual dynamic Ambisonic sound-field rotation.
24. Airpods (third gen.), Airpods Pro and Airpods Pro Max.
25. The interested reader might consult Bijker et al. (1987), Gibson (2014) and Latour (1996, 2007) as a starting point. Zagorski-Thomas (2014) offers a useful synthesis of many of these ideas with regard to record production.
26. The landscape is fast-moving, and despite the relatively recent case studies above, at the time of writing, FB360 support was dropped in May 2022, Magic Leap are advertising the second-generation HMD, and the Meta Quest three is said to ship in early 2024 (*VR* Expert, 2022).
27. In the presence of so many current implementations, the word *format* is not yet applicable.
28. Seok et al. (2019) have developed comparable technology in the ultrasonic spectrum for neuroscientific applications in small animals

References

Ahrens A, Lund KD, Marschall M, et al. (2019) Sound source localization with varying amount of visual information in virtual reality. *PLOS ONE* 14(3): e0214603. Public Library of Science. DOI: 10.1371/journal.pone.0214603

Alchemy Immersive (2019) Museum alive with david attenborough. In: *Alchemy Immersive*. Available at: https://alchemyimmersive.com/productions/museum-alive-mr/(accessed 11 July 2022).

Alchemy Immersive (2020) Micro monsters with david attenborough. In: *Alchemy Immersive*. Available at: https://alchemyimmersive.com/productions/micro-monsters-with-david-attenborough/(accessed 11 July 2022).

Alchemy Immersive (2021a) David attenborough's first life VR. In: *Alchemy Immersive*. Available at: https://alchemyimmersive.com/productions/david-attenboroughs-first-life-2/(accessed 8 June 2022).

Alchemy Immersive (2021b) Museum alive ar with david attenborough. In: *Alchemy Immersive*. Available at: https://alchemyimmersive.com/productions/museumalive/(accessed 26 May 2022).

Alchemy Immersive (2022) Kingdom of plants with david attenborough. In: *Alchemy Immersive*. Available at: https://alchemyimmersive.com/productions/kingdomofplants/(accessed 8 June 2022).

Armstrong C and Kearney G (2021) Ambisonics Understood. In: JL Paterson and H Lee (eds), *3D Audio*. 1st edition. Perspectives On Music Production. London and New York: Routledge.

Atalay TB, Gül ZS, De Sena E, et al. (2022) IEEE/ACM transactions on audio, speech, and language processing 30. In: *Scattering Delay Network Simulator of Coupled Volume Acoustics*. IEEE, pp. 582–593.

AT&T (2018) AT&T partners with magic leap to create the future of 5g entertainment. Available at: https://about.att.com/story/2018/magic_leap_partnership.html (accessed 4 January 2023).

Audfray R and Jot J-M (2019) Reverberation loudness model for mixed-reality audio. In: *2019 Audio Engineering Society on Headphone Technology*. San Francisco, CA: Audio Engineering Society. Available at: http://www.aes.org/e-lib/browse.cfm?elib=20515.

Audfray R, Jot J-M, and Dicker S (2018) Audio application programming interface for mixed reality. In: *2018 Audio Engineering Society Convention*. New York City, NY: Audio Engineering Society. Available at: http://www.aes.org/e-lib/browse.cfm?elib=19741.

Avarese J (2017) *Post Sound Design: The Art and Craft of Audio Post Production for the Moving Image*. London, UK: Bloomsbury Academic.

Ball M (2022) *The Metaverse: And How it Will Revolutionize Everything*. New York, NY, USA: Liveright.

Baltic Immersive Audio Network (2021) 3D audio: Free tools roundup. Available at: https://www.balticimmersive.net/blog/3d-audio-free-tools-roundup (accessed 18 July 2022).

Barratt EL, Spence C, and Davis NJ (2017) Sensory determinants of the autonomous sensory meridian response (ASMR): Understanding the triggers. PeerJ Inc.. *PeerJ* 5: e3846. DOI: 10.7717/peerj.3846

Behringer R, Klinker G, and Mizell D (eds) (1999) *Augmented Reality: Placing Artificial Objects in Real Scenes*. 1st edition. Natick, Mass: A K Peters/CRC Press.

Burtt B (2014) *Ben Burtt Interview: The sound of lightsabers*. Available at: https://www.youtube.com/watch?v=TJQ3_tipGEY (accessed 14 December 2022).

Bermejo C and Hui P (2021) A survey on haptic technologies for mobile augmented reality. *ACM Computing Surveys* 54(9): 184–191. DOI: 10.1145/3465396.

Bijker WE, Hughes TP, and Pinch TJ (1987) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA, USA: MIT Press.

Blauert J (1997) *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press.

Blauert J and Braasch J (2020) *The Technology of Binaural Understanding*. New York, NY, USA: Springer International Publishing.

Bosman IdV, De Beer K, and Bothma TJD (2021) Creating pseudo-tactile feedback in virtual reality using shared crossmodal properties of audio and tactile feedback. *South African Computer Journal* 33(1): 1. DOI: 10.18489/sacj.v33i1.883

Bregman AS (1994) *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge MA, USA: MIT Press.

Brinkmann F, Lindau A, and Weinzierl S (2017) On the authenticity of individual dynamic binaural synthesis. *The Journal of the Acoustical Society of America* 142(4): 1784–1795. Acoustical Society of America. DOI: 10.1121/1.5005606

Brinkmann W-P, Hoekstra ARD, and van Egmond R (2016) The effect of 3D audio and other audio techniques on virtual reality experience. In: BK Wiederhold, G Riva, and MD Wiederhold (eds), *Annual Review of Cybertherapy and Telemedicine 2015*. Amsterdam, Berlin, Washington DC: IOS Press.

Britannica The editors of encyclopaedia britannica (n.d.) david attenborough. *Encyclopedia Britannica*. Available at: https://www.britannica.com/biography/David-Attenborough (accessed 7 July 2022).

Chion M (2015) *Sound: An Acoulogical Treatise*. Durham, NC, USA: Duke University Press.

Coccia M (2018) Disruptive firms and industrial change. *Journal of Economic and Social Thought* 4(4): 437–450.

Copithorne D (2019) Lantos launches next-generation 3d ear-canal scanner. Available at: https://www.hearingtracker.com/pro-news/lantos-launches-next-generation-3d-ear-canal-scanner (accessed 5 January 2023).

Dagan G, Shabtai NR, and Rafaely B (2019) Spatial release from masking for binaural reproduction of speech in noise with varying spherical harmonics order. *Applied Acoustics* 156: 258–261. DOI: 10.1016/j.apacoust.2019.07.015

Devereaux A (2020) The digital Wild West: On social entrepreneurship in extended reality. *Journal of Entrepreneurship and Public Policy* 10(2): 198–217 Emerald Publishing Limited. DOI: 10.1108/JEPP-03-2019-0018

EBU (2017) Digital Video Broadcasting, "Specification for the Use of Video and Audio Coding in Broadcasting Applications Based on the MPEG-2 Transport Stream (ETSI *TS* 101 154, v2.3.1. European Broadcasting Union. Available at: https://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.04.01_60/ts_101154v020401p.pdf (accessed 13 July 2022).

Elvemo J-M (2013) Spatial perception and diegesis in multi-channel surround cinema. *The New Soundtrack* 3(1). Edinburgh University Press: 31–44. DOI: 10.3366/sound.2013.0034

Expert VR (2022) Meta quest 3: Release date, features, and cost: Everything you need to know. Available at: https://vr-expert.com/meta-quest-3-release-date-features-and-cost-everything-you-need-to-know/(accessed 19 December 2022).

FilmSound.org (n.d.) Diegetic and non-diegetic sounds. Available at: https://filmsound.org/terminology/diegetic.htm (accessed 15 February 2023).

Fırat HB, Maffei L, and Masullo M (2022) 3D sound spatialization with game engines: The virtual acoustics performance of a game engine and a middleware for interactive audio design. *Virtual Reality* 26(2): 539–558. DOI: 10.1007/s10055-021-00589-0

Füg S, Marston D, and Norcross S (2016) The audio definition model—A flexible standardized representation for next generation audio content in broadcasting and beyond. *2016 Audio Engineering Society Convention*. Los Angeles, CA: Audio Engineering Society.

Genovese A, Zalles G, Reardon G et al (2018) Acoustic Perturbations in hrtfs measured on mixed reality headsets. In: 2018 Audio Engineering Society Conference on Audio for Virtual and Augmented Reality, Redmond, WA USA, August 2018. Audio Engineering Society. Available at: http://www.aes.org/e-lib/browse.cfm?elib=19699.

Gibson JJ (2014) *The Ecological Approach to Visual Perception*. New York: Psychology Press. DOI: 10.4324/9781315740218

Goodwin SN (2019a) *Beep to Boom: The Development of Advanced Runtime Sound Systems for Games and Extended Reality.* 1st edition. New York, NY: Routledge.

Goodwin SN (2019b) *Interactive audio geometry. 2019* Audio Engineering Society Conference on Immersive and Interactive Audio. March 2019. York, UK: Audio Engineering Society. Available at: http://www.aes.org/e-lib/browse.cfm?elib=20430.

Gupta R, He J, Ranjan R, et al. (2022) Augmented/mixed reality audio for hearables: Sensing, control, and rendering. *IEEE Signal Processing Magazine* 39(3): 63–89. DOI: 10.1109/MSP.2021.3110108

Gutierrez-Parera P and Lopez JJ (2016) Influence of the quality of consumer headphones in the perception of spatial audio. *Applied Sciences* 6(4): 117. Multidisciplinary Digital Publishing Institute. DOI: 10.3390/app6040117

Hacihabiboglu H, De Sena E, Cvetkovic Z, et al. (2017) Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. *IEEE Signal Processing Magazine* 34(3): 36–54. DOI: 10.1109/MSP.2017.2666081

Hall ET (1963) A system for the notation of proxemic behavior. [American Anthropological Association, Wiley]. *American Anthropologist* 65(5), pp: 1003–1026

Hayden S (2019) Report: Apple ar headset could rely on iphone for rendering and connectivity. In: *Road to VR*. Available at: https://www.roadtovr.com/report-apple-ar-headset-rely-iphone-rendering-connectivity/(accessed 4 January 2023).

Hendrix C and Barfield W (1996) The sense of presence within auditory virtual environments. *Presence: Teleoperators and Virtual Environments* 5(3): 290–301. DOI: 10.1162/pres.1996.5.3.290

Hermann T and Ritter H (2004) Sound and meaning in auditory data display. *Proceedings of the IEEE* 92(4): 730–741. DOI: 10.1109/JPROC.2004.825904

Huang Y-H, Venkatakrishnan R, Venkatakrishnan R, et al. (2021) Using audio reverberation to compensate distance compression in virtual reality. In: *ACM Symposium on Applied Perception 2021*, New York, NY, USA, 16 September 2021, pp. 1–10. SAP '21. Association for Computing Machinery. DOI: 10.1145/3474451.3476236

Ihlefeld A and Shinn-Cunningham B (2008) Spatial release from energetic and informational masking in a divided speech identification task. *The Journal of the Acoustical Society of America* 123(6): p. 4380–4392. Acoustical Society of America. DOI: 10.1121/1.2904825

Immersive Audio Podcast (2022) Sennheiser AMBEO mobility (Part 2). Number 65 (4' 21"). Available at: https://open.spotify.com/episode/0KtTihC2B5yNRpZkK6zkz4 (accessed 15 July 2022).

Jot J-M and Lee KS (2016) *Augmented reality headphone environment rendering.* In: 2016 AES International Conference on Audio for Virtual and Augmented Reality. 21 September 2016. Los Angeles, CA: Audio Engineering Society. Available at: https://www.aes.org/e-lib/online/browse.cfm?elib=18506 (accessed 5 July 2022).

Jot J-M, Audfray R, Hertensteiner M et al. (2021) Rendering spatial sound for interoperable experiences in the audio metaverse. *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. Bologna Italy: IEEE, 1–15.

Lang B (2022) Apple's iphone will soon scan your ear to solve a big problem with spatial audio. In: *Road to VR*. Available at: https://www.roadtovr.com/apple-iphone-custom-hrtf-ios-ear-scan-spatial-audio/(accessed 5 January 2023).

Latour B (1996) On actor-network theory: A few clarifications. *Soziale Welt* 47(4), pp. 369–381.Nomos Verlagsgesellschaft mbH

Latour B (2007) *Reassembling the Social: An Introduction to Actor-Network-Theory.* Oxford, UK: Oxford University Press.

Leclère T, Lavandier M, and Perrin F (2019) On the externalization of sound sources with headphones without reference to a real source. *The Journal of the Acoustical Society of America* Acoustical Society of America, 146(4), p. 2309–2320. DOI: 10.1121/1.5128325

Lee H (2019) Capturing 360 audio using an equal segment microphone array (ESMA). *Journal of the Audio Engineering Society* Audio Engineering Society, 67(1/2), pp: 13–26..

Lee H (2020) A conceptual model of immersive experience in extended reality. *PsyArXiv*. doi: 10.31234/osf.io/sefkh.

Lee H (2021) Psychoacoustics of height perception in 3D audio. In: JL Paterson and H Lee (eds), *3D Audio. 1 edition. Perspectives on Music Production*. London and New York: Routledge.

Lee S and Lee J-H (2017) Does spatial attribute between 2d and 3d virtual spaces make different user immersion of audio-visual events? In: *Proceedings of the 9th International Conference on Computer and Automation Engineering*, New York, NY, USA, 18 February 2017, pp. 56–59. ICCAE '17. Association for Computing Machinery. DOI: 10.1145/3057039.3057092

Lewis PJ (2011) Storytelling as research/research as storytelling. *Qualitative Inquiry* 17(6): 505–510. DOI: 10.1177/1077800411409883

Litovsky RY, Goupell MJ, Fay RR, et al. (2021) *Binaural Hearing: With 93 Illustrations*. New York City, NY, USA: Springer Nature.

Llewellyn G and Paterson JL (2021) Towards 6DOF: 3D Audio for virtual, augmented and mixed realities. In: JL Paterson and H Lee (eds), *3D Audio. 1 edition. Perspectives on Music Production*. London and New York: Routledge.

Lowe Z (2021) Apple music's zane lowe explains how spatial audio will transform music. Available at: https://www.apple.com/newsroom/2021/06/apple-musics-zane-lowe-explains-how-spatial-audio-will-transform-music/(accessed 4 January 2023).

MacRumors (2022) Apple Glasses. Available at: https://www.macrumors.com/roundup/apple-glasses/(accessed 4 January 2023).

Maddox RK, Pospisil DA, Stecker GC, et al (2014) Directing eye gaze enhances auditory spatial cue discrimination. *Current Biology: CB* 24(7): 748–752. DOI: 10.1016/j.cub.2014.02.021.

Leap M (2020) Soundfield Audio | Magic Leap. Available at: https://developer.magicleap.com/en-us/learn/guides/lumin-sdk-soundfield-audio (accessed 11 July 2022).

Magic Leap (n .d. Spatial mapping overview and detail options. Available at: https://resources.magicleap.com/en-us/privacy/spatial-mapping-overview-and-detail-options (accessed 26 May 2022).

Moffat D, Selfridge R, and Reiss JD (2019) Sound effect synthesis. In: M Filimowicz (ed), *Foundations in Sound Design for Interactive Media*. Abingdon, UK and New York, NY, USA: Routledge, pp. 274–299. DOI: 10.4324/9781315106342-13.

Noise Makers (n.d.) Brickwall ambisonic limiter. In: *Noise Makers*. Available at: https://www.noisemakers.fr/ambi-limiter-hd/ (accessed 19 December 2022)

Nordahl R and Nilsson NC (2014) The sound of being there: Presence and interactive audio in immersive virtual reality. In: *The Oxford Handbook of Interactive Audio*. Oxford Handbooks. Oxford, UK: Oxford University Press. DOI: 10.1093/oxfordhb/9780199797226.013.013

Nykänen A, Zedigh A, and Mohlin P (2013) Effects on localization performance from moving the sources in binaural reproductions. In: International Congress and Exposition on Noise Control Engineering: 2013, pp. 3193–3201. ÖAL Österreichischer Arbeitsring für LärmbekämpfungInstitute of Acoustics. Innsbruck, Austria. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-31692 (accessed 12 July 2022).

Oculus TV (2020) Bug planet: Insects rule in 'micro monsters. Available at: https://www.oculus.com/blog/bug-planet-insects-rule-in-micro-monsters-now-available-on-oculus-tv/ (accessed 27 May 2022).

Olivieri F, Peters N, and Sen D (2019) Scene-based audio and higher order ambisonics technology overview and workflows. *Technical Review 2019* Technical Review. EBU Technology and Innovation. Available at: https://tech.ebu.ch/publications (accessed 13 July 2022).

Paterson JL and Lee H (eds), (2021) *3D Audio*. 1st edition. Perspectives On Music Production. London and New York: Routledge..

Paterson JL and Llewellyn G (2019) Producing 3-D Audio. In: R Hepworth-Sawyer, J Hodgson, and M Marrington (eds), *Producing Music. 1 edition. Perspectives on Music Production*. New York, NY: Routledge.

Pomper U and Chait M (2017) The impact of visual gaze direction on auditory object tracking. *Scientific Reports* 7(1). Nature Publishing Group: 4640. DOI: 10.1038/s41598-017-04475-1

Priscill BJ and Anandhavalli M (2018) Digital smell technology. *International Journal of Emerging Technology in Computer Science and Electronics* 25(5): 451–454.

Puronas V (2014) Sonic hyperrealism: Illusions of a non-existent aural reality the new soundtrack. *Edinburgh University Press* 4(2): 181–194. DOI: 10.3366/sound.2014.0062

Roffler SK and Butler RA (1968) Localization of tonal stimuli in the vertical plane. *The Journal of the Acoustical Society of America* 43(6). Acoustical Society of America: 1260–1266. DOI: 10.1121/1.1910977

Roginska A (2017) Binaural audio through headphones. In: A Roginska and P Geluso (eds), *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. 1st edition. New York, London: Focal Press.

Roginska A and Geluso P (eds), (2017) *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. 1st edition. New York, London: Focal Press.

Rumsey F and McCormick T (2009) *Sound and Recording*. 6th edition. Burlington, MA, USA: Focal Press.

Rumsey F, Zieliński S, Kassier R, et al. (2005) On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. Acoustical Society of America. *The Journal of the Acoustical Society of America* 118(2), p: 968–976. DOI: 10.1121/1.1945368.

Schärer Z and Lindau A (2009) Evaluation of equalization methods for binaural signals. In: *2009 Audio Engineering Society Convention*. New York City, NY: Audio Engineering Society. Available at: https://www.aes.org/e-lib/browse.cfm?elib=14917 (accessed 15 July 2022).

Schissler C, Nicholls A, and Mehra R (2016) Efficient HRTF-based spatial audio for area and volumetric sources. *IEEE Transactions on Visualization and Computer Graphics* 22(4): 1356–1366. DOI: 10.1109/TVCG.2016.2518134

Schütze S and Irwin-Schütze A (2018) *New Realities in Audio: A Practical Guide for VR, AR, MR and 360 Video*. 1st edition. Boca Raton, FL, USA: CRC Press.

Seok C, Yamaner FY, Sahin M et al (2019) *A sub-millimeter lateral resolution ultrasonic beamforming system for brain stimulation in behaving animals*. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Berlin, Germany: IEEE, July 2019, pp. 6462–6465. DOI: 10.1109/EMBC.2019.8857627

Sigaloff P (2022) The storytelling makeover: Metaverse, immersive and the multitude of possibilities. Available at: https://www.thedrum.com/opinion/2022/01/19/the-storytelling-makeover-metaverse-immersive-and-the-multitude-possibilities (accessed 4 January 2023).

Silzle A (2008) *Generation of Quality Taxonomies for Auditory Virtual Environments by Means of Systematic Expert Survey. Düren*. Germany and Maastricht, Netherlands: Shaker Verlag.

Sonnenschein D (2001) *Sound Design: The Expressive Power of Music, Voice and Sound Effects in Cinema*. Studio City, CA: Michael Wiese Productions.

Stephenson N (1992) *Snow Crash. Re-issue edition*. London: Penguin.

Stocco D (2022) Diego Stocco. Available at: https://diegostocco.myportfolio.com/home (accessed 14 July 2022).

Sunder K (2021) Binaural audio technologies – an introduction. In: JL Paterson and H Lee (eds), *3D Audio. 1 edition. Perspectives on Music Production*. London and New York: Routledge.

Susal J, Krauss K, Tsingos N, et al. (2016) Immersive audio for VR. In: *2016 AES International Conference on Audio for Virtual and Augmented Reality*, Los Angeles, CA, USA, 21 September 2016. Audio Engineering Society. Available at: https://www.aes.org/e-lib/browse.cfm?elib=18512 (accessed 11 July 2022).

Suzuki Y, Okamoto T, Trevino J, et al (2012) 3D spatial sound systems compatible with human's active listening to realize rich high-level *kansei* information. *Interdisciplinary Information Sciences* 18(2): 71–82. DOI: 10.4036/iis.2012.71

Szabó BT, Denham SL, and Winkler I (2016) Computational models of auditory scene analysis: A review. In: *Frontiers in Neuroscience 10*. Available at: https://www.frontiersin.org/articles/10.3389/fnins.2016.00524 (accessed 12 December 2022).

Thery D, Boccara V, and Katz BFG (2019) Auralization uses in acoustical design: A survey study of acoustical consultants. Acoustical Society of America. *The Journal of the Acoustical Society of America* 145(6), p. 3446–3456. DOI: 10.1121/1.5110711

Vazquez-Alvarez Y and Brewster SA (2010) Designing spatial audio interfaces to support multiple audio streams. In: *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, New York, NY, USA, 7 September 2010, pp. 253–256. MobileHCI '10. Association for Computing Machinery. DOI: 10.1145/1851600.1851642

Vazquez-Alvarez Y and Brewster S (2009) Audio minimization: Applying 3D audio techniques to multi-stream audio interfaces. In: *Poster: 4th International Conference Haptic and Audio Interaction Design (HAID)*, Dresden, Germany, September *2009*. Springer International Publishing. Available at: https://tinyurl.com/2p97yap3.

Vorländer M, Pelzer S, and Wefers F (2013) Virtual room acoustics. In: R Bader (ed), *Sound - Perception - Performance. Current Research in Systematic Musicology*. Heidelberg: Springer International Publishing, pp. 219–242. DOI: 10.1007/978-3-319-00107-4_9

Werner S, Klein F, Mayenfels T et al (2016) *A summary on acoustic room divergence and its effect on externalization of auditory events*. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). Lisbon, Portugal: IEEE, June 2016, pp. 1–6. DOI: 10.1109/QoMEX.2016.7498973

Whittington W (2009) *Sound Design and Science Fiction*. Austin, TX, USA: University of Texas Press.

Wiederhold BK and Wiederhold MD (2005) *Virtual Reality Therapy for Anxiety Disorders: Advances in Evaluation and Treatment. Virtual Reality Therapy for Anxiety Disorders: Advances in Evaluation and Treatment*. Washington, DC, US: American Psychological Association. DOI: 10.1037/10858-000

Wiggins IM and Seeber BU (2011) Dynamic-range compression affects the lateral position of sounds. *The Journal of the Acoustical Society of America* Acoustical Society of America 130(6), p: 3939–3953. DOI: 10.1121/1.3652887

Witmer BG and Singer MJ (1998) Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7(3): 225–240. DOI: 10.1162/105474698565686

Yost WA, Dye RH, and Sheft S (1996) A simulated "cocktail party" with up to three sound sources. *Perception and Psychophysics* 58(7): 1026–1036. DOI: 10.3758/BF03206830

YouTube (n.d. Use spatial audio in 360-degree and VR videos. Available at: https://support.google.com/youtube/answer/6395969?hl=en-GB (accessed 13 July 2022).

Zagorski-Thomas S (2014) *The Musicology of Record Production*. Cambridge, UK: Cambridge University Press.

Zammit A and Kenna T (eds), (2016) *International Conference 'ICiTy. Enhancing Places Though Technology', Valletta, Malta. Culture and Territory*. Lisbon, Portugal: Edições Lusófonas.

Zhou HY, Cheung EFC, and Chan RCK (2020) Audiovisual temporal integration: Cognitive processing, neural mechanisms, developmental trajectory and potential interventions. *Neuropsychologia* 140: 107396. DOI: 10.1016/j.neuropsychologia.2020.107396

Zotter F and Frank M (2019) *Ambisonics : A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Cham: Springer Nature. Available at: https://library.oapen.org/handle/20.500.12657/23095 (accessed 12 July 2020).