Spatial Audio Production for Immersive Media Experiences: Perspectives on practice-led approaches to designing immersive audio content

Abstract

Sound design with the goal of immersion is not new, however, sound design for Immersive Media Experiences (IMEs) utilizing spatial audio can still be considered a relatively new area of practice with less well-defined methods requiring a new and still emerging set of skills and tools. There is, at present, a lack of formal literature around the challenges introduced by this relatively new content form and the tools used to create it, and how these may differ from audio production for traditional media. This article, through the use of semi-structured interviews and an online questionnaire, looks to explore what audio practitioners view as defining features of IMEs, the challenges in creating audio content for IMEs, and how current practices for traditional stereo productions are being adapted for use within 360 interactive soundfields. It also highlights potential direction for future research and technological development and the importance of practitioner involvement in research and development in ensuring future tools and technologies satisfy the current needs.

Introduction

The last decade has seen a significant increase in the availability of consumer-grade technologies to facilitate Extended Reality (XR) experiences, including Virtual Reality (VR), Mixed Reality (MR), and Augmented Reality (AR), classed more generally as immersive media experiences (IMEs). This has caused an associated increase in the production of *immersive audio*, a term often used synonymously with 3D spatial audio, that aims to deliver the user a sense of being present within the IME environment (Quackenbush and Herre, 2021). It can be used both as an accompaniment to 360° and Six Degrees of Freedom (6DOF) audiovisual experiences, and as the sole content in standalone audio experiences. IMEs contain an inherent interactivity not commonly found within traditional media, which adds an additional layer of complexity to the requisite sound scene. The amount of interactivity can vary greatly based on the content and type of experience.

Designing sound to be interactive and/or immerse is not new and is well established within video game sound design practice (McMahan, 2003; Zdanowicz and Bambrick, 2020) with many approaches being potentially suitable for adoption within non-video game IMEs. Sound design for traditional linear media also has well documented practices and workflows (Sonnenschein, 2001; Murray, 2019). The process of immersive sound design for IMEs, however, can be still considered as a relatively new area of practice with less well-defined methods, and requiring a new, and still emerging, set of skills and tools. Hence there is currently a lack of formal literature around the challenges introduced by this new type of content creation, the tools used to create it, and what those creating the experiences see as the defining features that differentiate it from traditional media. There is also a question of whether the technology and tools being developed align with the needs of these content creators.

This paper, through the use of semi-structured interviews and an online questionnaire, aims to understand how individuals working within the sound design industry have responded to this new form of content by addressing the following research questions:

1. What are key features of IME content?
2. What are the current challenges in creating IME content and how do these differ from traditional content production?
3. How are current practices for creating IME sound content within traditional stereo productions being adapted for use within content allowing 360 interactive soundfields?

Through these research questions this paper will explore how established techniques for creating immersion within traditional media are being translated for a 360 interactive soundfield. This paper then looks to place these within the context of current research into immersion. Finally, the paper explores how some of these challenges may be addressed through technological interventions and serves to provide pracetice-led direction for future research and technological developments.

Background

Experiences delivered via XR technologies, such as training simulators (Zucchi et al., 2020), multi-channel music mixes/soundscape recordings (BBC, 2021), 360 video, and video games utilising VR and/or spatial audio (Hood, Knapp and Griliopoulos, 2021), are often said to be *immersive experiences*. However, this term is often used vaguely and interchangeably with related terms such as realism, naturalness, involvement, and presence (Francombe, Brookes and Mason, 2017). The inconsistency within the terminology and the differing definitions can cause confusion, both for consumers, and for those undertaking research in the area (Agrawal et al., 2019). This can be further complicated when taking into account the multisensory nature of many IMEs.

Although there is, as yet, no standard definition of immersion, current literature supports the idea that immersion is a multifaceted concept. A recent study by Eaton and Lee (2019) identifies two overarching categories of immersion: *passive immersion* and *active immersion*. Passive immersion is defined as being related to a feeling of presence or being in an environment (Eaton and Lee, 2019), and encompasses previously defined notions of sensory immersion (Ermi and Mäyrä, 2005) and perceptual immersion (Biocca and Delaney, 1995). This type of immersion has no requirement for the user to play an active role in the narrative and may include experiences such as non-interactive VR, non-interactive music, and soundscape recordings utilising 360° video and/or audio.

Active immersion relates to immersive media with an interactive/task-based element (Eaton and Lee, 2019), such as in video games, where the user needs to make a choice or be constantly attentive due to the task at hand (Adams and Rollings, 2006). For example, a game that presents the user with choices that must be made in order to progress through the experience.

One other cause of immersion that is arguably important when discussing IMEs is that of the narrative presented to the user (which they may or may not play an active role in), which results

in a shift in their attention towards the story and away from the physical environment they are in (Thon, 2008). This is related to the concept of imaginative immersion (Ermi and Mäyrä, 2005) where users relate or feel emotionally invested in the characters and the events within the experience. This dimension of immersion, though defined within the context of video games, can also be associated with the immersion one experiences when reading a well-crafted novel or listening to a radio play.

Immersive content will often combine or aim to elicit several dimensions of immersion, for example the BBC's binaural version of the Doctor Who episode 'Knock Knock' (BBC R&D, 2020) combines aspects of perceptual, narrative, and emotional immersion. The binaural audio serves to elicit perceptual immersion by providing a spatial soundscene, and traditional techniques within storytelling aim to provide the narrative and emotional immersion i.e cliff hangers to create tension and anticipation, and development of likeable (and unlikeable) characters.

Contrary to those who regard immersion as a cognitive phenomenon there are those who regard immersion as being an intrinsic objective property of a system, so the more advanced the system is at replicating the relevant perceptual stimuli the more immersive it is considered to be. Slater (2003) argues that the term immersion should be reserved, "to stand simply for what the system delivers", and does not see immersion as a subjective experience, an idea that is rejected by others in favour of regarding immersion as a psychological/cognitive experience that can be caused by technological processes.

For the purposes of this research, it was deemed appropriate to use a definition that is broad in scope and that captures the multidimensional nature of immersion. The following definition of immersion, as defined by Agrawal et al. (2019), is therefore adopted:

> Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world.

With respect to potential technological innovation, it is important to identify the spatial audio technologies used to target the elicitation of perceptual and sensory immersion and, as such, focus on IMEs that utilise these and related XR technologies. Within the broader exploration of practice, it is also interesting to understand how creators of IMEs utilise and target the other dimensions of immersion. Using this definition allows both areas of interest within the scope of the investigation to be considered.

The rise in popularity of this content has resulted in companies such as the BBC, Facebook, and Google, releasing tools (Firth, Bailey and Pike, 2020; Facebook; Google) and producing content for IMEs. However, a question could be asked as to whether the current tools cater to the needs, and wants, of content creators within the wider sector?

Methods

This study was approved by the University of York's Physical Sciences Ethics Committee (ref: Turner190919). As it was desired to interview those with professional experience in producing

IME who would have insight into current industry practices, potential participants were identified through nonprobabilistic, purposive sampling techniques. Participants were therefore all required to be working professionally within the industry and have experience working on productions requiring immersive audio. Participants were recruited via targeted emails utilising contacts within BBC R&D's audio team and The University of York's AudioLab. All those approached were given information which briefly outlined the purpose of the research study and included the name and contact information of the first author. 20 people were approached to take part in the study. 5 of these were interviewed and 3 completed an online survey. Although the number of participants is low, the experience in industry for those interviewed ranged from 3.5 - 27 years with the median being 10 years. Of those who accessed the survey two had five or more years experience and one had between three to five years experience. This resulted in a small, but highly experienced, pool of participants from which to collect data.

Data were collected between the months of February to September 2020 using an online survey with multiple choice and open-ended questions, and semi-structured, open-ended interviews using the survey questions as a guide. The interviewer asked participants to expand on their answers where appropriate. Participants could choose whether to complete the survey or partake in an interview with the first author. Originally it had been planned for the interview to consist of a mixture of face-to-face and video call, depending on the location of the participants. However, due to the COVID-19 pandemic and the restrictions put in place, all interviews were moved to audio-only or video call. This presented some challenges in the form of audio quality and issues arising from internet reliability.

All interviews were recorded, and responses transcribed using the NVIVO qualitative analysis software (QSR International Pty Ltd, 2020), while the online survey was created and administered through Qualtrics (Qualtrics, 2020). Transcripts were reviewed by the first author for accuracy and revised where necessary. Inductive thematic analysis was performed based on the methodologies and procedures presented by (Braun and Clarke, 2006) and (Thomas, 2006) utilising NVIVO. A code book was created by the first author in NVIVO and Microsoft Excel where all identified codes could be added, and any new codes could be cross-checked against the rest of the data. Analysis followed five stages: initial reading of transcripts; identify text segments related to research objectives; identification and definition of themes/subthemes; reduce overlap and redundancy amongst themes/subthemes; and interpretation of themes in relation to original research questions.

Themes

Three broad themes, each containing several subthemes (shown in Table 1), were generated through the analysis of the coded data: The Virtual Environment, Production Practicalities, and End User Experience. These themes reflect the current literature concerned with what constitutes a state of immersion and the psychological and physiological factors that can elicit a state of immersion (Biocca and Delaney, 1995; Ermi and Mäyrä, 2005; Thon, 2008; Ryan, 2003), from the perspective of those creating the content and sit within the context of professional practise.

Themes	Subthemes
The XR Environment	Localization
	Capturing/simulating reality
	Multi-sensory
	Timbre
	Spatial aspects of experience
Production Practicalities	Availability of resources
	Automatic processing
	Sound quality
	Tool functionality
	Working with non-experts
End User Experience	Interactivity
	Cognition
	Levels of immersion
	Novelty

Table 1: Themes and subthemes generated from inductive thematic analysis of interview and survey data

The XR Environment

An area highlighted across all interviews as being an intrinsic, yet challenging, feature of IMEs, was the creation of the XR environment and the ability to have it replicate the sensory signals the user would experience were they physically in that environment.

There was a consensus that one part of simulating a real environment required the auditory scene dynamically reacting to a change in the user's orientation, as would happen in the real world:

> which again, is all down to trying to model reality and create something, whereas [...] if I'm turning around, instead of the sound field staying still, [...] the sound field moves with my head and it doesn't lock itself to your head. [Participant 1]

Participant 2 noted that what is considered desirable in IMEs might be at odds with what is generally desired within traditional media content, e.g., room ambience captured as part of a recording.

> In 360 you might actually really want something to sound like it's off mic because then it's capturing more of the room that the sound source is in and actually closer to what the real thing sounds like. [Participant 2]

The ability to emulate distance between the user and an object was associated with creating a sense of presence for the user by enabling the externalisation of the content. This was seen as particularly relevant when using headphone rendering, as this method lacks the natural distance between the user and source inherent in loudspeaker systems.

> ...distance to me is a big thing, externalisation, and this seems to me like they are [...] the two main things for me to create the sense of presence in space.... [Participant 4]

> externalisation is a big one, [...] if you're delivering via headphones [...]. 'cause having access to a lot of speakers in an array is much more tricky, so assuming it's headphone delivered, having things sound like they're outside of you and not inside your head, is the hallmark of good immersive audio, because in the real world, that's what sound sounds like. [Participant 2]

Participant 4 noted that when simulating distance within a synthesised environment, the processing required would be dependent on how the object had been recorded and edited, particularly with reference to its loudness. Standards for distance processing are noted as being difficult to establish.

> If an object is one meter away in a virtual world, it doesn't mean it actually sounds one meter away because it depends [...] how loud it is going into the spatialiser and how you edited it. [Participant 4]

For some, this has led to a perceptually driven approach to distance emulation using plausible approximations that make subjective sense to the the content creators, even if the parameter settings used are not objectively accurate.

> [...]you basically use your own approximations, arbitrary figures that initially make sense, but then you tweak it to trick your brain. OK, that sounds believable. OK, that works for me. Even if [...] the figures [on the screen] not correct. [Participant 3]

It was noted by some that the technology presently available for emulating distance is simply not yet of the desired standard.

> I think this is where everything is falling short, where we actually can say, oh, this guy is three meters away and I feel it. So distance modelling is quite a hard thing to do. [Participant 5]

Although importance is placed upon both being able to place objects accurately within a scene, and faithfully recreating the tonal characteristics of an environment, participants felt compromises must be made with respect to these features depending on the aim of the audio at that particular instance. This was because the tools being used to enable finer control of object placement often did so at the expense of introducing greater tonal coloration.

> two things that I'm always looking for, precision or timbre. [...] when I need localisation, maybe I give up a little bit on the, on the sound quality of it, I know that timbre might not be there. But when I want everything to sound really nice and smooth maybe I give up of a bit of the localisation. [Participant 4]

Non-spatial aspects of realism were also noted as being important within interactive media, with Participant 1 noting the perceived realism of characters. This may involve greater efforts to have their behaviour, such as in game dialogue, less repetitive by giving them a wider range of possible responses. This can result in many lines of dialogue across all in-game characters.

> something [...] that we bump up against and it (sic) indirectly to do with immersion in as much, I suppose it's more to do with kind of being believable and not repetitive, is

> editing and organising dialogue lines as we have more characters. They say more things to try and give the illusion that these are real characters. [Participant 1]

Though all the participants were audio professionals, with none undertaking professional visual production work, all expressed the importance of multi-sensory stimuli as a key feature in immersive media.

> It's [...] including all their attention in many senses as technically possible. [Participant 1]

It can also be the combination of visual and auditory signal processing which can together provide a greater sense of depth and distance to an environment. The importance of visual quality should also not be understated as it is a key aspect of many immersive experiences.

> [...] the video is stereo so you've got [...] a sense of depth of vision and having that additional stuff audio wise enhances what you see. [Participant 2]

Responses from all participants focused on the goal of creating an approximation of reality when creating IMEs. This raises the question of what it means or what are the requirements for an audio object to sound *real*. Although this, and related, terms were used by participants throughout the interviews, no definition was established as to what they considered constituted a real-world experience. It could be argued that everything is a real-world experience as, even within an IME, our ears and eyes are responding to physical stimuli. One interpretation could be that content creators do not wish it to be apparent that the scene/object is being generated by some form of loudspeaker, rather than the physical object which it aims to represent. This would then account for the desire to simulate the *real-world* timbre/frequency response of a sound and avoid any artefacts that could cause the user to focus on the device producing the sound rather than the sound itself. Absolute accuracy, however, appears to not be required as it was acknowledged that often a compromise is needed between accurate auditory localisation and the tonal accuracy of the associated environment.

Production Practicalities

There were many frustrations and challenges associated with the production of IMEs spanning all aspects of the production process. This usually centred around the view that current processes were lacking, hindering content creators in delivering experiences as easily as they might if IMEs were more commonplace and the tools and processes more developed.

Some participants noted the lack of available material in spatial formats (such as Ambisonic's B-format), which meant they often had to record their own material.

> There's not a lot of Ambisonic source material around. A lot of the stuff we've used, we've recorded ourselves. [Participant 1]

When unable to access spatial material for the specific environment they were looking to create, some resorted to layering stereo ambiences of the target location with a spatial ambience of a similar environment to help give the scene cohesion.

> If I've got the ambience in stereo from a London street, then I can put it in the background, some random street ambience [in Ambisonic format] just to fill up the space. [Participant 4]

While in some cases they resorted to just using spatialised stereo material.

> A lot of the time it's actually constructing stuff out of stereo and then spatialising it ourselves. [Participant 1]

Some participants commented that common methods for spatialising objects and rendering spatial soundscenes are still quite difficult to work with and can not always deliver the desired results, specifically when rendering over headphones.

> In fact, for probably all of them [VR users], it's going to be on headphones. So I feel like VR brought kind of binaural into focus and trying to get binaural sounding good, which is I think the big challenge. [Participant 1]

Working with non-experts also poses challenges. Clients commissioning IME content often lack the language to clearly articulate their feedback and may not have the skills to pinpoint what is causing any perceived issues.

> Clients are a challenge. They are able to say, I don't like this. I don't know what's happening, but, if it's wrong, it's wrong. [Participant 5]

There can also be conflicting assumptions regarding the aesthetic goals in IME production when collaborating with production teams accustomed to creating traditional content. Some concepts of sound quality may differ between collaborators, for example, the desired ratio of direct and reverberant sound on a dialogue track.

> a lot of people talk about things sounding off mic, as sort of bad sounding TV mixes. In 360 you might actually really want something to sound like it's off mic because then it's capturing more of the room that the sound source is in and actually closer to what the real thing sounds like. [Participant 2]

When dealing with immersive content that has both 360° video and spatial audio, placement of the microphone in relation to the camera is also of importance when maintaining the correct perspective between the visual and auditory material.

> I have been given audio recorded fairly close to a camera, but just in the wrong place, and it all sounds completely wrong. [Participant 2]

With these immersive experiences still not yet being widespread within the industry, and game audio workflows already having established platforms and tools, there can be hesitancy in adopting new technologies that require new practices and tools

> not everybody's completely sold on Ambisonics. So people are still quite attached to a world that they feel they've got control over. [Participant 1]

Reliability was also felt to be a contributing factor in the adoption of new technology proposed to assist with immersive workflows. As production timelines are often strict, tools need to work first time and complete the task quicker than the content creator would be able to do manually.

> if something is not reliable. You can't use it because the, the timelines of production are so tight. [Participant 5]

End User Experience

Thoughts on end user experience seemed to be predominately two-fold: Firstly, aspects of the user experience directly delivered by the content, such as the interactivity afforded to the user within the environment; and secondly, the psychological aspects that occur within the user's own cognitive processes, usually as a result of the technological processes drawn out in previous themes. However, as Participant 5 noted, immersion is not exclusive to content delivered via a specific medium and can in fact be achieved without any technological intervention.

> what we're meaning is that we get people losing themselves inside the experience and that can happen in any kind of medium, of books, especially [...] which are [...] non technological. [Participant 5]

In terms of what separates traditional content from immersive content, it is often considered that the user should have a level of participation within the environment and/or narrative, as opposed to merely being an outside spectator.

> the main differences for immersive experience: You're creating a world for the players to participate in. [Participant 1]

Allowing the user to participate in the narrative is often associated with allowing them to make choices that affect the direction or flow of the story, and this in turn gives them a sense of agency within the experience and causes the user to become invested in the story they are now helping to shape:

> it's basically anything, [...]which enhances the player's investment in the experience and their sense of agency in the experience. [Participant 1]

Alongside participation, aspects of the narrative, such as its ability to compel and engage, were seen to play an important part in a user's potential to become immersed in an experience and was cited as something that should be considered carefully during the production process.

> a key feature, (is the) story, telling a convincing story. [Participant 5]

When experiencing traditional media content, the user may not have a definitive position within the action. Camera angles change, and the sound scene is not always constructed to be a realistic representation of each object's location in relation to the camera location or viewing perspective. This is especially the case in audio reproduction formats that are horizontal only.

> A definite listening position is another big difference, with immersive audio where in traditional stereo or, I guess even in surround really, there isn't a definite, you're not in a very fixed position as the listener [...] as a viewer, you can see things and you might hear footsteps just there for effect, but they don't have to be rendered in such a way that is true to life. [Participant 2]

Not all immersive experiences require a first-person perspective, both first- and third-person perspective are common in video games with some allowing the user to dynamically change between the two. Some participants felt this created differing levels of immersion, depending on

the perspective from which the user was experiencing the world. The user still maintains a sense of agency and active participation in the narrative, but with a third person perspective they can be said to be taking control of a character within the world, rather than being the character, as would be the case in a first-person experience. This was viewed more as an interactive cinematic experience.

> I think it's [the video game] a kind of more cinematic experience. I think for we're creating a real world. But I think we were creating a real world in terms of a kind of movie that you can interact with. [...] because you can see the character on screen. So obviously you are not the character. So it doesn't have that level of immersion. But you can control the character. [Participant 1]

An experience being believable, as opposed to *real*, as noted in section 4.1, was also seen as an important part of being able to elicit a state of immersion from users. This is interesting because believable does not always have to correlate with creating something that is exactly true to life. There are certain situations where aspects of the experience need to be overstated to have the desired impact and compensate for the fact that the experience is not a complete sensory one.

> They could be a little bit hyper-real. In as much as sometimes you might want to slightly amp up the experience [...] the goal is still for people to believe that, you know, that gun that you're picking up and manipulating is a real gun, and it feels like a real gun. Sound does as much as it can to make that thing feel like a real gun. [Participant 1]

In the real world, each naturally occurring sound is often unique, with even repetitive sound events, such as gunfire, differing in small almost imperceptible ways. The lack of these minute differences, as highlighted by Participant 1, is something that a user could become sensitive to, resulting in the immersion becoming broken.

> one of the things which I think breaks the immersion for games...it's repetition, [...] in a game where you're trying to simulate reality, any kind of repetition people are very, very sensitive to. [Participant 1]

Participant 5 noted that another important factor, in addition to the techniques employed by the content creator, is the user's perception of the uniqueness of the experience. This is something arguably outside the control of content creators.

> [the experience] needs to have a certain standard in order to convince people that they are experiencing something unique and special. [Participant 5]

Participant 5 also commented on the user's preparedness for undertaking an IME. The process and effect of taking the time to prepare oneself for an immersive experience could be just as important as the techniques employed by the content creator to elicit the state of immersion.

> If you go to the cinema, you're not just going to the cinema. You're not just sitting in the cinema and watching the film. [...] Making the decision to go to the cinema, travelling for something that is important to you and then going inside and buy a drink and some popcorn and getting in the mood for this whole thing and to be prepared [...] we're going to take time for this and we're going to turn off our phones and everything. We're going to be fully there and there's nothing else that is distracting

> us... [...]how do we get to the experience in order to be prepared to let ourselves go. [Participant 5]

The theme of End User experience encompasses both the cognitive aspect of immersion and how the attributes of the user experience differ from that of traditional media. Participant responses under this theme, relating to the defining features of IMEs, often focused on aspects of the experience that could be associated with the concept of *involvement*. In the literature involvement is often framed as a psychological state necessary for cognitive immersion (Ermi and Mäyrä, 2005), but within the interviews the term was arguably used as a synonym for participation and/or agency. This was highlighted by participants making a point of describing how the users should be able to interact with the experience and participate or have agency within the narrative. This participation can result in the user entering a state of involvement as described in the literature. Even within the IMEs where users are more passive a state of affective involvement can occur which represents the emotions resulting from the design and aesthetics of the experience itself (Calleja, 2007).

The amount of agency a user would have varies greatly between experiences, as do the differing perspectives the user could take of any unfolding narrative. Participant examples demonstrating these varying combinations of user perspective and user participation included the user having a first person view of a musical concert but being passive as an audience member; having a third person view but being in control of a character; being able to interact and make decisions within the narrative, as is the case with many video games. Which of these examples is more immersive will be dependent on the individual and their situation and goes beyond just the nature of the experience. The idea of user perspective could also be interpreted as being related to the importance some place on users being given a defined position within an experience, that would be true to life were they physically present in the environment.

Discussion

The XR Environment, Production Practicalities, and User Experience themes emerge from the perspectives of those professional practitioners who are creating the content and happen to closely reflect the current literature concerned with what constitutes a state of immersion and the psychological and physiological factors that can elicit a state of immersion (Biocca and Delaney, 1995; Ermi and Mäyrä, 2005; Thon, 2008; Ryan, 2003). Though the participants were all audio practitioners, it is interesting that much of the interview data presents a holistic view of IME production, and while spatial audio production plays a key role in defining this new form of content, it is inextricably interconnected with other aspects of the experience such as user interaction, quality of narrative, and visual content. This section explores common features that emerged across the themes and how these go towards addressing the research questions outlined in section 1.

Distance Perception

There are various well-established methods for placing audio objects around the listener, however, placing sounds at a distance from the listener is a commonly expressed area of difficulty, and is therefore related to the second research question. When using headphone-based audio systems a prerequisite to creating auditory distance is the ability for the system to

externalise the sound so it is perceived as being located outside of the listener's head, and this was seen as a defining aspect of immersive content. If this prerequisite of externalization is not achieved, then it is very difficult to create a sense of auditory distance comparable to a real-world experience. Head movement tracking, another technology highlighted as being key to producing immersive content, has been shown to play a significant role in providing externalisation due to facilitating the simulation of dynamic spectral cue changes and can be effective even in the presence of degraded binaural information (Brimijoin, Boyd and Akeroyd, 2013).

Even with externalisation achieved it was still seen as a challenge to simulate objects at specified distances, and often participants relied more on their own subjective approximation of distance and less on whether the parameter values applied using auditory software reflected accurate values. A possible reason is that our understanding of the mechanisms involved in auditory distance perception are lacking when compared to azimuthal localisation, and the reliability of cues vary with stimuli, environment, and source distance. (Kolarik et al., 2016). This introduces an added complexity for content creators to deal with. Some of these cues, such as direct-to-reverberant energy ratio and the overall level of a source, are signal attributes commonly manipulated via software to imply an approximation of distance. Given the extra psychophysical complexity involved in auditory distance perception, and the degree to which these estimates vary in accuracy depending on the individual, stimuli, and the environment, it is maybe not a surprise that as yet, a standardised way to effectively simulate distance has not been found.

Multi-sensory aspects

The multi-sensory aspects of the experience were also deemed vital in achieving immersion and sit within the findings of our first research question. Alongside the quality and accuracy of the audio reproduction it was also felt that visual quality was an important factor, with techniques such as stereoscopic video helping to reinforce a sense of distance when combined with audio signal processing. The inclusion of multiple sensory stimuli better replicates what would be experienced in the real world, assuming no sensory impairments, further supporting the idea of perceptual/sensory immersion. It can also, given the well documented ability of our visual system to influence auditory perception, assist in achieving a greater quality of experience than current audio technology alone can deliver. This raises the possibility that it may be harder to achieve the same level of spatial plausibility with audio only content.

An important point to consider briefly is that, by their very nature, video games are designed to be immersive. Interactivity is a base requirement for all video games, but the ways in which video games have progressed in recent years, including the rise in popularity of spatial audio and the increased computational power of technology platforms that host them, means they are now often aiming to offer a multifaceted experience of immersion. Many non-video game IMEs model the interactivity found in video games through involvement in the narrative or affording the participant some degree of agency within the environment.

Immersion factors

The ability for a user to become fully immersed within an IME was intrinsically linked by participants to the quality of all aspects of the production, both technological (e.g., ability to replicate accurate sensory information) and non-technological (e.g., quality of narrative). It was

also said that the experience required a certain standard to convince the user they are experiencing something unique and special, although the exact implied meaning of this being vague. The idea of being required to present something that the user finds unique and special could suggest that the user's perception of novelty, and their prior experience with the medium, may have an impact on the level of immersion they experience. For users inexperienced in IME environments there may be a greater inclination to suspend disbelief and engage with the experience (McArthur, 2016), and this may cause them to be more likely to ignore/not notice quality issues that may be apparent to those more experienced. If this is the case, it raises the question of how long this "novelty effect" might last for, and once users become more accustomed to the experiences will it become increasingly difficult to elicit the same perceived quality of immersion?

Tools and assets

A lack of available or adequate resources and tools were seen as barriers to the adoption of immersive audio within the wider industry. Though multi-channel microphones are becoming more readily available, making in-house production easier, there is still a lack of sound effects libraries containing spatial 360 audio content when compared to mono and stereo content.

The adoption of new technology can often be a challenge as it requires experimentation and adaption in order to be refined, but often due to the tight production schedules and the inherent risk involved it can be difficult to undertake that experimentation outside of a research and development context. All participants referred to a commonplace requirement to capture bespoke spatial audio recordings as part of their work. This substantially increases the time taken to complete tasks, such as creating atmospheric audio beds for a scene, due to either the need to record a specific soundscape or create an artificial soundscape by layering existing mono/stereo material and then applying spatialization. This highlights areas of the traditional production workflow that are being adapted in order to be applied to IME production. In the context of video games, which have high levels of interactivity, it can make production vastly more complicated when trying to implement a format such as Ambisonics into workflows that have been built around channel-based audio. It was noted that practitioners are often much more comfortable using tried and tested methods given the intense time pressure involved in producing modern games. Those working within 360 video and VR seemed to approach the requirement to create a bespoke project based individual audio archive as part of the process when working in this area.

Potential for New Tools and Technologies

This section presents areas for future technological development related to spatial audio production for IMEs.

Automatic panning

Some of the challenges presented by immersive content production may be addressed by the further development of current production tools, and in some cases may require the development of new tools and technologies. Ensuring spatial congruence between visual and auditory objects has been highlighted as time consuming, especially when the objects' locations are not static within a scene. Some tools to automate this process have already been developed, (e.g., the object tracker within the Facebook 360 Spatialiser plug-in (Facebook)), however, responses from

participants suggest general issues with reliability. The utilisation of computer vision techniques explored by some of the authors previously (Turner, Pike and Murphy, 2020) could improve object tracking within a scene and provide object classification that may reduce the time taken to select appropriate sound effects from a chosen repository.

Distance emulation

The desire to simulate auditory distance is not new. The manipulation of digital audio signals to simulate the psychoacoustic cues for distance have well established methods within audio production and signal processing (Zölzer, 2011). While current methods may be enough to approximate the general perception of distance, the accurate simulation of distance with standardised techniques is something considered lacking in current tools, according to participants. There have been recent developments to enable estimation of sound source distance in known environments based on convolutional recurrent neural networks (Yiwere and Rhee, 2020). Since a person's ability to estimate distance becomes more accurate in situations where congruent visual and auditory stimuli are present (Anderson and Zahorik, 2014), it poses the question of whether applying both auditory and visual data would allow machine learning algorithms to develop a more complex representation of the problem space. If this is possible then they may have the potential to be used to inform cross-adaptive audio processing (Reiss and Brandtsegg, 2018) within an audio-visual context. An example might be a sound producing object within a visual scene having its distance estimated from associated visual information. Based on this prediction, parameters for EQ, reverberation, and level are then set according to features mapped from a prediction in the audio space corresponding to the distance estimated. This is similar to SoundNet (Aytar, Vondrick and Torralba, 2016) that utilised the natural relationship between sound and vision to learn acoustic representations from videos for the purpose of acoustic scene classification.

Upmixing

It may also be of benefit to further explore the possibilities of upmixing mono/stereo content to non-channel-based formats, such as B-format, as this would alleviate some of the issues surrounding the lack of spatially recorded sound libraries. Much of the current research into upmixing methods focuses on channel-based formats (Kraft and Zölzer, 2015; Park, Chun and Kim, 2016; Choi and Chang, 2021) and are based on decomposing the original signal into into its primary-ambient components and then applying processing specific to the target loudspeaker configuration (Walther and Faller, 2011). Laitinen (2014) proposes a method for converting two-channel stereo to B-format, but as an intermediary signal for the purposes of being reproduced using directional audio coding (DirAC). However, there are no formal perceptual testing results reported and no comment on whether the additional signal processing required during the conversion process would impact its use with other rendering methods. Some commercially available plug-ins offer upmixing to Ambisonics, such as Blue Ripple (2020) and Nugen (2020). No technical information is available on the latter, and Blue Ripple, held in high regard by all the participants interviewed, offer plug-ins to include a variety of channel-based formats into a 3D mix. However, at present, there appears to be no commercial offering to enable the synthesis of B-format signals from a stereo recording, that would be comparable to those generated had a multi-channel microphone been used to capture the original source material.

Limitations and future work

There are a number of limitations of the present study, which may impact the context and the interpretation of results, and the wider generalisability of the analysis. Effort was made to engage with as large a sample size as possible, although, the focus on those working professionally in immersive content production and able to respond with the study timeframe limited the pool of potential respondents. All participants, however, were highly experienced industry professionals and thus have detailed insight into the area this research looks to address. Therefore, while not providing a large enough dataset to allow wider generalisation it does offer a basis for further research.

The use of interviews and questionnaires to explore working practices have been criticised, with Luff et al. (2000) proposing that often there is a difference between what participants report they do and what they actually do. This difference is due to some activities being performed on such a regular basis that they become second nature and may not be at the forefront of a participant's thoughts when asked about the subject. Further work in the area may involve undertaking in-situ ethnographic style studies similar to that of Baume's (2015) exploration of radio production practice. This would enable the collection of firsthand data on current working practices and would not rely on participants having to take time out of their schedule to participate.

Conclusion

This paper, through the use of semi-structured interviews and an online questionnaire, explored how individuals within the sound design industry have responded to spatial audio production for IME content by addressing the research questions outlined in section 1. Thematic analysis was used to identify underlying patterns within the data and how these provide answers to these questions.

Immersive experiences aim to provide a user with a more intimate experience than traditional media, often placing them either within narrative or allowing them a truer to life perspective. Alongside the use of technologies utilised to create these experiences, it was felt that the difference in end user experience is what defined this type of content. Specifically, it enabled the user to feel present in the XR environment through the presentation of sensory stimuli comparable to that which would occur in a physical environment, with interactive content providing the user with a further sense of agency and involvement within the narrative.

Though there is sometimes a difference in semantics, clear associations can be drawn between what professional practitioners feel is important in generating immersion, and the different dimensions of immersion as explored in more academic literature.

Many of the challenges faced by immersive content producers are technological in nature: current audio tools are unable to replicate complex psychoacoustic phenomena such as distance, and those designed to assist in the spatialisation of audio associated with objects in a visual scene can be unreliable. However, with IMEs being new to many users there may be a novelty effect masking some of the current inadequacies of the technology as highlighted by participants. The question raised is how long such a potential novelty effect might be sustained and will immersive production tools and practices advance ahead of users' awareness of and desire for increased quality. There are also challenges associated with working with non-experts, both in the context

of clients commissioning IMEs and other practitioners that are new to the area. While this kind of challenge may fade as the medium becomes more established, education initiatives, like those available through the BBC Academy (2020), may not only help to alleviate this, but may also assist the speed at which the wider industry adopts this new form of content.

Spatial audio production for IME content might still be considered to be in its infancy, having only in the last decade started to come into its potential with the rise of affordable consumer level XR technology. This study has highlighted challenges for some of those working in the field and their view on what defines immersive content. The challenges highlighted could be seen as areas that warrant future research and interventions to further refine the tools available and increase the ease and quality of production in this fast-growing area. It is felt that while there is much research on-going this paper demonstrates the value in collaboration with professional practitioners in identifying directions for future research and tools/technology development that satisfy the current needs. It is also paramount that the user experience is at the forefront of any IME quality assessments due to the ability to both elicit, and subsequently experience immersion being so very dependent on the individual.

References

Adams, E. and Rollings, A. (2006). *Fundamentals of game design*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Agrawal, S. et al. (2019). Defining Immersion: Literature Review and Implications for Research on Immersive Audiovisual Experiences. *AES 147th Convention*, pp.1–11.

Anderson, P. W. and Zahorik, P. (2014). Auditory/visual distance estimation: Accuracy and variability. *Frontiers in Psychology*, 5 (SEP), pp.1–11. [Online]. Available at: doi:10.3389/fpsyg.2014.01097.

Aytar, Y., Vondrick, C. and Torralba, A. (2016). *SoundNet: Learning Sound Representations from Unlabeled Video*. (Nips). [Online]. Available at: http://arxiv.org/abs/1610.09001.

Baume, C., Plumbley, M. D. and Calic, J. (2015). Use of audio editors in radio production. *Journal of The Audio Engineering Society*.

BBC. (2021). *BBC Soundscapes for Wellbeing aims to bring nature to everyone*. [Online]. Available at: {https://www.bbc.co.uk/mediacentre/2021/soundscapes-for-wellbeing}.

BBC Academy. (2020). *Spatial audio: Where do i start?* [Online]. Available at: https://www.bbc.com/academy-guides/spatial-audio-where-do-i-start.

BBC R&D. (2020). *Sounding Special: Doctor Who in Binaural Sound*. https://www.bbc.co.uk/rd/blog/2017-05-doctor-who-in-binaural-sound.

Biocca, F. and Delaney, B. (1995). Immersive virtual reality technology. In: *Communication in the age of virtual reality*. Lawrence Erlbaum Associates, Inc.

Blue Ripple Sound. (2020). *O3A Upmixers*. https://www.blueripplesound.com/products/o3a-upmixers.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3 (2), pp.77–101. [Online]. Available at: doi:10.1191/1478088706qp063oa.

Brimijoin, W. O., Boyd, A. W. and Akeroyd, M. A. (2013). The contribution of head movement to the externalization and internalization of sounds. *PLoS ONE*, 8 (12), pp.1–12. [Online]. Available at: doi:10.1371/journal.pone.0083068.

Calleja, G. (2007). Revising Immersion: A conceptual Model for the Analysis of Digital Game Involvement. In: *3rd digital games research association international conference: Situated play*. 2007.

Choi, J. and Chang, J.-H. (2021). Exploiting Deep Neural Networks for Two-to-Five Channel Surround Decoder. *Journal of the Audio Engineering Society*, 68 (12), pp.938–949. [Online]. Available at: doi:10.17743/jaes.2020.0020.

Eaton, C. and Lee, H. (2019). Quantifying Factors of Auditory Immersion in Virtual Reality. In: *International conference on immersive and interaction audio*. 2019. York.

Ermi, L. and Mäyrä, F. (2005). *Fundamental components of the gameplay experience: Analysing immersion*.

Facebook. facebook360. *Facebook360*. [Online]. Available at: https://facebook360.fb.com/.

Firth, M., Bailey, R. and Pike, C. (2020). Binaural EBU ADM Renderer. *Google ARCore*, Google. [Online]. Available at: https://www.bbc.co.uk/rd/blog/2020-10-ear-next-generation-audio-software-tools.

Francombe, J., Brookes, T. and Mason, R. (2017). Evaluation of spatial audio reproduction methods (Part 1): Elicitation of perceptual differences. *AES: Journal of the Audio Engineering Society*, 65 (3), pp.198–211.

Google. *Google ARCore*. [Online]. Available at: https://developers.google.com/ar.

Hood, V., Knapp, M. and Griliopoulos, D. (2021). *Best VR games 2021: The top virtual reality games to play right now*. [Online]. Available at: {https://www.techradar.com/uk/best/the-best-vr-games}.

Kolarik, A. J. et al. (2016). Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, and Psychophysics*, 78 (2), pp.373–395. [Online]. Available at: doi:10.3758/s13414-015-1015-1.

Kraft, S. and Zölzer, U. (2015). Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain. *DAFx 2015 - Proceedings of the 18th International Conference on Digital Audio Effects*, pp.1–6.

Laitinen, M. V. (2014). Converting two-channel stereo signals to B-format for directional audio coding reproduction. *137th Audio Engineering Society Convention 2014*, pp.314–319.

Luff, P. J., Hindmarsh, J. and Heath, C. (2000). *Workplace studies: Recovering work practice and informing system design.* Cambridge, UK: Cambridge University Press.

McArthur, A. (2016). Disparity in horizontal correspondence of sound and source positioning: The impact on spatial presence for cinematic VR. *AES Conference on Audio for Virtual and Augmented Reality*.

McMahan, A. (2003). Immersion, engagement, and presence: A method for analyzing 3-d video games. In: Wolf, M. J. P. and Perron, B. (Eds). *The video game theory reader*. January 2003. 1st ed. Routledge. pp.67–86. [Online]. Available at: doi:10.4324/9780203700457-10.

Murray, L. (2019). *Sound design theory and practice: Working with sound*. 1st ed. Milton: Routledge.

NUDEN Audio. (2020). *Halo Upmix*. https://nugenaudio.com/haloupmix.

Park, S. Y., Chun, C. J. and Kim, H. K. (2016). Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks. *2016 International Conference on Information and Communication Technology Convergence, ICTC 2016*, (2015), pp.377–380. [Online]. Available at: doi:10.1109/ICTC.2016.7763500.

QSR International Pty Ltd. (2020). *NVivo (released in March 2020)*. [Online]. Available at: {http://www.s3a-spatialaudio.org/wp-content/uploads/2019/10/userdoc-0.12.0.pdf}.

Quackenbush, S. R. and Herre, J. (2021). MPEG standards for compressed representation of immersive audio. *Proceedings of the IEEE*, pp.1–12. [Online]. Available at: doi:10.1109/JPROC.2021.3075390.

Qualtrics. (2020). *Qualtrics (2020), Provo, UT, United States*. [Online]. Available at: {http://www.s3a-spatialaudio.org/wp-content/uploads/2019/10/userdoc-0.12.0.pdf}.

Reiss, J. D. and Brandtsegg, Ø. (2018). Applications of Cross-Adaptive Audio Effects: Automatic Mixing, Live Performance and Everything in Between. *Frontiers in Digital Humanities*, 5 (June). [Online]. Available at: doi:10.3389/fdigh.2018.00017.

Ryan, M. L. (2003). *Narrative as virtual reality: Immersion and interactivity in literature and electronic media*. Baltimore, MD, USA: The John Hopkinds Univerity Press.

Slater, M. (2003). *A note on presence terminology*. In: 2003.

Sonnenschein, D. (2001). *Sound design: The expressive power of music, voice, and sound effects in cinema*. Studio City, Calif.: Studio City, Calif.: Michael Wiese Productions.

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27 (2), pp.237–246.

Thon, A., J-N. (2008). Immersion revisited: On the value of a contested concept. In: *Extending experiences. Structure, analysis and design of computer game player experience*. Rovaniemi, Finland: Lapland University Press.

Turner, D., Pike, C. and Murphy, D. (2020). Content Matching for Sound Generating Objects within a Visual Scene Using a Computer Vision Approach. In: *Proc. Of the 148th AES convention*. 2020. Online. pp.1–10.

Walther, A. and Faller, C. (2011). Direct-ambient decomposition and upmix of surround signals. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.277–280. [Online]. Available at: doi:10.1109/ASPAA.2011.6082279.

Yiwere, M. and Rhee, E. J. (2020). Sound source distance estimation using deep learning: An image classification approach. *Sensors (Switzerland)*, 20 (1). [Online]. Available at: doi:10.3390/s20010172.

Zdanowicz, G. and Bambrick, S. (2020). *The game audio strategy guide: A practical course*. In: New York: Routledge, Taylor & Francis Group.

Zölzer, U. (2011). *DAFX: Digital audio effects*. Chichester: Wiley.

Zucchi, S. et al. (2020). Combining immersion and interaction in XR training with 360-degree video and 3D virtual objects. In: *2020 23rd international symposium on measurement and control in robotics (ISMCR)*. October 2020. IEEE. pp.1–5. [Online]. Available at: doi:10.1109/ISMCR51255.2020.9263732.

Acknowledgements

Thanks first and foremost must go to those who participated in the study. I am grateful for your valuable time and your willingness to assist me in my work. This research is supported by an EPRSC iCASE PhD Studentship in partnership with BBC R&D.