Exploring narrative possibilities of audio augmented reality with six degrees of freedom

Master's Thesis for Master of Arts (Art and Design)

Matias Harju
Sound in New Media
Master's Programme in New Media
Department of Media
Aalto University
2021

Aalto University
School of Arts, Design
and Architecture

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Master of Arts thesis abstract

Author Matias Harju

Title of thesis Exploring narrative possibilities of audio augmented reality with six degrees of freedom

Department Department of Media

Degree programme Master's Programme in New Media – Sound in New Media

Year 2021 **Number of pages** 90 **Language** English

Abstract

Audio augmented reality (AAR) is gaining momentum with the current renaissance of augmented reality (AR) and mixed reality (MR). Since the 1990s when the term AAR was first introduced, the medium has been researched extensively, and with new technologies and wearable devices, interesting AAR applications are getting available to the wider audiences. However, the narrative possibilities of AAR tend to be still an under-explored territory. This is especially true for AAR that utilises six-degrees-of-freedom (6DoF) positional tracking. In 6DoF AAR the user can freely move in a space while hearing spatially synchronised virtual sounds embedded in the environment. Conceivably, the medium has interesting storytelling potential in, for instance, museums.

This thesis explores the narrative possibilities of 6DoF AAR, concentrating on what is arguably characteristic of the medium: use of spatialised virtual audio, interplay between real and virtual, and interactivity based on user's location and movements. The topic is approached from a content creator's perspective through oscillation between conceptualisation and practical work.

By analysing literature and related AAR experiences, 1) a series of narrative techniques characteristic to 6DoF AAR has been identified, and 2) a prototype of a 6DoF AAR setup has been crafted. The design and creation process of the prototype has been discussed to better understand the possibilities and limitations the technology may set for the narrative use of the medium. Five demonstrative scenes have been created and used to test some of the identified techniques in practice.

This thesis presents a proposal for a list of narrative techniques characteristic to 6DoF AAR. The techniques are accounted for being a useful tool set for the author when designing the demonstrative scenes. Further, observations are disclosed on designing and building a 6DoF AAR setup capable of plausible auditory illusions and immersion. Challenges are reported related to, for example, registration errors in tracking and unexpected environmental sounds potentially hindering immersion and disrupting the narrative. For lack of user testing, however, more definite presumptions of the effectiveness of the techniques and the prototyped scenes cannot be made. On the other hand, it is suggested that with off-the-shelf components and authoring tools it is nowadays relatively easy for anyone with knowledge on sound design, programming, and storytelling to create gripping, spatially dynamic AAR experiences.

Keywords audio augmented reality, augmented reality, mixed reality, 6DoF, room-scale, immersion, narrative, interactivity, spatial audio, positional tracking

Aalto-yliopisto
Taiteiden ja suunnittelun korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Taiteen maisterin opinnäytteen tiivistelmä

Tekijä Matias Harju

Työn nimi Exploring narrative possibilities of audio augmented reality with six degrees of freedom

Laitos Median laitos

Koulutusohjelma Master's Programme in New Media – Sound in New Media

Vuosi 2021 **Sivumäärä** 90 **Kieli** englanti

Tiivistelmä

Lisätyn todellisuuden (augmented reality, AR) ja laajennetun todellisuuden (mixed reality, MR) nykyisen suosion myötä myös auditiivinen lisätty todellisuus (audio augmented reality, AAR) on nosteessa. Siitä lähtien kun käsite AAR esiteltiin 1990-luvulla, aihetta on tutkittu paljon, ja uusien teknologioiden ja kannettavien laitteiden myötä mielenkiintoisia AAR-sovelluksia on tullut laajan yleisön käyttöön. AAR:n kerronnalliset mahdollisuudet ovat kuitenkin jääneet vähemmälle huomiolle. Tämä pitää paikkansa erityisesti sellaisten AAR-kokemusten kohdalla, joissa käyttäjä voi liikkua vapaasti tilassa virtuaalisen maailman ja sen äänten ollessa spatiaalisesti synkronoituja ympäristöön (six-degrees-of-freedom, 6DoF). Tällaisilla sovelluksilla voidaan nähdä olevan mielenkiintoista narratiivista potentiaalia esimerkiksi museoissa.

Tämä opinnäyte tutkii 6DoF AAR:n kerronnallisia mahdollisuuksia keskittyen sen luonteenomaisiin piirteisiin: spatialisoidun (kolmiulotteisen) virtuaaliäänen käyttöön, todellisen ja virtuaalisen väliseen suhteeseen sekä interaktiivisuuteen, joka perustuu käyttäjän sijaintiin ja liikkeisiin. Aihetta lähestytään sisällöntuottajan näkökulmasta liikkuen käsitteellistämisen ja käytännön työn välillä.

Analysoimalla kirjallisuutta ja muita AAR-kokemuksia 1) sarja 6DoF AAR:lle luonteenomaisia, kerronnallisia tekniikoita on tunnistettu ja 2) prototyyppi 6DoF AAR -järjestelmästä on rakennettu. Prototyypin suunnittelu- ja rakentamisprosessia on analysoitu, jotta voidaan ymmärtää, mitä mahdollisuuksia ja haasteita teknologia tarjoaa taiteenlajin narrativiiselle käytölle. Viisi esimerkkikohtausta on luotu kerronnallisten tekniikoiden testaamiseksi käytännössä.

Opinnäyte esittelee ehdotuksen sarjasta kerronnallisia tekniikoita, jotka ovat luonteenomaisia 6DoF AAR:lle. Tekniikat ovat osoittautuneet hyödylliseksi työkalupakiksi tekijälle hänen suunnitellessaan esimerkkikohtauksia. Opinnäyte esittää myös havaintoja äänelliseen illuusioon ja immersioon kykenevän 6DoF AAR -kokemuksen rakentamisesta. Haasteista raportoidaan liittyen mm. järjestelmäviiveisiin, seurantaepätarkkuuksiin ja ennakoimattomiin ympäristöääniin, jotka saattavat heikentää immersiota ja häiritä tarinankerrontaa. Ilman käyttäjätestejä syvempiä oletuksia tekniikoiden ja esimerkkikohtausten vaikuttavuudesta ei kuitenkaan voida tehdä. Toisaalta opinnäyte ehdottaa, että valmiita komponentteja ja ohjelmistotyökaluja käyttäen nykyään on suhteellisen helppoa kenelle tahansa äänisuunnittelua, ohjelmointia ja tarinankerrontaa tuntevalle luoda mukaansatempaavia, tilallisesti dynaamisia AAR-kokemuksia.

Avainsanat audio augmented reality, lisätty todellisuus, laajennettu todellisuus, 6DoF, room-scale, immersio, kerronta, interaktiivisuus, tilallinen ääni, paikkaseuranta

ACKNOWLEDGEMENTS

I would like to thank my supervisor Antti Ikonen for his valuable perspective and support with this thesis, as well as the opportunities within new media sound I would have missed without him.

I am also grateful to my advisors; Professor Sebastian Schlecht for giving structure to thinking and to-the-point feedback during the last mile of the project, and Marko Tandefelt for suggesting many of the concepts tackled in this thesis and encourage to think big.

I want to address my special thanks to Emilia Lehtinen, a person with imagination without limits, for the invaluable storytelling ideas for 6DoF AAR. I am also grateful to Eva Havo for great book starts.

I would like to thank Matti Niinimäki for support at Aalto University with his inspirational hacker-artist attitude, and John Lee for his invaluable help with programming.

Thank you, Hannes Gamper and Steffen Armbruster, for sharing your expert insight.

Warm thanks go to the fellow students and other friends who have tested different prototype versions along the project.

Finally, a very special thank goes to Kaisa for arranging time for this crazy project while being wise as always, and encouraging in every turn.

LIST OF ACRONYMS

AAR	Audio Augmented Reality
AI	Artificial Intelligence
AR	Augmented Reality
ARA	Augmented Reality Audio
DAW	Digital Audio Workstation
DSP	Digital Signal Processor
GNSS	Global Navigation Satellite System
HMD	Head-Mounted Display
HRTF	Head-Related Transfer Function
IEM	In-Ear Monitoring
ILD	Interaural Level Difference
IMU	Inertial Measurement Unit
IPS	Indoor Positioning System
IR	Infrared
ITD	Interaural Time Difference
LOS	Line of Sight
MR	Mixed Reality
OSC	Open Sound Control
SBC	Single-Board Computer
SLAM	Simultaneous Localization and Mapping
UWB	Ultra-Wide Band
VAD	Virtual Auditory Display
VR	Virtual Reality
WFS	Wave Field Synthesis
2D	Two-dimensional
3D	Three-dimensional
6DoF	Six Degrees of Freedom

TABLE OF CONTENTS

1. INTRODUCTION

This thesis explores the narrative possibilities of audio augmented reality (AAR) with six-degrees-of-freedom (6DoF) positional tracking. In a 6DoF AAR experience the user can freely move in a real-world environment while hearing virtual audio content embedded in the environment, rendered binaurally through headphones. Using accurate location and head-orientation tracking the augmented sounds can be spatially synchronised with the environment. For instance, when the user walks forward and turns head right, the virtual world moves and rotates to the opposite directions, creating an appearance that the soundscape is attached to the real world. This 6DoF approach potentially creates an illusion of another acoustic reality coexisting with the real world. The tracking system can also be used as an input method for interaction with the virtual world: audio events and dynamic narration can react to user's location, head-orientation, and movements.

6DoF AAR carries many intriguing possibilities for storytelling and immersive experiences. For example, it can be used to convey an alternative narrative of a certain place through acousmatic sounds interplaying with the real world. The medium is also powerful in creating plausible illusions of something happening out of sight of the user, for instance, behind or inside an object. Unlike in traditional augmented reality (AR) with a visual display, in AAR the user's sight is not disrupted at all. This may be beneficial in places where situation awareness is important such as museums, shopping centres and other urban environments (Kurczak et al., 2011).

However, it seems that there are still not many 6DoF AAR experiences made with narrative content, and only little research has been done on the narrative use of AAR and particularly 6DoF AAR. Also, from the storyteller's perspective the threshold to start creating content for the medium might be rather high because there are no standardised technical solutions to choose from, and information on building an own setup must be gathered from different sources. Furthermore, the narrative language of the medium is still evolving, making the storyteller a pioneer in many ways.

This thesis will take part in developing the language by suggesting a list of narrative techniques that are characteristic of the medium. It will also discuss the design choices and challenges in assembling a 6DoF AAR setup as well as in creating narrative content for it. The work reflects the possibilities of creating auditory illusions with additive audio content: immersing the user

into a story world through her ears, without filtering anything out of the real world but rather binding the auditory narrative into it.

To help to understand the topic research and theories on some of the key concepts are discussed, including AR and AAR, virtual audio, interactivity, and narrative techniques. Also, four related experiences are analysed in terms of what means they use to convey their narrative ideas to the user and how the experiences are technically realised.

Many of the topics and issues discussed in this thesis may apply to other forms of AAR or auditory MR such as audio-only games and audio walks. There is also a lot of common ground with video games, virtual reality (VR), and other multimodal media. However, the focus of the thesis will be on the 6DoF AAR experiences, and the other media will be discussed only if necessary. In these cases, some alternate use of terms may occur, however being usually interchangeable; for instance, depending on the environment the user in the centre of the experience can be called player, reader, visitor, spectator, or audience.

1.1. Background

AAR has been an interest of research at least since the 1990s when the term 'audio augmented reality' was first introduced (Bederson, 1995; Arth et al., 2015, p. 6). It has been studied extensively from the perspectives of technology and psychoacoustics (e.g., Blauert, 1997; Härmä et al., 2003; Liski et al., 2016; (Jacuzzi, 2018), user perception and interaction (e.g., Sundareswaran et al., 2003; Veltman et al., 2004; Larsson et al., 2010; Rovithis et al., 2019), and potential applications (e.g., Mcmullen, 2014; Albrecht, 2016; Boletsis & Chasanidou, 2018). Some studies on the narrative use of AAR have been made (e.g., Gampe, 2009), however the scale being much smaller than with the other areas.

This is not to say that AAR has not been used to tell stories. One starting point for the narrative use of augmented audio can be set to early 1950s when the first handheld museum audio guide utilising shortwave radio was designed and taken into use in Amsterdam, Netherlands (Sandvik, 2011, p. 189). Since then, museums with their personal audio guides have probably been the biggest platform for storytelling using mediated audio vitally grounded in the surrounding environment.

Thanks to the geolocation capabilities of the modern smartphones there are nowadays many location-based audio applications available for the public. For instance, locative audio story and

audio guide apps have been popular for a long time and they are widely used with a plenty of content created around the world (e.g., *NoTours*, 2015; *Echoes*, 2020; *SonicMaps*, 2021; *Soundtrails*, 2021). The fitness game *Zombies, Run!*, launched in 2012, engages the user into auditory stories reacting to user's exercise pace (*Zombies, Run! Wiki*, 2021). Other, more experimental projects have also been initialised, examples being a location-aware musical album (Myers, 2011), and a 3D audio mapping project for the visually impaired (*Microsoft Soundscape*, 2018). However, geolocative AAR experiences with head-tracking and six-degrees-of-freedom are rare. Even though smartphones are capable of sensing orientation, using them to track user' head-orientation is inconvenient without an external sensor or special headwear.

To enable more immersive auditory experiences, several manufacturers of consumer audio products have developed headphones and other wearables with head-tracking capabilities (e.g., *OSSIC X*, 2016; *Bose Frames*, 2019; *Audeze Mobius Headphones*, 2021), and also other 'AAR' features such as binaural or 'spatial' sound as well as 'active listening' or 'transparency mode' which is a way of controlling how the environmental sounds are mixed with the programmatic audio content (e.g., *Sennheiser AMBEO Smart Headset*, 2019; *AirPods Max*, 2020). However, the development or production of many of the devices targeted for AAR use has been terminated soon after their launch, including the *Bose AR* programme (*Bose AR Public Beta Closure FAQ*, 2020). One can hope that the many promising applications and ideas developed for these devices (e.g., Gordon, 2019) will find a new life and users on other AAR platforms.

Although there have been some challenges in the customer market of AAR, from now and then artists and researchers have crafted their own spatial audio experiences. Some examples are *Growl Patrol*, an audio game by Queen's University in Ontario utilising head-tracking combined with smartphone geolocation (Kurczak et al., 2011), and *Sounds of Things,* an audio installation by Holger Förterer using accurate 6DoF tracking with infrared cameras (Förterer, 2013). Also, several companies are developing narrative AAR tools and environments for businesses and public instances. *Spatial* is one example, being a software toolkit enabling creation of interactive, virtual soundscapes for public spaces using installed loudspeakers (*Spatial*, 2021). Another example is *Usomo*, a 6DoF AAR system using trackable headphones to provide spatialised virtual audio content for multiple simultaneous users. Usomo has been used in multiple exhibitions and installations around Europe to create spatial audio experiences (*usomo*, 2019).

Whereas 6DoF AAR is taking its first steps towards a narrative art form, the storytelling mechanisms and techniques of theatre, literature and many other media are established and researched over the decades and even centuries. Since the coeval analysis of theatre and oral storytelling in ancient Greece, a vast number of narrative techniques have been identified (*Literary Devices and Terms*, 2020), many of which are universally applicable while some are specific for the particular media. Further, interactivity has brought an additional set of techniques at disposal and to be adapted to 6DoF AAR. Using these already established narrative means and, at the same time, developing its own unique techniques, the narrative possibilities of the medium are undoubtedly getting unveiled.

1.2. Scope

Within this work, I will focus on the narrative possibilities of 6DoF AAR. I see the medium as one of the most interesting branches of mixed reality (MR), enabling an interactive, subjective experience for each user while simultaneously allowing interaction based on other users' movements and actions.

To explore narrative possibilities on a more concrete level, I am interested in finding out what are some of the techniques that are characteristic to 6DoF AAR for enabling unique storytelling and auditory illusions. Hence, my first research question is:

> *What are characteristic narrative techniques of 6DoF AAR?*

Based on theories, related work, and my own experience, I have identified a set of narrative techniques potentially characteristic to 6DoF AAR. My main areas of interest in terms of techniques are:

1. use of spatialised (3D) virtual audio
2. interplay between real and virtual environments and objects
3. user interaction based on user's location, head-orientation, and movements of these two

While identifying the narrative techniques, I have designed and constructed a prototype of a 6DoF AAR setup. The prototype has served two purposes: Firstly, it has helped in understanding any possibilities and limitations the technology may set for the narrative use of the medium. The construction project has naturally influenced the process of identifying the narrative techniques, thus creating oscillation between the practice and conceptualisation.

Secondly, the prototype has worked as a platform to test the narrative techniques in practice. For that I have created five short demonstrative scenes based on some of the identified techniques. All in all, the prototype project has attempted to answer the second research question:

> *How to design and build a 6DoF AAR experience demonstrating some of the narrative possibilities of the medium?*

In this thesis, I will discuss and analyse the process of designing and assembling the prototype and the scenes. I will describe the demonstrative scenes with brief analysis on what functions the chosen narrative techniques have in them and what narrative ideas are explored through them. Finally, I will present my personal observations on the performance of the prototype, the usability of the chosen narrative techniques in this context, and the challenges I have faced with my technological and narrative choices.

2. THEORETICAL FRAMEWORK

This field of work potentially touches multiple research areas such as perception, cognition, hearing, spatial audio technologies, augmented reality, positional tracking technologies, immersion, illusion, and interactive storytelling. However, in the context of this thesis and its approach to the subject, I will focus on selected areas that may help in understanding the topic at hand.

I will first discuss augmented reality (AR), audio augmented reality (AAR), and mixed reality (MR). Then I will discuss the key concepts and components of an AAR setup, including spatial audio and headphones, positional tracking, and interactivity in narrative environments. Finally, I will analyse some of the previous experiences related to 6DoF AAR by other authors.

2.1. Augmented reality

According to Azuma (1997, p. 356), AR 'allows the user to see the real world, with virtual objects superimposed upon or composited with the real world.' Whereas VR is trying to immerse the user completely inside a virtual environment, 'AR supplements the reality, rather than completely replacing it'. Azuma talks about image-based AR, almost totally neglecting auditory and other sensory modalities, hence his use of word 'see' instead of 'sense' or 'experience'. However, he does not explicitly exclude audio from being a part of AR in multi-modal applications, or audio as the only augmented channel.

Azuma describes AR as any system that has the following three characteristics:

1. Combines real and virtual
2. Is interactive in real-time
3. Is registered in three dimensions (Azuma, 1997, p. 356)

'Real' means the surrounding world the user can multimodally sense (Schraffenberger & van der Heide, 2016) with, for example, sight, hearing, smell, touch, and perhaps taste. In addition to the traditional five senses, proprioceptive senses may help to understand the surroundings due to signalling e.g., body position and movements (Proske & Gandevia, 2012). This sensation of the real forms the basis of AR, and the virtual content is in relation to the real, embedded to it in a way or another. As Jacuzzi (2018, p. 1) puts it, Augmented Reality is 'grounded in reality'.

Azuma's (1997) definition on AR works as a useful starting point to understand AAR. Still, auditory and visual modalities differ so much from each other that in many respects AAR could

be considered as a medium of its own. Whereas human's sight is able to register only a limited view at a time, hearing is omnidirectional and can thus pick up sounds from any direction regardless of the head orientation. You can look away, but you cannot 'listen away'. Also, while a visual image can be viewed for as long as it is visible, sounds are temporal or transient in nature: once something is heard it is not possible to re-hear it. Further, hearing is always active, even when sleeping, unlike sight that can be 'switched off' by eyelids. (Sarter, 2006, p. 441) Also, unlike light, sound can normally travel around obstacles (Kolarik et al., 2016, p. 373), and whereas walls prevent from seeing to the next room, sounds can often be heard through.

2.1. Audio augmented reality

Audio augmented reality (AAR) is a relatively new concept, having its roots in the virtual reality boom of the 1990s. The term was first used by Benjamin Bederson in 1995 (Arth et al., 2015, p. 6) when he presented a prototype of an automated tour guide for museums playing voice information based on the user's location (Bederson, 1995). The idea of augmented reality (AR)—superimposing virtual objects on the real world—had been described a few years earlier by Myron Krueger in 1991. For Bederson, one of the advantages of AR was clearly the fact that it allowed enriching museum and other experiences without isolating people from each other like used to be the case with Virtual Reality (VR) at least at the time. Hence, a personal location-based audio tour guide was a logical concept in that respect: it allowed visitors to hear voice guidance when looking at the exhibition pieces, and by simply stepping away from the items, enabled people to interact socially and talk with each other. (Bederson, 1995, p. 210)

Since that, the term audio augmented reality and its abbreviation AAR have steadily been established in academic literature and commercial applications, although the term 'augmented reality audio' (ARA) is sometimes used, too (e.g., Rovithis et al., 2019). AAR can refer to a number of different concepts (Krzyzaniak, Frohlich and Jackson, 2019, p. 1). Gamper's (2014, p. 23) definition seems to represent a rather common way to understand what AAR is: 'a technology that aims to embed virtual auditory content into the real environment of a user.'

However, as Krzyzaniak et al. (2019, p. 1–2) note, not many attempts to define AAR have been conducted. In consequence, they present a suggestion for a user-centred taxonomy of different types of AAR. They are not trying to define AAR, but rather list concepts that have been, or could be, referred to as AAR. Their working assumption has been that the 'essential elements of AAR are (1) access to some local physical analogue sound, direct or indirect, and (2) some digitally

mediated addition or modification to the sound that provides some benefit.' With these assumptions their taxonomy embraces a multitude of concepts of which a few examples being telephone, enchanted football, karaoke, electric guitar pedal, geolocated sound and 'sound attached to virtual objects in AR games'. The list is intentionally permissive, including non-interactive and non-immersive concepts to it. (Krzyzaniak et al., 2019, p. 2) However, it shows how common a phenomenon it is to combine 'real' sound with 'virtual' sound, despite of whether they are called AAR or something else.

For a more concrete grasp on AAR we can analyse Azuma's (1997, p. 356) three characteristics of AR from the perspective of audio.

1. Combines real and virtual

Regardless of the augmented sensory channel(s), the 'real' refers to the multimodally experienced environment, centre of which being the user. The 'virtual' may, however be a more flexible concept. Although Azuma (1997) himself tried to avoid limiting his definition to any specific technologies, virtual nowadays commonly refers to computer-generated simulations. For image-based AR, the most convenient way to create images that could be manipulated three-dimensionally in real-time is to model and run them using computer software. However, with other modalities the situation may be different, like with audio: sound sources can be created and embedded in the environment without the use of a computer. In Bederson's (1995) audio tour the 'virtual' component was the museum guide's voice describing the exhibition items to the visitor through headphones, played back from a modified Sony MiniDisc player. In modern AAR experiences with spatialised (3D) sounds the virtual audio material is usually still recordings of real sounds. Hence, we can probably safely extend the definition of virtual to be, e.g., 'digitally mediated' as Krzyzaniak et al. (2019, p. 2) have done.

Having said that, one can imagine an AAR experience without any digital mediation by, for example, using hidden speakers in a real-world environment, controlled by an analogue interactive system reacting to user's movements or other actions. Maybe there could even be a human being controlling the experience. With this in mind, shackling ourselves to tight definitions may not be very fruitful, although the definitions are useful in understanding the phenomenon and finding new angles to it.

2. Is interactive in real-time

This characteristic is essential as the aim is to try to augment life and create a subjective experience where the environment reacts to user's actions. Many traditional museum audio tours would not fulfil this characteristic since their interaction is often sequential instead of real-time: a new audio track starts to play when the museum visitor manually selects the right track on her device. However, location-based tours and audio-walks, such as Bederson's (1995) prototype and many after that, can be seen being interactive in real-time as they track the user's position and start and stop the otherwise sequential audio tracks accordingly, sometimes fading the material in and out, or manipulating it otherwise. The same applies to 6DoF AAR with location and head-orientation tracking, thus rotating and moving the augmented audio world in synchronisation with the real environment.

3. Is registered in three dimensions

The third characteristic of AR requires that virtual objects were combined with the real environment in 3D. Hence, a two-dimensional visual overlay projected in the field of user's sight would not be considered as AR, even if it was interactive in real-time (Azuma 1995, p. 356). Applying Azuma's definition to AAR, a museum audio tour with guide's voice playing 'inside' the visitor's head, or a geolocated audio walk with interactive but head-locked soundscapes would not qualify since the sounds are not combined with the real 3D world. However, one can argue that the voice and soundscapes would still be in *relation* to the three-dimensional environment (or often two-dimensional, omitting the vertical axis): they are attached to a certain position or area in the real 3D (2D) space even if they are not acoustically rendered in 3D.

It is left to experimenting to what extent three-dimensional rendering and spatial synchronisation is required in order to gain the effect of the objects coexisting with the real world. For example, in the *Sounds of Silence* exhibition (2018–2019) in Bern Museum of Communication, a visitor heard head-locked soundscapes through headphones, triggered according to her location in the space. Even without any other 3D synchronisation, the locative system created an illusion of the sounds belonging to the real-world spaces and thus coexisting with the reality. Should the visual and other sensory stimuli from the environment support the auditory information, the effect would probably be even stronger.

2.1.1. Auditory Mixed reality

AR is often considered as a sub-genre of mixed reality (MR), a wider term describing everything between the full reality and the full virtuality. The seminal 'reality-virtuality continuum' by Milgram et al. (1995, p. 283) illustrates these relationships (Figure 1).

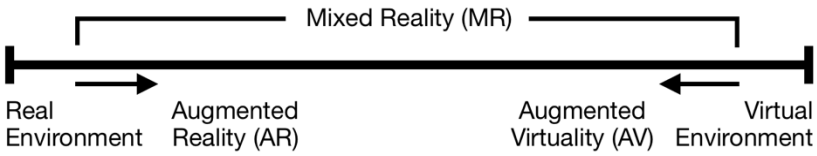

Figure 1: Reality-Virtuality Continuum by Milgram et al. (1995).

The MR concept works as an umbrella for a wide variety of approaches to the mixing of virtual with real. New applications of AR or AAR are constantly being developed with current technologies, ideas and multi-modal approaches. Therefore, the old definitions may not always be useful in understanding what is going on, or they may restrict communication of new ideas and approaches. The MR concept may be in help with that. In fact, some scholars, such as McGill et al. (2020), use the term 'auditory mixed reality' when discussing AAR with varying levels of 'acoustic transparency'. Such AAR would take use of different combinations of sensory modalities and levels of virtuality, such as 'auditory AR with visual reality, or auditory VR with visual AR, etc' (McGill et al., 2020).

One example where an AAR experience is stepping out of the traditional definitions of AR, and could be called MR, is a case where auditory stimuli are replaced with virtual auditory content while the other modalities are not manipulated at all. As an example, in the *Sounds of Silence* the visitors were wearing closed-back headphones, blocking the quiet real-world soundscape almost entirely. The virtual audio content heard in the headphones was synchronised with the visitors' locations in the space, so the audio could be experienced as embedded in the surrounding and thus co-existing with the real.

McGill et al. (2020, p. 10) have reported similar experiences with test users wearing noise-cancelling headphones in outdoors and still perceiving birdsong appear real. The researchers note that such experience 'could be described as an augmented reality experience, despite being rendered on the equivalent of an auditory VR headset.'

In the examples above, one could say that the augmentation process happens cross-modally. Even if the audio is 'VR', other sensory modalities are grounded in reality: the physical

environment is sensed with a combination of sight, smell, touch, and proprioception (sense of space). Since the virtual sounds are attached to and synchronised with the same environment, they add something to the sensed reality that was not there in the first place. The augmentation does not happen on the auditory channel but complements other sensory data from the surroundings.

It is also worth suggesting, that if the real environment is naturally completely silent and anechoic, then blocking the auditory stimuli from it would not change the perceived reality much. However, that is quite hypothetic, unless the exhibition happens inside an anechoic chamber or an acoustically treated sound recording booth. Then again, if the user's subjective conception of the real environment is silent or she does not tend to pay much attention to reverberations and small sounds like footsteps or clothing noises, then the end result using acoustically isolated 'VR audio' would not differ much from 'AR audio' where the subtle environmental sounds would be heard in the background.

If we apply cross-modal augmentation to sight instead of hearing, one can safely compare it to wearing a VR headset without headphones: the visual sensory channel would be blocked and replaced with a computer-generated virtual image, spatially synchronised with the environment thanks to the tracking system. All the other senses would keep receiving information on the surroundings. However, to complete the analogy, the virtual world projected to eyes should relate to the surrounding real environment.

Examples of such experiments are rare, but Microsoft's *DreamWalker* (2018–19) is one. In DreamWalker the test users walked at a large campus area, wearing a VR headset and tracking system, while seeing the streets of downtown Manhattan with cars and other pedestrians in front of their eyes. The system presented the user a computer-generated image of the virtual city, synchronised spatially with the real environment. Real-world obstacles were displayed as traffic cones or other obstacles in the virtual world. The experiment did not generate any sounds corresponding the virtual world, so the only mediated modality was sight. Hence, it was analogous to the *Sounds of Silence* example. The test users of DreamWalker were reported to have rated the experience as immersive despite the fact that the visual stimuli were asynchronous with the other sensory stimuli. (Yang et al., 2019)

One practical use for replacing a sensory channel instead of augmenting it would be in case where the real-world environment is too busy with stimuli, and it would be overwhelming to add more content on it. If selectively removing or masking of real-world objects is not possible,

one option could be to completely replace the channel with virtual content. For example, guidance voices in geo-located AAR systems with acoustically transparent playback devices such as *Bose Frames* would be easily audible in a peaceful environment, but in a busy street they would potentially get easily masked by the real-world sounds.

2.1.2. Immersion

To use AAR in telling stories, the ability to immerse the user in a story world—or other layers of reality—is essential. To get the user committed to the narrative, the story needs to be presented so that the user can 'suspend disbelief'. Suspension of disbelief is a concept by an early-19th century philosopher-poet Samuel Taylor Coleridge, referring to the user's attitude to overlook improbable story components that may threaten the coherency of the narrative. With modern virtual and augmented realities suspension of disbelief can be understood broadly as sensory immersion, or a 'haptic sensation of simulation'. (Karhulahti, 2012, p. 6)

In AAR that would mean that immersion requires, at least, coherence between the narrative and the sensed world, which is a combination of the real environment and augmented auditory layer. The story dependent on these two realities should not contradict itself, otherwise the immersion is at risk of breaking (McErlean, 2018). Since in AAR, with transparent auditory display, it is challenging to control how the real-world sounds, the augmented sounds need to match with the narrative while being simultaneously grounded in the environment. In MR with a virtual audio channel there is obviously more freedom to play around with the sounds to make them support the user's suspension of disbelief.

Sounds have an important role in 'making a reality seem alive' (McGill et al., 2020). Larsson et al. (2010) refer to one study where war veterans with profound hearing loss told that the world felt 'dead' with 'strange and unreal quality', and another study where participants had experienced background sounds as 'important for the sensation of being "part of the environment"'. Arguably, a fundamental element in feeling immersed is the sense of presence: presence of oneself in the situation as well as presence of other subjects and objects in it. Lombard and Ditton (1997) define presence as 'perceptual illusion of nonmediation', i.e., an illusion that there is no mediating technology between the user and the experience.

With modern virtual auditory displays (VAD) with high-quality spatial rendering and head tracking, it is becoming more difficult to distinguish between real and virtual sounds (McGill et al., 2020). According to Larsson et al. (2010) an auditory system in MR has the ability to induce

both the sensation of the user being surrounded by the space ('spatial presence') and the location of an object in the space ('object presence'). Some of the parameters contributing to inducing presence are

- externalisation (sounds appear to come from outside of the user's head)
- localisation (sounds appear to be attached to the real world)
- spaciousness (reverberations match the real-world space)
- sound quality (sound pressure levels (SPL) and frequency responses are consistent with what the user sees)
- possible mismatch between visuals and audio
  (Larsson et al., 2010, pp. 159–160)

In addition to making the sounds and soundscapes appear plausible by paying attention to the parameters mentioned above, one could argue that an important factor in creating an immersive experience is to make the virtual environment interactive and alive, reacting to the user's movements and actions. (Geronazzo et al., 2019, p. 4)

A third factor in creating suspension of disbelief and hence immersion is the story itself (Karhulahti, 2012). As McErlean (2018) notes, immersion does not need the user to be physically inside the story world or sense a three-dimensional projection of the world around her, but it can be created only with words. Since the assumption is that 6DoF AAR and AAR in general can be used to tell almost any stories, it is comforting to think that immersion is possible even without perfect technical illusions of the nonmediation, but can be achieved by literal narrative means, e.g., with believable characters and events.

In fact, not everything in a simulation needs to be a perfect copy of the existing environment in order to make it plausible: a suitable reproduction is sufficient with enough features for a given application (Blauert, 1997, p. 374; Lindau & Weinzierl, 2012, p. 804). Chion (1994, p. 107) argues that in cinema the spectator observes and resynthesises sounds according to the coded conventions of the medium. Although AAR has probably not yet developed such strong conventions, one can assume that, in order to create verisimilitude, hundred-percent authenticity is not necessary even in AAR.

Then again, unlike in cinema or VR, in acoustically transparent AAR experiences the real-world sounds are always present, and virtual sounds are compared against them. It would be interesting to study whether the quality of the virtual auditory simulation should match the real-world soundscape in order to maintain immersion, or would virtual sounds with very

different characteristics plausibly coexist with the reality. In other words, would artificial sounds work well in AAR as do cartoon characters in the film *Who Framed, Roger Rabbit?*

2.2. Virtual auditory display

In order to create a plausible illusion of an acoustic reality we have to find a way to trick the user's auditory perception to believe that the virtual sounds are real. For this we need a system that is able to playback or synthesise audio events located around the user in a virtual space and render them to the user's ears so that their propagation mimics the real world. With 'direct augmentation', i.e., loudspeakers hidden in the real-world environment (Normand, Servières & Moreau, 2012), the task is rather easy since the virtual sounds transform into real sound waves when emanating from the speakers. However, with headphones, a lot more simulations and calculations are needed: the acoustic properties of the room must be modelled, the position and movements of user's head needs to be tracked, and finally the user's head and ear pinnae shapes must be taken into account in order to calculate how sound would propagate and get filtered before arriving in the ear canals.

A system where audio spatialiser is used to binaurally render virtual audio object to the user's ears is sometimes called a 'virtual auditory display' (VAD), a term by Shinn-Cunningham (1998) (Xie, 2013, p. 37). Although technology has advanced a lot since the 1990s, the principles of a virtual audio setup have not changed significantly. For example, the AAR setups described by Blauert (1997, p. 386) and Bederson (1995, pp. 210–211) had already all or most of the main components used in modern AAR experiences (e.g., Albrecht, 2016b, pp. 22–23). However, there are yet no standards for AAR setups or VADs, and the selection of components depend greatly on the chosen technical approach and the desired level of immersion and interaction. The lack of standards and universal configurations might be, in fact, a reason the 3D audio market has been growing slowly and is still a niche (*3D Audio Market - Global Industry Analysis 2018 - 2026*, 2018).

In addition to using hidden speakers or headphone-based systems one interesting way to produce spatial auditory images around the user would be using wave field synthesis (WFS). In WFS a large number of loudspeakers are used to produce 'holophonic' sounds inside the listening space without requiring the users to wear any devices (Daniel et al., 2003). However, since WFS relies on a large array of speakers surrounding the 'play area', it is probably too complex and challenging solution for many AAR experiences.

2.2.1. Spatialiser

A key component of a VAD is spatialiser. Firstly, it simulates the acoustic properties of the real environment, or the desired illusionary environment. Then, being aware of the spatial locations of the virtual sound sources in relation to the user's position, it simulates the propagation of the sound waves from the audio objects and their environmental reflections all the way to the ear canals of the user, binaurally rendering the audio to the headphones.

Humans' ability to localise sounds is mainly based on comparing acoustic information between two ears, called binaural hearing. The two key mechanisms of binaural hearing are interaural time difference (ITD) and interaural loudness difference (ILD) (Blauert, 1997). In addition, small head movements and spectral colourisations caused by outer ears seem to help in localisation. (Xie, 2013, p. 16).

When a sound source is, for example, on the left of the listener, the sound waves will first reach the left ear since it is closer to the sound source. After a small delay the right ear will register the sound. The maximum ITD occurs when the sound is 90 degrees to the side, and is around 660 microseconds (in air). (S. Kim, 2015, p. 147) However, when the sound is coming straight from front, back, up, or down, i.e., along the 'median plane', both ears will receive the sound waves more or less simultaneously. Therefore, ITD is useful mainly with left-right localisation. Also, when the sound frequency increases, the wavelength gets shorter and thus the phase difference between ears gets more difficult to perceive. Hence, ITD can provide clues of the sound position mainly for sounds under 1500 Hz. (Akeroyd, 2006, p. 26)

In addition to ITD, the human spatial hearing registers difference of sound pressure levels between two ears, the ILD. When a sound is produced on the left of the user, the left ear receives the sound waves without obstruction, but due to the 'acoustic shadow' created by the head, the sound intensity gets reduced at the further ear. Small deviations are also produced by torso, shoulders, and ear pinnae with its cavities and other shapes, all unique to each individual. Whereas ITD is perceivable with the lower frequency spectrum of sounds, the opposite is true with ILD that works better with higher sounds; low frequency sounds with a longer wavelength diffract through obstacles with the size of the head, but with shorter wavelengths the head starts to have an occluding effect. (Akeroyd, 2006, p. 27)

For perceiving the elevation of the sound as well as whether it is coming from behind rather than front, it seems that the spectral colourisations caused by the pinnae have a role: depending on the direction the sound gets scattered, reflected and resonated slightly differently. (Akeroyd,

2006, p. 28; Xie, 2013, p. 17). Also, studies suggest that small head movements help to distinguish whether the sound is coming from front or back, up or down (Xie, 2013, pp. 13–14). However, the resolution of spatial hearing along this median plane is still less accurate than on the horizontal plane (Blauert, 1997).

In perceiving distance, the human hearing performs poorer than with directions. Distance hearing is considered being based on multiple cues, such as loudness of the sound when the sound source is familiar, high-frequency absorption by air for long-distance sounds, ILD for near-field sounds, and reflections in an enclosed space (Xie, 2013, p. 19).

It is possible to make a model of how a person's ears receive sounds coming from different directions, taking account all of the spectral changes mentioned above. For that, a microphone is placed near each eardrum, and while in an anechoic chamber, short sound impulses are played from all desired distances and directions around the person. Using the recorded signals, Head-Related Transfer Functions (HRTF's) can be generated. These HRTF's can be then used to emulate the person's three-dimensional hearing when feeding audio straight to ears using headphones. (Akeroyd, 2006, p. 28)

The spatialiser, usually a computer software plugin, receives information on the spatial locations of the sounds together with their actual audio content. It gets these from the scene generator programme, often running on a game engine. The spatialiser can also incorporate an acoustic model of the space for simulating reverberations and other surface reflections. With all this information, using either built-in HRTF's or HRTF's loaded by the user, the spatialiser makes a binaural render of the audio content for each sound source and its reflections. This process creates an illusion of sounds emanating from the environment, outside of the user's head. This effect of 'externalisation' is a key factor in making the virtual audio plausible, in contrast to 'laterilised' audio appearing to sound inside the head. (Akeroyd, 2006, p. 28)

Since the shapes of head and pinnae vary a lot from person to person, their HRTF's would also be different. Using another person's HRTF may result in errors especially with up/down and front/back detection (Akeroyd, 2006, p. 28). However, due to practical reasons, many spatialiser use a generic set of HRTF's with a 'one-size-fits-for-all' approach. Depending on the manufacturer the results may vary, although some spatialisers such as DearVR seem to perform rather well in this regard (Laamanen, 2018, pp. 36–42).

Analysing each user's head, ears and upper body before starting an AAR experience would be difficult especially in public exhibition circumstances. However, techniques and services are

being developed for fast HRTF creation for individuals (e.g., *Aural ID*, 2020), so perhaps in the near future the user can be scanned in the beginning of the experience and the unique HRTF's can be immediately applied in the spatialiser.

2.2.2. Headphones

Headphones feed the binaurally rendered audio to the user's ears and enable a subjective auditory experience in contrast to using loudspeakers embedded in the environment. A few attributes of headphones may be especially relevant in the context of AAR: acoustic transparency, audio quality, and comfortability.

The structure of the headphones dictates their acoustic transparency, i.e., how much of the real-world sounds gets passed to ears. 'Open-back' headphones are usually relatively transparent, and can be seen as equivalent to optical AR systems such as AR glasses and 'head-up display' (HUD), where the virtual world is rendered on top of the real one using optical combiners (Azuma, 1997, pp. 361–363). However, the slight colourisation of the real-world sounds caused by the headphone structure might be inadmissible for some critical applications (Jacuzzi, 2018, p. 2). There are also acoustically totally transparent solutions with nothing covering the ears such as *Bose Frames*, a pair of sunglasses with small speakers in each temple feeding the audio to ears (*Bose Frames*, 2019), or bone-conductor headphones using electromechanical transducers to vibrate the bones of the skull.

'Closed-back' and 'in-ear' headphones, respectively, isolate the user from the real-world sounds to some extent, especially at higher frequencies. This may be optimal in situations where environmental sounds would be too disturbing in relation to the virtual audio content. However, due to the acoustic structure the frequency response in closed headphones may not be as even as in open ones, although that can be compensated with digital audio processing. Larsson et al. (2010, p. 152) also note that closed and in-ear headphones may increase user's awareness of self through hearing own bodily sounds in a similar manner as with earplugs. This, in consequence, can contribute to feeling less connected with the surroundings. The opposite can be argued about acoustically transparent headphones: they would allow the user to retain the sensation of 'being part of the environment', a fundamental concept of AR and MR experiences (Larsson et al., 2010, p. 152).

Transparency can also be achieved electronically with closed headphones, using a pseudoacoustic system: a microphone array is installed on the outer surface of the earcups, and

the stereophonic audio image of the environment is fed to the user's ears (e.g., Liski et al., 2016). Noise-cancelling headphones utilise this 'mic-through' feature (McGill et al., 2020). However, without a headphone manufacturer releasing their application programming interface (API), it may be difficult for content creator to utilise this feature in AAR experiences, e.g., routing the environmental audio out of the headphones for external processing and feeding it back.

A mic-through approach allows a similar experience with AAR as see-through AR offers with visual modality where augmented images are rendered on top of a video image from one or several cameras (Azuma, 1997, pp. 362–363). Compared to an acoustically transparent AAR system, a pseudoacoustic AAR system is more complicated to build and run as the real-world sounds need to be processed and fed to the headphones, but as a benefit it enables removing and replacing the real-world soundscape, making way to MR experiences. A possible theoretical application for such a system could be an attempt to ease the registration problem of alignment mismatch caused by the end-to-end system latency by delaying the real-world sounds in synchronisation with the augmented sounds when turning head fast.

The quality of audio may affect the plausibility of the virtual sounds, being a source of registration errors for AAR, similarly as optical distortions may cause errors in visual-based AR (Azuma 1997, p. 368-369). Whereas one can argue that the headphones should be as high quality as possible, they are not the only source of possible reduction in audio quality: at least the original recorded or synthesised audio material, possible audio packaging and processing in the computer and playback system, and the amplifier system feeding audio to the headphones may have an effect in altering the spectral reproduction, dynamics or adding background hiss. Nevertheless, it seems that the human hearing is capable of adapting to changing acoustic conditions (Khalighinejad et al., 2019). Hence, it may be that auditory illusion and immersion are possible to gain even with imperfect audio quality.

Finally, the weight of the headphone unit might be a consideration for at least two reasons: lightweight headphones are more comfortable to wear than heavy ones, but they are also easier to 'forget' which potentially helps the user to immerse into the experience.

2.2.3. Positional tracking

To keep the virtual sounds fixed in their three-dimensional positions regardless of the spatial location and orientation of the user, a tracking system is required. According to the head

movements in the real world the virtual environment is rotated and moved to keep it stationed with its physical counterpart.

When an object can freely move in a three-dimensional space it is said to have 'six degrees of freedom' (6DoF): three degrees refer to position coordinates in Euclidean space (x, y and z axes) while three others refer to orientation angles (yaw, pitch and roll) (Mazuryk & Gervautz, 1999, p. 19) (Figure 2). With 6DoF tracking of the user a fully immersive AAR experience can be potentially constructed as the user can freely move in the space while the system takes care that all the sounds are emitting from their intended positions.

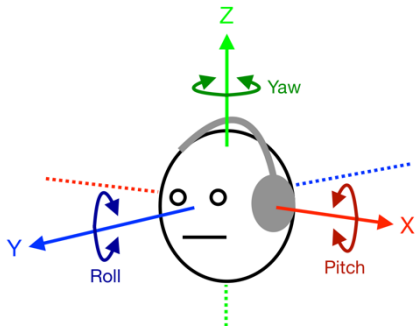

Figure 2: Six degrees of freedom (6DoF).

Depending on the AAR application not all degrees need to be tracked, however. If the augmented soundscape consists of mainly atmospheric sounds, a plausible illusion may be achieved without any head-orientation tracking, like is the case in many audio walks (e.g. Koivumäki, 2018, pp. 120–141). Yet, with point-source sounds the tracking becomes important: human hearing can notice if the position of sound changes about 1 degree (Blauert 1997, pp. 38–39).

Positional tracking can also be the primary input mechanism for the user to interact with the system, especially if no buttons, voice recognition, or other input methods are used. User's location can trigger interactive transitions (Gamper, 2014, p. 29), which is used in, for example, interactive museum audio guides to playback different narratives based on user's location in the museum (e.g. *SFMOMA App*, 2018). Also, head orientation can be used for interactive purposes, e.g. user's nods can be interpreted from the orientation data and used as interactive cues (e.g. Gampe, 2009, p. 3-5), and similarly the direction of gaze can be registered and used.

There are two main functional approaches to tracking: the system can deliver absolute data with total positional values, or relative data with changes from the last state (Mazuryk & Gervautz, 1999, p. 19). Most of the tracking systems applicable to AAR deliver absolute data, such as systems based on fixed anchors or cameras, as well as the Global Navigation Satellite System (GNSS). However, for instance, an inertial measurement unit (IMU) provides primarily relative data, and hence requires calibration more often due to drifting. Since there are multiple technologies available for positional tracking, a brief presentation of some of them follows with evaluation on their usability for 6DoF AAR applications.

UWB technology

UWB (Ultra-Wide Band) tracking uses fixed-position 'anchors' sending radio beacon signals to the environment, and a 'tag' on the trackable item, receiving signals from the anchors and using them to calculate its location in a three-dimensional space. UWB utilises a broad spectrum of gigahertz radio frequencies to transfer data in very short pulses, making it capable of estimating distances accurately. It is gaining popularity as an indoor positioning system (IPS) due to its accuracy and still low cost and low power consumption. (Dardari et al., 2015)

Other wireless radio technologies

As Dardari et al. (2015) note, several short-range wireless radio communication standards have been used for location tracking purposes, even though they have not been designed for that. Common ones are Wi-Fi, RFID, NFC, Bluetooth, and ZigBee. Not only is their accuracy poor, usually about 1-5 meters, but they present other challenges, too, including reduced possibilities for scalability.

In addition to the short-range systems, cellular networks have been used for tracking. However, since their localisation resolution is based on the network cell size that can vary from meters to kilometres, they are out of use for 6DoF AAR applications. (Dardari et al., 2015) However, with 5G technology, the indoor localisation accuracy may get better than 1 meter (Zhang et al., 2017, p. 28), making the technology feasible for at least some AAR applications.

Global satellite navigation systems (GNSS)

Satellite navigation is based on a constellation of satellites orbiting the Earth. They send positioning and timing data which is received by the trackable devices. The devices use the data to interpret their location relative to Earth. (*What Is GNSS?*, 2016) Four GNSS systems with their own satellites are operational (Galileo, GPS, GLONASS, and BDS), and modern receivers can interpret data from all or most of them simultaneously. GNSS works properly only for outdoor tracking, and whereas the accuracy is normally around 1 meter at its best, with correction data decimetre range can be reached, whereas there are technical means to reach even centimetre accuracy (Fernandez-Hernandez et al., 2018). With such accuracies GNSS would open highly interesting possibilities for outdoor 6DoF AAR experiences in a global scale.

Image-based camera localisation

In image-based camera localisation, 2D or 3D cameras are used to determine the location and/or orientation of the trackable item (Wu et al., 2018). The system can use 'inside-out' tracking where cameras are attached to the trackable item, interpreting its location according to the changes in the patterns in the environment. Alternatively, external cameras can be used in an outside-in' configuration. (*Inside-out v Outside-In*, 2017)

The inside-out approach is getting popular in many applications varying from MR displays (e.g., *Varjo XR-3*, 2020; *Microsoft HoloLens*, n.d.) and robotic vacuum cleaners (e.g., *iRobot Roomba S9+*, n.d.) to medical instruments (e.g., Busam et al., 2018).

The advantages of image-based tracking are its flexibility and low cost. Further, the same system can be used to track both location and orientation. Also, with the development of computer vision it has become possible to let the tracking device know its location even in an unknown environment using SLAM (simultaneous localization and mapping) technologies (Wu et al., 2018). Like with decimetre or centimetre-accurate GNSS, SLAM would allow creating 6DoF AAR experiences that are not location-based but can be experienced anywhere. Perhaps the virtual audio objects and their positions could adapt to the new environment or be generated procedurally.

Optical tracking

Optical tracking is usually based on infrared (IR) light and two or more IR cameras. The object may be equipped with IR LEDs, or alternatively it is decorated with retro-reflective markers on its surface which are illuminated by a ring of IR LEDs around the camera lens. The cameras detect the IR light emanating or reflecting from the object and calculate its location on a 2D plane as well as its orientation based on the marker patterns. With data from multiple cameras the 3D position can be derived. (*Optical Tracking Explained*, 2019; Förterer, 2013)

Optical tracking can be very accurate (even 0.1 mm), and with reflective markers there is no need to equip the trackable object with active electronics and power source. However, the tracking requires line-of-sight (LOS) and optimal lighting conditions. (Ungi et al., 2015, p. 472) Therefore, it may not be feasible for multi-user experiences or venues that have natural light. Also, in larger spaces the camera system required to cover the whole area may become complex.

Inertial measurement unit (IMU)

In an IMU there are usually three sensors built-in: accelerometer, gyroscope, and magnetometer. With the combined data the trackable object's rotation angles (pitch, roll, yaw) can be determined, and with the accelerometer it is also possible to measure linear acceleration. IMUs are very widely used in many applications, for instance, smartphones, aircrafts, video game controllers and head-tracking headphones. The data is relative, so drifting tends to happen over time and thus calibration is needed. (Ahmad et al., 2013) However, with a combination of multiple sensors ('sensor fusion') and use of filters and algorithms the tracking biases can be minimised.

Other methods

There are many other tracking methods which are not discussed here, including infrared laser scanning used in, for instance, *HTC Vive* VR system (*Roomscale 101*, 2017), laser rangefinding (LIDAR) used in, for example, automated cars, and acoustic localisation using ultrasound (e.g., *Alps*, 2018; Krekovic et al., 2016a).

2.2.4. Scene generator

An essential component of an AAR setup is a system that a) controls the audio material located in the spatial mapping of the environment and b) feeds the audio to the user according to the interactional rules and sensory data. When discussing augmented reality Azuma (1997, p. 363) uses the term 'scene generator' to describe such a system, which is usually a computer software. The scene generator can, for instance, be running on a game engine, be a purpose-built authoring service such as an audio-walk application, or it can be programmed specifically for the purpose.

An easy and versatile way to authorise 6DoF AAR experiences would be using a game engine such as *Unity* or *Unreal Engine*. A game engine is an integrated development environment (IDE) that has a vast number of pre-programmed elements and functionalities to help and speed up especially game making process, although it can be used to develop many other interactive and digital art forms, too. Game engine also frees the creator from many programming tasks, enabling concentration on other areas of the creative process. (Buttle, 2020)

The creation of an AAR scene would start from modelling the real-world space in the game engine editor. To have walls, furniture and other essential elements mapped serves two

purposes. Firstly, it helps to position virtual sounds so that their placement corresponds with the real environment. Secondly, the structures placed in the game engine can be used by the audio spatialiser to simulate how sound propagates in the space, for instance, how it reflects from surfaces and gets attenuated when occluded. This potentially makes the virtual space, and the individual sounds appear as real. (Sinclair, 2020)

In most cases, the virtual sounds would then be placed in the modelled 3D space as individual, monophonic sound emitters. Scripting or other methods would be used to start, stop, and otherwise manipulate the playback of the sounds. Also, their position can be changed and moved through scripting. The user would be represented in the virtual 3D space by an 'avatar', or merely a head of the avatar. It receives the positional data from the tracking system and moves accordingly. In the avatar, where the real person's ears would be, is placed the game engine's 'listener', a component that acts as virtual ears.

The audio spatialiser knows the locations and orientations of the sounds and the listener. It may be aware of the basic geometry of the room, or it can even analyse all the surfaces and obstacles in the space. With these, it will simulate the sound propagations in the space to the listener, adding appropriate surface reflections. For some sounds there is no need for the spatial audio treatment, such as internal voices and music with which normal stereo or mono sounds would be sufficient. Also, Ambisonic sounds can be used for, for instance, static 3D soundscapes and room tones. (Sinclair, 2020)

Interactivity with triggers, ray casting, and other methods can be created similarly as in video game development using scripting (e.g., C# programming language in *Unity*) or graphical coding (e.g., *Blueprints* in Unreal). Finally, the project may be compiled into an executable programme for the target platform (Mac, Windows, Android, iOS, etc.).

2.2.5. End-to-end system delay

To trick the user to believe the virtual objects coexist with the real environment, an AR system needs to keep the virtual world spatially aligned with the real world and playback the virtual audio objects to the user's ears as authentically as possible.

Azuma (1997, pp. 368–370) lists registration errors that may break this illusion, categorising them in static and dynamic errors. If we adapt Azuma's list of static errors to the context of AAR, it may look like this:

- Acoustic distortion / Imperfect audio quality of the system
- Errors / inaccuracy in the tracking system
- Mechanical misalignments, e.g.
    - User's headphones not placed properly around ears
- Incorrect listening parameters, e.g.
    - Offset between the location of the head tracker and the user's ears
    - HRTF model of the binaural rendering not suitable for the user
    - Sound volume adjusted too soft/loud

Dynamic errors, according to Azuma, occur due to the latency in the system measured from the moment the user's movement is measured to the moment the changed soundscape corresponding to the movement is played to the user's ears. Azuma calls this 'end-to-end system delay' (Azuma 1997, p. 370), whereas nowadays it is often called 'motion-to-photon latency'. Since Azuma's term is more universal and works with AAR, too, I will use it in this thesis.

The end-to-end system delay causes problems to AR experiences since it makes virtual projections slip away from their real-world locations when the user's perspective changes due to turning head or moving. (Azuma 1997, p. 367) This delay would remind the user that the world is partly computer-generated and mediated, potentially breaking the immersion. (McErlean, 2018)

According to Azuma (1997, p. 367), the demand for accurate synchronisation is more elemental in AR than in VR, because in VR the possible 'registration errors result in visual-kinaesthetic and visual-proprioceptive conflicts' which are less noticeable than visual-visual conflicts in AR. Whereas Azuma concentrates on visual-only AR with one augmented modality, in AAR these conflicts would happen both unimodally (auditory-auditory) and cross-modally (e.g., auditory-visual, and auditory-proprioceptive).

In an AAR setup the latency would accumulate as follows: When the user turns head or moves in the space, it takes some time for the tracking system to interpret the new orientation and position and transmit that information to the computer. This is sometimes referred as update rate, and it can be, for instance, 60 Hz resulting a delay of nearly 20 milliseconds (ms) (*Scaling the Creator System*, 2021). The computer software or mobile device application will add some more latency if data is filtered before used to rotate the virtual world. Then, the newly oriented auditory soundscape is rendered and outputted through the audio system of the computer. Depending on the processor power, audio driver and system configurations this stage may generate a substantial amount of latency to the chain from a few milliseconds to nearly 200 ms

(*Low-Latency Multichannel Audio in Unity*, 2019). Finally, the possible wireless audio transmit to the headphones may add some delay to the signal. If the total accumulated latency is small, the virtual soundscape will follow the user's movements quickly enough for the user not to observe any misalignment. However, with higher latencies, the virtual world will always react with a delay and appear as being momentarily out of place after each head-turn.

Nevertheless, compared to visual AR, auditory AR may be a bit more forgiving in this regard: The angular accuracy of visual perception is very high, less than 1/60 degrees (1 arcminute) (Azuma 1997, p. 368; Blauert 1997, pp. 38–39). It is equivalent of noticing when something in the distance of 756 meters moves by the length of a soccer ball (22 cm) (*'Minute and Second of Arc',* 2021). Hence, there is only little marginal for augmented virtual objects to get misaligned with the real environment before it is observed by the user. However, the spatial resolution of human hearing is much less accurate: according to tests people can notice an angular difference of about 1 degree (Blauert, 1997, pp. 38–39). That is roughly equivalent of the width of index finger when hand is extended to arm's length (*Positions and Sizes of Cosmic Objects, n.d.*).

The AR systems of the 1990s had a typical end-to-end system lag of about 100 ms. According to Azuma (1997, p. 371), with 'a moderate head rotation rate of 50 degrees per second' that would lead to the angular dynamic error of 5 degrees. An error such large would mean any virtual objects at arm's length would drag behind the real-world by palm's width when turning head moderately. With faster movements the off-sync would be even bigger. Such a latency being practically unusable for visual AR, it is not far from the theoretical requirement for AAR: with the 'localization blur' of about 1 degree for hearing (Blauert 1997, p. 38) the maximum unnoticeable system delay would be around 20 ms. Some modern VR head-mounted displays (HMD) are reported to reach baseline end-to-end latencies as low as ~2 ms (Feng et al., 2019). Even though those performance figures were obtained using heavy optimisation, they suggest that technology is available to cater the needs of AAR rather well. However, with tools and technology available for average content creators the system delays may still be one or two orders of magnitude higher.

As Azuma (1997, p. 367) notes the visual-proprioceptive and visual-kinaesthetic registration conflicts are not be as noticeable as visual-visual conflicts. As an explanation he suggests a phenomenon called 'visual capture', a term by Robert Welch (1978), where the brain tends to believe visual information over all the other senses. Chion (1994, p. 70) talks about the same phenomenon in cinema where film image 'spatially magnetizes' the sound: even when the

sound is played from a single loudspeaker behind the screen, it is perceived to emanate from different directions according to the information the image is providing. In films, for clarity, dialogue is normally mixed to the centre speaker regardless of the character's position on the screen. However, if a character is talking on a side of the screen, the audience will perceive the voice coming from the character's direction, not from the centre of the screen (unless intentionally paying attention the true source of the sound).

This 'spatial magnetization' or 'visual capture' effect may forgive some of the dynamic errors in AAR (Azuma 1997, p. 367), especially if the sounds have a visual counterpart in the real world. With sounds that are not attached to any visible object in the environment, the effect would of course not apply.

2.3. Interactivity

One can see a 6DoF AAR experience as an interplay between the user and computer with four steps: (1) The user's actions are tracked by motion sensors. (2) The computer programme reacts to this sensory data and manipulates the virtual world. (3) The auditory elements of the virtual world are projected to the user's ears. (4) The user reacts to what she hears, and the cycle starts again. (Figure 3)

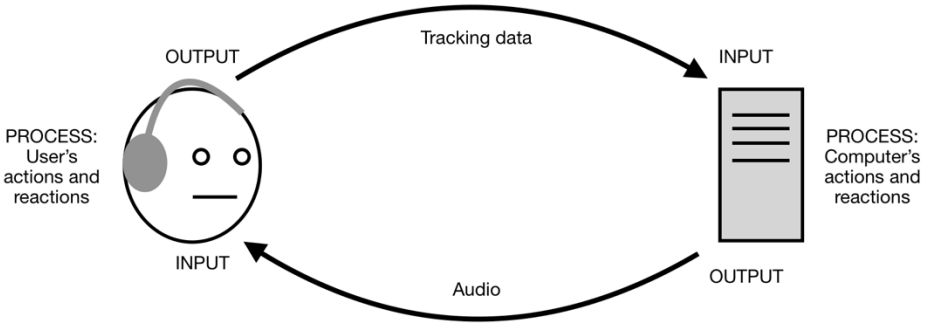

Figure 3: 6DoF AAR as an interplay between user and computer.

Thus, a 6DoF AAR experience is a type of conversation with the computer. Following Chris Crawford's (2012) definition, it has the three requirements for interactivity: input, process, and output. Or, as Crawford puts it, the system is able to 'listen, think, and speak' (Crawford, 2012). The amount and quality of interactivity incorporated in 6DoF AAR experiences is up to the creator, but technically nothing stops from using the medium for fully interactive, game-like experiences.

2.3.1. Interactive vs. reactive

It seems that the definitions for 'interaction' and 'interactivity' are quite commonly understood, but there is still rather much variation in how they are used in the context of media and art. According to the Cambridge Dictionary, interaction refers to communication or reaction between two or more people or things. Interactivity, subsequently, is exchange of information between users and computers. In this paper, I will use the words 'interaction' and 'interactive' when discussing the concept in general, and 'interactivity' when specifically referring to human-computer interaction.

With human beings in question, it is easy to make the distinction between interaction and reaction: when a person tells a joke and the other person laughs to it, we have a reaction. However, as soon as the joker reacts to the laughter—let it be a gesture or an extra pause before the next joke—we witness interaction. A two-way connection is established between the two agents.

However, with narrative media pieces the definitions tend to get blurry. The adjective 'interactive' seems to be used whenever the user can manipulate the piece. There are children' books where the reader can find new pictures when opening small flaps. There are 'gamebooks' where the reader can choose her own story path from the narrative branches. In many museum exhibitions visitors can press buttons to start a video or see lights turning on in a diorama. These examples are often called interactive books and interactive exhibitions.

However, some scholars think that a piece needs to truly react to user's actions in order it to be called interactive. This, of course, requires a computer programme, human or some other sophisticated system in order to understand the input and generate a reaction to that. (Rouse & Holloway-Attaway, 2020, p. 9)

Crawford (2012) emphasises that interactivity requires two-way communication. It is 'a cyclic process - - - in which each agent alternately listens, thinks, and speaks - - - '. Hence, in his opinion, traditional films, paintings or books are not interactive even though they manage to evoke reactions in audience and spark discussion. The art pieces are just 'speaking' without having conversation; the pieces do not participate in the interaction in any way. (Crawford, 2012)

2.3.2. Viewpoint interaction

Interactivity in narrative contexts often refers to the audience's or user's ability to manipulate the storyline through their actions and reactions (Crawford, 2012; Glassner 2017; McErlean 2018). However, it is often the case in interactive pieces that the audience can interact with narrative elements other than the storyline itself. A typical example is what I call 'viewpoint interaction': by interacting with the world through exploring and changing perspectives the user discovers new information, which can change the way characters, subjects or the story are interpreted. This is common in interactive books throughout their long history (Rouse & Holloway-Attaway, 2020), or interactive museum exhibitions. Holger Förterer's installation *Sound of Things* (2013) is also an example of this approach, to be discussed later. 'Viewpoint

interaction' is also used in many video games where the player can find additional story content that deepens the lore but does not necessarily affect the course of the main narrative, *The Turing Test* puzzle game being an example.

Kelly McErlean (2018) notes that interactive storytelling enables different interpretations of the text and a unique experience for each viewer. A shift of viewpoint can dramatically change the interpretation of the narrative, especially if pre-planned by the author. On the other hand, since the experience is interactive, the user can decide not to change the perspective and rather experience the narrative on the surface level.

Crawford (2012) presents a concept of 'storyworld' where the storyteller is not feeding the player with pre-planned events forming a 'picture-perfect view of the truth', but instead where the player can interact with ideas, observe them from various angles and try different things: 'Storyworld presents truth in three dimensions, including the less elegant angles.'

Although Crawford speaks metaphorically of the three-dimensional storyworld, one could draw an analogy to the built-in 3D nature of the 6DoF AAR experience. The player can move around, walk closer to sound sources, or further from them, allowing her to concentrate on the neighbours arguing behind the wall instead of listening to the narrator's voice (unless the narrator is deliberately mixed so that one cannot but listen to it). Even though the virtual characters and objects in the storyworld cannot react to each action of the player, the change of perspectives alone may enable the player to experience and 'read' the story differently from other players. Temporal freedom, or ability to move and listen at own pace, may also have a similar effect.

Another example of a potentially interactive element is atmosphere. Perhaps not as common as viewpoints, atmosphere can however be another way to change the way the audience experiences the story. In 'Emotive VR', an experimental 'neuro-interactive' 360 film by Marie-Laure Cazin, the viewer is wearing an EEG headset which measures her valence and arousal levels during a scene in the film. I was involved in the project as sound designer and composer and was commissioned to construct the music for the scene so that it reacts to the emotional data in real-time, changing the feeling of the music accordingly: negative valence triggers a dissonant variation of the music, while positive valence produces a major key version. It can be suggested that with the changes in music the atmosphere of the scene changes, too, which consequently may affect the viewer's emotions, starting the cycle again. (Harju, 2019)

2.3.3. From one-way to two-way experiences

One way to approach interactive storytelling is to classify the experiences according to their level of interactivity. Glassner (2017, pp. 21–25) presents a continuum of 'reactive environments' when discussing interactive storytelling. The list starts from static 'one-way' experiences such as novels that do not change at all. In the middle of the list there are 'rail ride' and 'riding the current' type of linear narratives that allow some control for the player. The list ends with 'two-way' experiences that require a lot of participation and conversation, including participatory stage plays and role-playing games.

Glassner (2017, p. 25) points out that participatory experiences are demanding to the players. According to him, that may be the reason why one-way experiences such as books and films are mainstream, whereas role-playing games and other two-way media 'are still picking up speed.' While computers and technology are making the participatory experiences easier to create and enjoyable for more people, Glassner sees that the further we go on the list the more important it becomes to have other human participants in the interactive experience. As he puts it, 'Computer are getting better at playing chess, but they still can't hold a decent conversation.' (Glassner, 2017, p. 25)

Participatory, interactive stories require a lot also from their creators (Crawford, 2012). The difficulty of the artform is probably another reason for the slow take-off of the truly two-way media. It can be argued that many 'interactive' art pieces actually provide quite limited possibilities for the audience to interact with or manipulate the story itself, but rather offer ways to customise how the experience looks or sounds like in a tightly set framework. Glassner (2017, p. 18) points out that 'The problem with giving the audience control of the story is that they will often, quite reasonably, act to reduce tension and avoid conflict. But tension and conflict are at the heart of great stories.' Whereas that may be true, the decision to limit audience interaction may be due to limited productional resources and lack of experience and made well before even facing the problems Glassner presents.

2.4. Narrative techniques

A story, or narrative, is 'a chain of events in cause-effect relationship occurring in time and space' (Bordwell & Thompson, 1997, p. 90) situated and interpreted in a specific discourse context (Herman, 2009, p. 14). In addition, a story typically conveys the experience of 'what is it like' to be living through the events in the storyworld (Herman, 2009, p. 14). Hence, the

storyteller's task is not only to explain the progress of events and educate, but to create a world around the events and characters and immerse the audience into that storyworld (McErlean, 2018).

The noun 'narrative' is commonly used synonymously with 'story'. Yet, narrative also carries its own distinctive meaning as a version of story events, often to create a manifestation or drive an agenda (Halverson, 2011; *Story vs. Narrative*, 2014). For simplicity, however, in this thesis, the terms are treated as interchangeable since the focus will be more in the way stories and narratives are told, the narrative techniques.

Depending on the medium the practical ways to fulfil the storyteller's tasks vary. In literature, the techniques available for the author to express her ideas and strengthen the narrative are sometimes called literary devices (*45+ Literary Devices and Terms Every Writer Should Know*, 2020), literary techniques, figurative language (*The English Literary Techniques Toolkit for The HSC*, 2018), narrative techniques or narrative devices (*Narrative Techniques in Writing*, n.d.). A vast number of techniques can be identified, pertaining to at least setting, plot, perspective, style, theme, character, and genre. Examples of the techniques range from stylistic approaches such as 'satire' to word-level rhetorical devices (*Literary Devices and Terms*, 2020).

Whereas many of the narrative codes in other media are based on the centuries-long tradition of the literature, each media still has its own techniques to convey the narrative elements. Sound designer of a radio drama utilises sound design techniques to, for instance, create a caricature of a sound to make it more understandable or change the way a narrative event is interpreted; some of the techniques applied could be equalising the sound's frequency content, manipulating its amplitude envelope, or layering multiple sounds on top of each other (Beck & Des, 1996).

In established art forms the techniques and their effects, the narrative codes, are familiar to the audience: for example, spectators of conventional cinema have no difficulties in following stories presented through the use of montage (Kuhn, 1985, pp. 208–209) even though the effect of the technique is rather detached from the everyday human experience. With younger media the narrative codes may be still developing, which is arguably true with MR and VR experiences in general, and namely with 6DoF AAR. Although these particular media have been developed since the 1990's, only the recent technological advancements have enabled almost anyone to start experimenting with MR and VR and try out their own narrative approaches; what works and what does not. Presumably, these virtual media are borrowing and utilising a combination

of techniques from other art forms such as video games, cinema, literature, sonic media, and theatre. However, due to the special nature of MR and VR using spatially grounded, three-dimensional sensory illusions to convey the stories, it is assumably inevitable that these media will eventually form their own, universal narrative codes and techniques to efficiently serve the audience.

3. RELATED AAR EXPERIENCES

In this chapter I will discuss four projects that represent or are related to 6DoF AAR and have similarities to my own prototype experience. Two of the projects I have experienced personally; two are discussed based on descriptions in pertinent academic reports.

3.1. Sound of Things

Sound of Things (Der Klang der Dinge) is an interactive AAR installation by a German sound designer Holger Förterer. It is the only one of the four experiences in this chapter utilising full 6DoF tracking and binaural sound throughout the experience. The installation was originally made in 2013 as Förterer's final project at the Karlsruhe University of Arts and Design, Germany (Förterer, 2013). I personally experienced the installation at *klingt gut!* sound symposium in Hamburg, June 2017.

The installation was set in a darkened room. In the middle of the room there was a wooden table with a desk lamp casting some light on items laying on the table: a wine glass, notebook, ashtray, candle, etc. When walking around the table and leaning closer to it the visitor could hear sounds emanating from the items: for example, the notebook produced sounds of scribbling whereas sounds of celebration emerged from the wine glass. As Förterer describes, '[m]ost of the sounds were created by striking, tapping, rubbing and kneading the very things on the table. The sounds of some items are further augmented with noises that are acoustically or socially in close relationship with the respective object.' (Förterer, 2013) The spatial positioning of the virtual sounds was extremely accurate, feeling almost magical.

The installation did not take into account any real-world environmental sounds, other than the natural room acoustics that the visitor could feel while immersing into the augmented sound world. The experience had no overall narrative but consisted of these individual items with the sounds embedded on them. The visitor was free to let her imagination wonder and form a story around the items. The approach was lyrical and associative rather than strictly narrative. However, the installation showed, in my opinion, what huge potential AAR with full 6DoF tracking would have for storytelling and other immersive needs.

Two visitors could experience the installation simultaneously wearing wireless, open-back headphones installed with three infrared (IR) LED's. Optical tracking with eight IR cameras were used to deduce location and orientation of the headphones. All the processing and sound

generation, including binaural rendering, happened in a software programmed by Förterer himself using C++. (Förterer, 2013)

3.2. Sounds of Silence

Sounds of Silence was a temporary exhibition at the Bern Museum of Communication, Switzerland, running between 9.11.2018 and 7.7.2019. As the name suggests, the exhibition explored silence in the modern world filled with noises. It was realised by a Basel-based company Idee und Klang Audio Design. They used a 6DoF-capable virtual audio system called *Usomo* by a German company FRAMED immersive projects. The exhibition used an MR approach with a virtual auditory channel instead of acoustic transparency. Also, one of the trackable dimensions (elevation) was intentionally left unused, hence making the experience 5DoF rather than full 6DoF. Still, with these deviations the exhibition was an extremely relevant reference and closely related to the topic of this thesis.

Figure 4: Visitors at Sounds of Silence exhibition at Bern Museum of Communication.

I visited the exhibition on 13 January 2019 (Figure 4), after which I conducted an email interview with one of the designers at FRAMED to learn more about the practical realisation and design principles behind the exhibition.

The exhibition covered a large hall separated into multiple rooms with curtains and light walls. Each museum visitor was handed a pair of closed, around-ear headphones (*Sennheiser HD 200 Pro*) equipped with an Usomo tag (Figure 5). Along with the headphones the visitors were given

an Android mobile phone running the content. The Usomo system uses UWB technology for location tracking, and inertial sensors for head orientation tracking. According to FRAMED, their system has no limits for simultaneous users, and in fact, there were dozens of headphone units available at the start of the tour.

Figure 5: Usomo tag.

Multiple UWB anchors were installed on the ceiling. Even though the system technically allows full 6DoF tracking, the designers had limited location tracking to horizontal axes only, thus omitting the elevation data. According to the interviewed designer, measuring the z axis caused problems for the sound designers as there were "too many possibilities".

Scenes changed automatically based on the visitor's location in the exhibition space. In many spots location tracking was also used for interaction within the scene. Several scenes consisted of head-locked soundscapes with what sounded like stereophonic sound, instead of being binaural, or at least externalisation was very mild. That too may have been a design choice since according to the interview, the designers valued movement-based interaction over the binaural effect.

Each visitor was able to experience the exhibition at their own pace. Due to the use of closed-back headphones the outside voices were muffled, which provided a peaceful experience that helped to focus on the sounds and silences. In the context of the theme of the exhibition this 'VR audio' approach was justified. Since the environment looked and felt artificial with abstract visualisations and video projections, and the acoustic reality in the space was quite silent and steady, using acoustically transparent VAD's would probably not have provided much extra value.

The overall style of the sound design was realistic, and no or little non-diegetic music was used. A narrator's first-person voice guided the visitor throughout the exhibition. It was mainly played monophonically, appearing to sound inside the visitor's head. One could compare it to the first-person narrator in literal narrative techniques. Lateralised audio, especially speech, is also something people are nowadays used to hearing whenever they are wearing headphones while having a telephone or internet call, listening to a podcast or radio, etc. Although in 'normal life' hearing a voice inside head is not too common, it has become an everyday experience with the wide spreading use of headphones.

In two occasions the narrator's voice was externalised and spatially locked with the environment. That happened, for instance, in a virtual concert of John Cage's 4'33" where the invisible narrator was whispering while sitting next to the visitor. Sounds were also attached to physical objects a few times: in one scene there was a group of hanging Styrofoam balls, each of which emitting a different sound and demonstrating how sound pressure levels differed from each other. This technique utilised the full tracking capabilities of the setup with the head-orientation tracking and not just the location tracking. However, due to the end-to-end latency in the system, the tracking could not always keep up with head movements causing the virtual sound to detach from its object. Perhaps this was one of the reasons why the soundscapes in the exhibition were designed head-locked instead of being spatially synchronised: constant registration errors may have been nauseating in the long run.

Here is a brief overview of some of the typical scenes of the exhibition. The names of the scenes are my own.

Snow landscape

When stepping to the first room the visitor sees a big video screen showing a video of a calm winter landscape. A soundscape loop starts, its content and feeling matching the image, with additional footsteps in the snow and some other sounds not shown on the screen. The narrator starts with introductory words (in German or French; language chosen in the beginning of the tour): "Hello. Don't say anything, just stay here..." The big video projection and the soundscape serve the purpose of displacing or immersing the visitor into another place. Visitor's movements have no effect on the sounds since they are not spatially synchronised. Walking outside of the room fades out the soundscape and switches the scene.

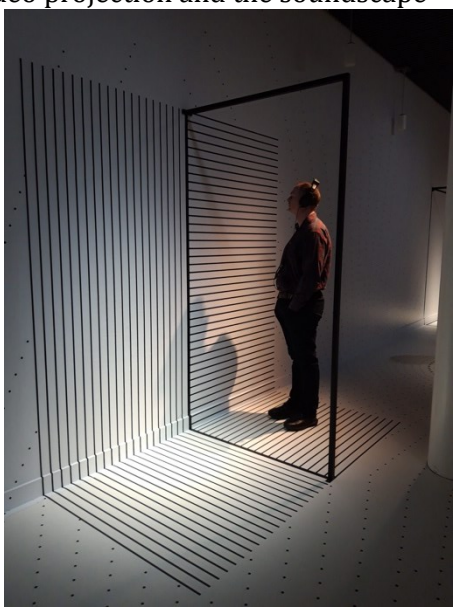

Figure 6: User listening to night sounds in the Day/Night scene.

Day/Night

There is a simple door frame through which the visitor can walk (Figure 6). On one side the visitor hears domestic daytime sounds, on the other side nigh time sounds. The setting is visually and spatially very abstracted; a symbolic door between day and night which themselves are just 1m^2 areas drawn with lines on the

floor and wall. The sounds are truly augmenting the abstracted spaces into something more, trusting the imagination of the visitor to 'see' the daytime and nigh time home around her.

Airplane

When visitor sits on a sofa, she hears the soundscape of an airplane interior with loud hum, people talking, children crying... When bending backwards the sounds get muffled as if noise-cancelling headphones were put on. This was a creative and impressive example of using a tight trigger zone in the 3D space with accurate location tracking of the Usomo system. Otherwise, it was rather similar to the *Day/Night* scene.

Hanging balls

Each grey-painted Styrofoam ball emits a sound with a different sound pressure level. The sounds are spatially attached to the balls, making it possible to walk around them, although their timbre or other qualities stayed the same regardless of the listening angle. The balls do not match contextually with the sounds other than with their size: the bigger the ball the louder the sound. As with many other scenes, this one was a non-linear, temporally free demonstration where the user could explore the sounds without any fixed time frame.

John Cage 4'33

Each visitor can enter a "concert hall" at their own time. In the room there are chairs arranged in rows, and the front wall is covered with a photograph of a concert hall (Figure 7). When sitting on a chair the performance starts. The invisible narrator is sitting next to you, spatially locked in the environment, and sound externalised using binaural processing. The concert plays back a surround recording of a real performance of the 4'33" by the Stuttgart city orchestra with the sounds of the audience moving in their chairs and coughing from time to time.

Figure 7: Visitors listening to 4´33" in their own temporal universes.

This scene really utilised the possibility of each visitor's individual temporal universe while still providing the illusion of everyone enjoying the concert in the common space at the same time. It also played interestingly with multiple layers of reality and time: The present-day real-world

environment in Bern was layered with the temporal and spatial reality of the original concert event in Stuttgart. These two were then layered with the narrative of the Sounds of Silence exhibition, running temporally independently in each visitor's headphones while sharing the visual and spatial cues of a common concert hall experience.

3.3. Growl Patrol

Growl Patrol was a geolocative audio game utilising head tracking and binaurally spatialised audio on a horizontal 2D plane. It was created for research purposes at the Queen's University in Ontario, Canada, and reported by Kurczak et al. (2011). This brief description of the experience is based on the authors' article.

The game was played outdoors in a park. In the premise of the game a number of animals, cats, dogs, and birds, had escaped from the local pet shop. They were now running around the park, and the player was charged with catching the animals and bringing them back. A hungry tiger presented an obstacle attempting to steal the animals from the player.

The game was audio-only; no visual display was used. The player was wearing headphones and carrying a smartphone with the game content. The phone tracked the user's position with GNSS ('GPS'), and the user's hat was equipped with a head-tracking system comprising a gyroscope and compass. The animals were represented by distinct, repeating sounds: cats meowing, dogs barking, and birds twittering. The tiger was growling low when patrolling and roaring fiercely when chasing the player. The player needed to come within 10-meter radius around the escaped animals in order to catch them. When an animal was caught, a homing beacon sound guided the player to the pet shop.

Although the game was created to test the usability and immersion of ambient audio versus spoken commands or visual display, it demonstrates some of the narrative possibilities of an outdoor AAR experience with location and head-orientation tracking. One prospect is the ability to create immersion through combining auditory-only information with physical movements. Regardless of registration errors caused by slow tracking and blurry directionality of the audio spatialisation, interacting with invisible animals by only hearing them apparently managed to create a sense of immersion among the players. The intensive gameplay loop and the fact that the game was located outdoors in a real-world environment must have contributed to that.

Also, interestingly, it seems that most or all of the game objects did not have a real-world counterpart nor were they attached to any physical objects in the park. However, they were still embedded in the environment and moving there among the tangible reality. This potentially creates a narratively fascinating world of another reality with escaped ghost animals coexisting and interacting with the physical world.

3.4. Hyperkuulo

Hyperkuulo was a geolocative audio story walk premiered in April 2013 in Tampere, Finland. It was realised by Ari Koivumäki with three students of Tampere University of Applied Sciences (TAMK). The experience was created using *noTours* audio walk authoring platform by *escoitar.org* collective. The users downloaded the app and the corresponding audio content to their smartphones and wore their own headphones. The phone's GNSS tracking was used for positioning. As is the case with most audio walks, the audio was head-locked since such experiences are usually made for anyone to experience them independently with their own devices and no dedicated equipment can be provided. Even though Hyperkuulo did not utilise head-tracking, it is discussed here since it demonstrated some interesting narrative techniques and concepts. The descriptions of the story walk here is based on Koivumäki's (2018) doctoral dissertation where he construes the project with detail.

In the premise of Hyperkuulo the user takes part in a scientific experiment where her brain has been implanted with a device providing extraordinary hearing capabilities, thus the name Hyperkuulo ('Hyper Hearing'). The experiment is conducted by a megacorporation whose artificial intelligence (AI) is guiding the user through voice commands in the headphones. The story is set in a park, and the voice gives the user different tasks to test the hearing implant, making the user proceed in the park from one geolocative spot to another while following the story. The tasks may involve the user to do some actions that engage the user in a sympathetic way but obviously cannot be verified by the audio walk system. For instance, the user is asked to hug a tree to hear the sounds of birch sap (spoiler: no birch sap sound is played), walk around a fountain counterclockwise while humming 'omm', and make noises on an outdoor bandstand to create reverberations.

Due to the enhanced hearing sensitivity, the user can hear, for instance, grass grow, rats squeak and people bath inside a distant sauna. The user can also hear sounds of a past in the form of a

half-minute sound effects narrative of a 1930s shipwrecking as well as sonic visions of the future and its threats.

Hyperkuulo presents a few interesting narrative approaches and techniques worth mentioning. Firstly, it forms a coherent, linear story. It is sequentially interactive in the sense that the user can walk it through in her own pace, and the story progresses only when the next location trigger is entered. Also, there are tasks that invite the user for interaction with the surroundings, although these actions cannot be registered and reacted by the system. Setting a story in a real-world environment imposes some challenges, especially when the user needs to go through trouble to find the next story point in the terrain while comprehending the narration and sounds in the headphones. As Koivumäki (2018, p. 123) notes, to design such a pervasive story requires good planning so that the listener stays interested in the journey and waits for new story events, otherwise there is a risk that the environment grabs attention and the user abandons the story.

One interesting narrative technique Hyperkuulo uses is auditory 'zooming'. Thanks to the hearing implant, the user is able to acoustically zoom into a distant sauna and hear the sounds inside of it. Without head-tracking and knowledge of the user's orientation in relation to the sauna, the sounds cannot be spatialised to a certain direction. Hence, the zooming effect probably needs further explanation for the user to understand where the sounds are originating from and why. However, with a 6DoF VAD the zooming effect could potentially feel quite natural without a further account, and rather easy to realise, too. The zooming effect is also conceptually interesting since it forces the user to eavesdrop without a possibility to look away as is the case with visual displays (Koivumäki, 2018, p. 129).

Especially in transitions between tasks music is used in the background. It is apparently diegetic as if being a part of the experiment to calm down the test subject. In that sense, it supports the narrative of the megacorporation manipulating the user. Music is also used a few times as a way to direct the user to a right location. However, being a head-locked experience, it seems that the loudness of the music is the only directional cue.

The fact that the user has to wear headphones is cleverly motivated by the story as an essential part of the experimental setup. Consequently, all the sounds heard through the headphones can be explained as being generated by the experiment. Thus, there is no narrative need to make the virtual sounds feel authentic, albeit that would potentially enhance immersion. The

experimental setup would also partly explain why all the sounds, even the ones enhanced by the implant, are head-locked.

As is the case with geolocative audio stories and walks, the content creator has very little control on what kind of headphones the user is wearing, whether they are acoustically transparent, blocking or something between. Hyperkuulo apparently avoided the problem by not counting too much on the interplay between real-world sounds and virtual sounds, but instead investing in, at least, cross-modal interaction with the real world and interesting narrative.

4. EXPLORING NARRATIVE POSSIBILITIES OF 6DOF AAR

This thesis explores audio augmented reality (AAR) with six-degrees-of-freedom (6DoF) and its narrative possibilities through the following steps:

1. Identifying a series of narrative techniques characteristic to 6DoF AAR
2. Designing and constructing a prototype of a 6DoF AAR setup
3. Analysing and discussing the design and construction process of the prototype
4. Testing the usability of the techniques and assessing the capabilities of the prototype through five demonstrative scenes made for the prototype and utilising some of the identified techniques

The process of identifying and describing narrative techniques of 6DoF AAR has been a constructive exploration to the ways narrative elements could be presented in the medium. It has attempted to answer the first research question:

> *What are characteristic narrative techniques of 6DoF AAR?*

The identification work has been based on analysis of other related experiences in their use of narrative devices, namely *Sound of Things* by Holger Förterer (2013), *Sounds of Silence* at the Bern Museum of Communication (2018–2019), *Graw Patrol* by Queen's University, Ontario (2011), and *Hyperkuulo* by Tampere University of Applied Sciences (2013). In addition, literature has been revisited, explicitly Michel Chion's *Audio-Vision: Sound on Screen* (1994), and Ronald Azuma's *A Survey of Augmented Reality* (1997) together with other writings on AR.

The methodological approach has been mainly inductive and descriptive, although I have included some techniques not empirically discovered but rather deductively reasoned based on my earlier experience and knowledge around the topic.

While identifying the narrative techniques, I have been simultaneously designing and constructing the prototype of a 6DoF AAR setup. From the research perspective the prototype building project has had two purposes:

> a) It has been the subject of observation itself, providing data on the design and creation process to better understand any possibilities and limitations the technology may set for the narrative use of the medium. The construction project has also influenced the process of identifying the narrative techniques, creating natural oscillation between the practice and conceptualisation.

b) It has enabled the creation of five demonstrative scenes for testing some of the identified techniques in practice.

Hence, the process of designing and constructing the prototype has attempted to answer the second research question:

> *How to design and build a 6DoF AAR experience demonstrating some of the narrative possibilities of the medium?*

By reflecting upon my own diary notes, photographs, audio files, authoring tool project files and code snippets, I have discussed and analysed the prototype building project from the initial idea of summer 2017 up to the current version of spring 2021. The five demonstrative scenes are presented and explained with brief analysis on what functions the chosen narrative techniques possess in them.

Finally, my personal observations on the prototype process and the use of narrative techniques in the demonstrative scenes are listed. However, within the limited scope of this thesis, no data from outside test users has been gathered and analysed. Therefore, more definite presumptions of the effectiveness of the techniques and the prototyped scenes cannot be made.

4.1. Narrative techniques of 6DoF AAR

Based on the theories and related AAR experiences discussed above, I have identified some narrative techniques I consider characteristic to AAR, and particularly AAR with six-degrees-of-freedom (6DoF). I will propose a list of techniques mainly based on the spatial audio capabilities of the medium and interplay between real and virtual.

Many of the listed techniques are not exclusive or unique to AAR; they may be applicable to other multi-modal media such as cinema, video games, visual-based AR, and VR. In fact, when elaborating the techniques, I will be frequently referring to theories of cinematic audio, namely concepts described by Michel Chion (1994) since he has seminally analysed many phenomena and techniques related to not just cinematic audio but mediated sound in general. However, the strong visual storytelling component of the projected image in many other multi-modal media would probably have a strong effect on how the techniques would work if applied as they are. Hence, I consider the identified techniques characteristic to AAR with audio being its primary channel of conveying information about the storyworld.

The techniques listed are not contextually 'equal', for example, some are related to static audio positioning while some describe methods for triggering interactive events. Also, many techniques of, for instance, sound design are left out of this listing since they are common in other media, too. An example of such techniques would be juxtaposing of an unrealistic or imaginary sound with a realistic real-world object, something that was done in *Sound of Things* with the enchanted wine glass emanating sounds of people having a party.

My proposal for narrative techniques characteristic to 6DoF AAR is:

Spatial positioning techniques

- Attachment
- Detachment/Acousmêtre
- Location affiliation
- Spatial offset
- Within-reach
- Out-of-reach
- Spatial asynchronisation

Contextual techniques

- Match
- Mismatch
- Additive enhancement
- Masking
- Manipulation

Examples of dynamic techniques

- Change between attached and detached/acousmatic sound
- Change between matched and mismatched sound
- Zooming

Input methods

- Location
- Ray casting

4.1.1. Spatial positioning techniques

Attachment

In this technique the sound is spatially attached to a physical real-world object. The attempt is to create an illusion of the object emanating the sound. This was the main technique used in *Sound of Things* where objects on the table emitted virtual sounds. The technique was also used in *Sounds of Silence* with sounds attached to, for example, hanging Styrofoam balls. In the same exhibition, the 'concert' of John Cage's 4'33" was a surround recording of a live concert, spatially locked in relation to the room of which front wall was tapestried with a large photograph of the concert hall. Although the individual virtual sounds were not attached to any real-world objects, they were loosely attached to the objects portrayed in the picture and imagined existing around the user. This created the effect of synchronising two sensory modalities with each other albeit the stimuli on both being more or less virtual.

The technique is analogous to what Chion (1994, p. 71–72, p. 78) calls 'visualized' in the context of cinema. With visualized he refers to sounds that have a corresponding object on the film screen. However, in the context of AAR I prefer the term 'attached' over 'visualized' since the user can sense the environment with multiple sensory modalities, not just sight.

Sounds can also be narratively attached to real-world objects without using spatial synchronisation, something that was used in *Hyperkuulo* with, for instance, the growing grass and sauna scenes. Although the user may imagine the non-spatialised sounds coming from an external object, perhaps through 'spatial magnetization' Chion (1994, p. 70), the technique falls out of this categorisation since it is not really using the spatial audio capabilities and has a strong dependency on the context it is presented in.

It is also worth mentioning that the sound source of an attached sound can be either 'at sight' or 'out of sight', or more generally speaking it can be either directly sensed or not. For example, a sound of burning wood could be attached to a gas heater behind a corner. The user could sense the heater by infrared radiation and thus verify it is real, although the augmented sound would suggest it is a fireplace.

Detachment/Acousmêtre

In this opposite technique, sound is not attached to any physical object or location in the environment, but it is attached to a static or moving point in the 3D space. The object emanating

the sound exists only in the virtual world. The sound is 'acousmatic', a term with Greek roots theorised by Jérôme Peignot and Pierre Schaeffer describing a sound that can be heard without seeing its cause (Härmä, Jakka, Tikander, Karjalainen, & Lokki, 2004). Chion (1994) calls this 'acousmêtre' and gives two examples: ghosts, and *HAL2000* in *2001: A Space Odyssey*. As is the case with HAL2000, the sound source of the computer is visually revealed at one point in the film, thus the sound being 'de-acousmatised'. The same can be applied to AAR with a detached sound transforming to an attached sound when the sound source is revealed.

In the beginning of the *Sounds of Silence* the third-person narrator is talking inside of the user's head, mixed monophonically and head-locked without spatialisation. We know that the voice is not our own, but, at the same time, we do not know where it is coming from, where is its source. However, after a while, the voice moves out of our head and starts to talk from a painted dot on the wall. It has now 'revealed' itself as a spatial being and attached itself to the environment. Still, we cannot see, smell or touch it; we can just register the dot on the wall. Later in the exhibition, in the *4'33"* scene, the narrator comes out of our head again and sits next to us, again invisible and untouchable. Now there is no real-world object the voice is attached to, so it is still detached but at the same time de-acousmatised.

The voice of the AI and the background music in *Hyperkuulo* were acousmatic in a sense that the user could not see their sources. However, as the story's premise was that the user wore headphones in order to hear the voice commands and soothing music generated by the computer, the source of the sounds could have been localised into the headphones or perhaps the smartphone carried by the user.

Location affiliation

Here, a soundscape or a group of sounds is affiliated with a physical real-world location whereas individual sounds are not necessarily bound to any particular object (detached/acousmatic). In other words, the sounds appear when the user enters a defined area and disappear when exiting the area. This is the de facto technique in geolocative audio walks, *Hyperkuulo* being an example. This technique was also used extensively in *Sounds of Silence* with soundscapes linked to certain zones in the exhibition space, often marked and framed with visual indicators or physical objects. The virtual soundscapes existed only in those locations, being embedded to the abstracted real-world indicators and objects, much like in the *4'33"* but the environment being even more suggestive.

Spatial offset

In Spatial Offset, sound is attached to a physical object, but it is spatially offset. An example would be an airplane passing by: in real life its sound would be appear being behind the visual object due to the speed of sound being slower than light, but for artistic reasons it may be necessary to place the sound closer to the airplane's real position. With acoustically transparent headphones, offsetting real-world sounds would of course be difficult, although possible with e.g., masking the original sound with a louder virtual sound.

Within-reach

In this technique the sound is placed inside the 'play area': user can walk around the sound source and observe it from different angles. For example, musical instruments often emanate different sound qualities to different directions, and with large instruments such as piano this is easy to notice when walking around it while someone is playing. The technique can be applied to attached or detached/acousmatic sounds, and it was used in *Sounds of Things*, and *Sound of Silence.* In *Growl Patrol* all the acousmatic animal sounds were within-reach in a sense that they were inside the play area. However, since the animals were caught as soon as the player entered the 10-meter radius around them, it may actually be that the user was never able to actually reach the sounds.

Out-of-reach

In contrast to the previous technique, here the sound is outside of the "play area" i.e., user cannot walk around it. For example, the sound is heard behind a (real-world) window, or even through walls, thus spatially extending the story world (Chion 1994, p. 86). This phenomenon also happens as a 'side effect' when the sound system registers user's head-orientation but not location. For example, in *4'33"* of *Sounds of Silence* the surround soundscape was head-tracked but not 6DoF and thus the user could never reach any of the sounds and walk around them, because the soundscape was ego-centric: when walking towards the individual sounds they always escaped like a rainbow.

Spatial asynchronisation

This technique refers to intentionally synchronising the spatial reference point of the audio, the origin of the virtual world coordinate system, with something else than the surrounding environment. Even with a full 6DoF experience it may be sometimes desired to, for example,

headlock the audio so that the sounds would be relative to the user's own head rather than the environment. That would be desired when, for example, the experience requires non-diegetic music which is not a part of the storyworld and hence should not be spatially locked with the surroundings. Also, voice overs may benefit from being head-locked for the same reasons. That was the case with most of the scenes in *Sounds of Silence*. In general, head-locked sounds may be useful when attempting to communicate that certain sounds or soundscapes are internal, belonging only to the user.

Other reference points other than the real-world environment or user's head would potentially break the immersion easily since the sounds would appear as moving quite unnaturally in relation to the surroundings. If the origin of the virtual world was another person in the installation, the user would hear the surrounding sounds move and rotate arbitrarily whenever the other person is moving. Nevertheless, spatial asynchronisation may be an interesting technique to experiment with.

4.1.2. Contextual techniques

Match and Mismatch

With this technique the augmented sound contextually either matches or mismatches with the real-world object. However, it is often apparent that whether something is considered as matching, or mismatching, depends heavily on the narrative approach and genre. For example, in a story that is 'realistic' in style, adding a virtual purring sound on a real, silent cat would be considered as a match, potentially resulting in illusion that the cat purrs even if it does not. Then again, adding to the cat a sound of someone talking would be a mismatch since cats do not speak. However, in fairy tales talking cats are regular, so not just the purring would match but also the speech.

Further, if the cat's lips are not moving according to the virtual speech, that would be a mismatch should the cat be portrayed as speaking aloud. However, if the speech is portrayed as the cat's inner thoughts, then the dualistic division to match and mismatch would be softened. In *Sounds of Silence*, when the narrator's voice moved to a painted dot on the wall, an interpretation against the context of strict realism would be to call that a mismatch since painted dots do not speak. Yet, if one thinks that the narrator is an invisible ghost-like entity capable of attaching to anything, then the match/mismatch division gets blurry again.

One can argue that using mismatched sounds enable the content creator to make real-world objects something more or else than what they are. An old, unfunctional ship engine in a museum can be put back to life when adding a sound to it (e.g., Kaghat & Cubaud, 2010). Although the artificially engineered engine sound matches the museum item in principle, since the machine parts are not moving as they should when the machine is operating, the sound is functionally a mismatch. However, with the help of museum visitor's imagination the illusion may work, and the visitor can dip into the immersion where the engine is actually running.

Like some other techniques, Match and Mismatch are applicable to any other multi-sensory media, but I see them as very characteristic techniques to AAR due to their strong relation to the real world.

Having discussed attached and detached as well as matched and mismatched sounds, it may be useful to contemplate some qualities of AAR relating to these concepts. Let us imagine our story needs the cat mentioned above. In visual-only AR we could simply project a 3D image of the cat sitting on the coach. The user would believe the cat is there even if it did not produce any sounds since cats can be silent. However, in AAR we are limited to projecting a sound of the very same cat sitting on the coach. We can make the cat purr. Now, we would hear the sound of a cat purring coming from an empty coach. Is it a ghost cat? Or imagination at work? Creating an illusion with just acousmatic sounds becomes difficult when the other senses are registering the real-world conflicting with the intended auditory illusion.

However, if the story dictates the cat purring behind the coach, out of sight, the auditory version of the cat becomes potentially plausible since there would be no mismatch between the modalities.

The mismatch may cause a challenge the other way around, too. If in visual-only AR a virtual person is shouting but there is no sound, the scene must feel strange. However, the illusion of the virtual person coexisting there in the real world may be still work if we think the person is mute for a reason or another. The situation may be weird and even frightening, but the power of sight over other senses (Azuma, 1997, p. 367) may apply here, too.

In AAR, attaching the virtual augmented sound to a physical object may solve the illusion problem, like was done in both *Sound of Things* and *Sounds of Silenc*e. The object does not need to represent the source of the sound but can be anything, like the painted dot on a wall. Our imagination takes part of the play and attaches the sound and object together, 'magnetises' them (Chion, 1994, p. 70).

Interestingly, for the advantage of the audio, it is usually much easier to create an authentic acoustic reproduction of an object than model it photorealistically in 3D and animate it with all the necessary steps. Consequently, when the object is out of sight without any visual reference, but still audible, the illusion may potentially be more believable than when paired with a computer-generated image.

Additive enhancement

In this technique a real-world sound is additively enhanced or manipulated by the virtual audio display (VAD). For example, additive effects can be applied, such as reverberation or some other effect. Also, an additive sound can be overlayed on the existing one, mixing virtual sound with a real one. With this, new timbres and narrative meaning scan be created, for instance, on a perfectly running car engine sound one could add the sound of a squealing belt, breaking the sense of comfort and smooth driving experience. If the actual real-world sound cannot be used, a reproduced version of it can be used, trying to match the original as accurately as possible. A very simple additive enhancement using a reproduced sound would be an increase of the real-world sound's volume.

To add effects on real-world sounds, an external microphone setup can be used. For example, the headphone rig can be equipped with a microphone to capture user's footsteps and other sounds. The microphone signal can then be used to add room reverberations on the footsteps to match them with the other virtual audio content.

Masking

This other additive technique masks the real-world sound with a virtual sound. It is equivalent to 'painting' over in visual-based AR as described by Azuma (1997, p. 361). Since the technique is primarily applicable to AAR experiences using acoustically transparent headphones or direct augmentation, virtual sound needs to be louder or have other properties in order to mask the real sound. With acoustically isolated systems there is naturally no need to mask the real-world sounds, unless a mic-through system is used, and all of the surrounding sounds are passed to the user's ears. In that case the situation resembles the one with acoustically transparent systems.

In *Hyperkuulo* background music was used in some places to partially mask environmental sounds, likely to draw listener's attention to the narrative of the megacorporation represented by the music and the voice of the AI. However, since there was no control of the type of

headphones the listener was using nor the audio volume, the technique may have had different effect depending on the person.

Manipulation

Here, the real-world sound is manipulated, e.g., pitch-shifting is applied, or the sound is attenuated. This technique requires a VAD that is able to suppress the original sound and replace it with a manipulated one. For isolating individual sounds from the environment and manipulating them requires technology that is not yet available at least to public. However, during the time of writing this, Apple has filed a patent for an MR headset with an array of multiple microphones capable of 'directional audio detection', which enables isolating sound sources and rejecting ambient noise and reverberation. (*Two New Apple HMD Inventions*, 2021) With such system manipulation of individual sounds may become possible.

4.1.3. Examples of dynamic techniques

The techniques listed here are merely examples of some of the dynamic ways to use the above-mentioned techniques. Hence, the list is not trying to be comprehensive.

Changing between attached sound and detached sound

An example of this technique would be music that is first heard coming from a loudspeaker in the scene, but then 'expands' to the environment and starts to play from 'everywhere'. Analogous to this is a common technique in cinema where diegetic music changes to non-diegetic 'pit music' and vice versa (Chion 1994, p. 80).

Changing between matched and mismatched sound

An example of this could be seeing a person talk on a computer screen with lip-synced speech heard coming from the computer speakers; after a while the lip sync goes off-sync and the speech changes to more personal level, suggesting that we are now hearing the person's inner thoughts. Both sounds are still attached to the person on the screen, but their 'degree of match' changes.

Zooming

When zooming into a sound, the user is given an opportunity to hear a distant sound closer than it is, as if using 'audio binoculars' or a parabolic microphone. Technically, the effect can be

achieved by simply increasing the loudness of the sound while attenuating the reflections from the environment. Nearly the same effect can be achieved by moving the sound source closer in the 3D space.

In *Hyperkuulo* the hearing implant enabled the user to hear the imagined sounds inside a distant real-world sauna. However, being a head-locked experience, it was not possible to keep the direction of the zoomed sounds synchronised with their assumed source.

Zooming can be used as an additive enhancement technique: even with an acoustically transparent system the original sound could be masked with a louder version of the same—or nearly the same—sound.

4.1.4. Input methods

Input methods here refer to the ways for the user to interact with the experience, for example triggering various auditory and sequential events. Like the narrative techniques described above, the input methods listed below are not unique to 6DoF AAR as they are also available for other MR experiences, VR, or games. However, they are listed here as they are in my opinion characteristic to 6DoF AAR; being based on the utilisation of the positional tracking system already in place they may be the only input method if no other additional input devices are used.

Location

In this method the location of the user is used to launch events, e.g., using trigger zones. This method was used in *Sounds of Silence, Growl Patrol,* and *Hyperkuulo*. Also, user's pace, paths, time spent in stationary location, among others, can be measured and used as triggers.

Ray casting

Here, events are triggered based on detecting where the user's face is pointing at, suggesting the approximate direction of gaze. For example, a painting on a wall can start to talk when user is looking at it, regardless of where the user is standing. For more control a maximum distance or other conditions can be set.

Another application for the technique could be to detect whether another exhibition visitor is behind the user: when that happens, a whisper or some other sound can be played, attached to the visitor as if produced by her. When the user turns around, or the visitor walks to the user's

field-of-view, the sound playback could be prevented in order not to create a mismatch and risk breaking the illusion.

The examples above can be realised in the scene generator (e.g., game engine) using a ray casting method where a trace is projected forward from the face of the user's avatar: when the trace hits with the desired object an event is triggered. Ray casting can also be used to detect user's head movements such as nodding, although the head-orientation data can also be interpreted for the same purpose (e.g., Gampe, 2009).

Many other input methods could be used in addition to the ones based on positional tracking such as microphones to analyse speech or user-produced sounds, or buttons, sensors, etc, placed in the environment (Jacuzzi, 2018, pp. 6–7). However, since there is virtually no limit of the possibilities, they will not be included in here.

4.2. The prototype

This chapter will expose the process of designing and building a prototype of a 6DoF AAR experience. The prototype has been built in order to learn and understand the possibilities and limitations of the technology through practice. It has also served as a test bench for the identified narrative techniques.

In the prototyped AAR setup, the user is able to move in a room wearing open-back headphones and hear virtual audio objects superimposed three-dimensionally onto the physical real-world environment. According to the user's location and head-orientation, different narrative cues and interactions are applied.

For scoping reasons, I made the experience only for a single user, whereas the setup is modifiable to some extent to allow multiple simultaneous users. That would not only make it suitable for public use in museums, shopping centres, amusement parks, etc., but also open interesting possibilities for narrative interaction based on the users' movements relative to each other.

4.2.1. Early prototype: Invisible Voices

The spark for the project ignited when I experienced Holger Förterer's installation Sound of Things in Hamburg, June 2017. It felt magical how accurately the virtual sounds were spatialised and attached to the real-world items. With its intriguing feel and lyrical approach, the installation immediately triggered me to start developing narrative ideas for such a medium and wanting to build a similar setup of my own.

The following autumn, when I started my studies in the Sound in New Media master's degree programme at Aalto University, I began experimenting with a similar idea. My first prototype was based on a system using headphones equipped with an inertial measurement unit (IMU) to track the user's head-orientation, and two *Kinect* depth-sensing cameras by Microsoft to track the user's location. Since Kinect has a rather narrow angle of view, and the tracking reaches only to a few meters, I decided to try run two Kinects simultaneously to widen the trackable area.

The virtual scene was built in *Unity* game engine using the *DearVR* audio spatialiser plugin. Unity was chosen because I felt comfortable using a game engine as an authoring tool, and Unity was the only game engine supporting DearVR. I had tested other spatialiser plugins from Microsoft, Oculus and Steam, but in my opinion, DearVR was the only one that produced a

plausible spatial illusion. Soon after my own experiments, Laamanen (2018) in his master's thesis conducted a user study with 78 participants on six audio spatialiser where DearVR clearly outshined the others in its directional precision.

The content of this early prototype consisted of spatially positioned sound sources attached to real-world objects, such as a woman whispering a story under a table, a music box playing Tchaikovsky's *Swan lake*, a *Nokia Tune* ringtone emanating from a cellular phone, knocking on a door, distant chattering of people in a corner of the room, forest ambience behind the windows, etc. There was no narrative coherence or progress; the scene was only demonstrating the concept of creating auditory illusions with individual augmented sounds.

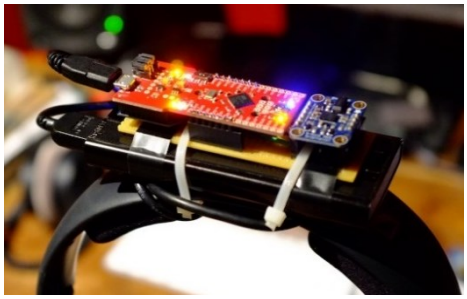

Figure 8: Tracking sensor system of the early prototype mounted on headphones.

In the prototype I used wireless open-back headphones *(Sennheiser HDR 120)* utilising analogue frequency-modulated (FM) radio with a very low latency but quite an audible background hiss and a narrow stereo image. However, the sound quality was good and clear. A small Arduino-compatible microcontroller board *(SparkFun Fio)* with an IMU *(Adafruit BNO055)* was mounted on the headphones (Figure 8). The board was connected to the computer with a radio link *(Digi Xbee)*. A small 5 V USB power bank was also attached to headphones, powering the board. The data from the two Kinects was construed in two separate computers each running an instance of Processing development environment with SimpleOpenNI library. Using Kinect's skeletal tracking method, the location of 'head' was read and converted to an Open Sound Control (OSC) messa‿

The OSC message from the 'slave' computer was sent to the 'host' computer over ethernet and combined in a machine-learning software Wekinator. From Wekinator the interpreted positional data in OSC format was sent to Unity. All in all, the setup was rather complicated.

I presented the prototype for fellow students on 7 December 2017 at Kallio Stage, Helsinki. The double-Kinect setup with its complexity was, however, quite unreliable, constantly losing the track of the user. Also, the empty 'black box' stage was not an ideal place to demo an AR concept since it was intentionally clean of every-day objects and functions, making it hard to get the experience 'grounded in reality'.

After the Kallio Stage demo I decided to develop the prototype further, however using a single Kinect. I prepared a new demo with a more interesting real-world environment for the MediaLab Demo Day on 19 December 2017 in Dipoli, Otaniemi (Figure 9; Figure 10). Around ten people tested the demo, and it worked surprisingly well, even though the trackable area was very small, and the tracking was still unreliable at times. Some virtual sounds turned out to feel very realistic and fascinated many of the testers, especially the whisper coming from between the rocks.

Figure 9: Visitor listening to virtual whispering appearing to come between the rocks.

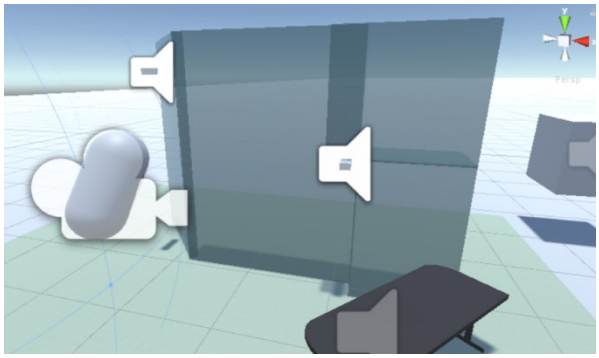

Figure 10: Unity scene with sound sources represented as loudspeaker symbols. Cylinder on the left is avatar with audio listener, representing the user's head and ears.

4.2.2. Developing the idea

During the following spring I kept developing the idea, hoping to be able to increase the trackable area and create more narrative and dynamic content for it. I had four design principles in mind:

1. The user should be able to experience as authentic illusion of added auditory reality as possible.
2. The experience should be personal without other people in the same space getting distracted. Also, the user should be able to adventure the experience at her own pace.
3. The user should be able to move freely in a room-sized area, preferably across multiple rooms, in a museum or similar venue.
4. The system should be doable by me as my master's thesis project using the available or acquirable resources.

System reliability, ease of setting up, or scalability were not major concerns for me yet at this prototyping stage, although they would become important issues later if outside collaborators and venues would be approached. Interactivity was also not a concern since I was confident that

once the tracking works well, scripting dynamically responsive events in game engine will be routine, at least from the technical point of view. Narratively, on the other hand, creating interactive content is very demanding and would have taken too much time to realise well in the scope of this thesis.

In the regard of the first principle, I wanted to create a true augmented reality experience so that the user will be able to keep hearing the real-world sounds as naturally as possible. For that I needed an acoustically transparent auditory display system, options being mainly open-back headphones and noise-cancelling headphones with a mic-through function. The open-back headphones seemed an easier choice for me, although they add a degree of colourisation to the external sounds due to the speaker element and other material distracting the sound waves. However, with open headphones the sense of surroundings stays unmediated and, at least in my opinion, feels more real compared to the reproduced soundscape of noise-cancelling headphones.

For the virtual sounds to appear as realistic as possible I needed clean and high-quality, close-perspective recordings of the source sounds, high-quality room simulation and binaural rendering system, professional-quality audio interface and transmission devices, high-quality headphones, and of course well-thought sound design choices. In terms of these requirements, I was rather comfortable since I was working professionally in sound design and audio technology.

The second principle of the experience being personal required the use of headphones or other personal auditory displays over, for example, direct augmentation with hidden speakers. With direct augmentation everyone in the space would hear the virtual sounds at the same time. Closed-back headphones with mic-through functionality would enable the sonic experience of the environment while limiting the audio leakage to outside. However, the leakage from open-back headphones is also quite minor, and that is reportedly the case also with acoustically fully transparent options such as Bose Frames (Carnoy, 2019) and bone-conduction headphones (J. Kim, 2021). Therefore, I regarded the choice of headphones more of a matter of acoustic experience and comfort than privacy.

The third principle about the size of the area with freedom to move around ruled out the Kinect approach due to its narrow and shallow field of view. I had also difficulties in getting reliable data out of the Kinects. Therefore, I started to look at other tracking methods based on various technologies.

The fourth principle about the project being doable by me with the available resources necessitated for 'artist-friendly' technology, i.e., thought-out, off-the-shelf systems that are easy to implement without deep technological knowledge and long development process.

During the spring and summer 2018 I also developed narrative ideas. I established a strong bearing to use augmented sounds as a hidden layer to the reality, revealing a new perspective to a common topic. A valuable help with the narrative ideating were conversations with Emilia Lehtinen, a writer and filmmaker with experience in audio books and other audio-only media. With her we came up with many ideas on how 6DoF AAR could be utilised narratively.

Virtual audio display

First, I considered using *HoloLens* by Microsoft as the VAD since it was an all-in-one solution with inside-out tracking and internal computer. With such a self-contained system there would be no need to install beacons, antennas or other fixed devices for tracking and audio feedback, an issue worth considering with e.g., historical venues. However, HoloLens had certain limitations, a major one being the visor potentially distracting and reducing immersion. It also turned out to be difficult to acquire one for testing. There was a *Meta2* helmet by Meta available at Aalto University with similar inside-out tracking features as in HoloLens. However, Meta2 lacks a built-in computer, and since I had no knowledge of whether it was possible to connect the device to an outboard computer wirelessly, I considered this headwear out of the question.

In late October 2018 I was in contact with a researcher at Microsoft who was working on spatial audio and HoloLens, and I asked for advice on how to borrow a HoloLens for this kind of project. He could not help with acquiring a device but gave valuable ideas for technical solutions. He also thought that HoloLens may actually not be ideal for this kind of application, mainly due to the visor distraction and short battery life. He also pointed out that while HoloLens is comfortable to wear, it takes a while to get used to it. He gave a few suggestions for alternative approaches such as other Microsoft MR headsets with custom modifications, the Kinect approach I had already tried, outside cameras tracking visible markers on the user, and strapping a wireless VR controller to user's head.

Following the email exchange, I abandoned HoloLens and continued exploring other solutions, taking account some of the suggestions. I investigated an approach using optical tracking with an infrared (IR) LED mounted on the headphones tracked by an array of IR cameras. An IMU attached to headphones would track the head-orientation. That would have been a variation of

Förterer's solution for 'Sound of Things'. The idea sounded definable although not very easy. I did some simple tests with LED tracking using a web camera and a simple tracking algorithm running in Processing. However, I quite soon realised such a setup would have required a lot of work just by acquiring right components. Knowing my limited programming skills and finite amount of time available I decided to look for a more ready solution. Also, I learned that IR-based systems are problematic in spaces with varying or bright lighting conditions. That would have radically limited the options for possible venues.

I also tested the idea of strapping a VR controller to headphones, using HTC Vive 2.0 (Figure 11). The Vive system uses IR laser beacons for accurate positional tracking, and with the Vive 2.0 update the reach was increased to 10 m x 10 m area according to the specs. In my test I managed to get very accurate positional data, but the tracking was extremely vulnerable when losing line-of-sight (LOS) making it quite unusable in a room with other people or equipped with furniture.

Figure 11: Testing with a Vive 2.0 controller mounted on headphones.

Pozyx Creator

I soon learned about Ultra-Wide Band (UWB) tracking technology, and a Belgian company Pozyx selling their UWB tracking kits aimed for creators. Based on the information on their website the solution seemed almost perfect for my prototyping purposes. According to them, with four fixed anchor units the *Creator* system could track up to five moving tags, reaching a 10 cm accuracy. Additional four anchors could be added for additional trackable area and/or accuracy. Conveniently, the tags were integrated with an IMU chip providing orientational information in the process without a need for extra hardware or coding, and these tags and anchors were compatible out-of-the-box with *Arduino* and *Raspberry Pi*.

Aalto University MediaLab kindly purchased a Pozyx Creator kit, and in January 2019 I got to start working with it. The data was first very bumpy, but that may have been due to inaccurate measurements of the anchor positions and/or interference from the structures of the building. I bought a laser distance meter and measured the anchor positions again, improving the data quality to some extent, however not making it as accurate as I was hoping (Figure 12). One of the fellow-students helped me with parsing and filtering the serial data coming from the Pozyx and going into Unity. Only much later had Pozyx added a notion to their website that six anchors

are recommended when for 3D positioning purposes; in the Creator kit there were only four anchors. Only at a quite late stage of the project, I found out how to use tags as anchors, thus being able to improve the performance to some extent using six anchors instead of four.

Figure 12: Accurately measuring the positions of Pozyx anchors and their antennas.

I also noticed that the orientation was drifting along time for some reason, but I could not find out why that happened. That was not an issue addressed at Pozyx website, so I suspected some interference inducted from the headphones, but did not research that further. My solution was to keep calibrating the tag often by placing the headphones in a known position and orientation every now and then. For possible public showcases I should, of course, solve the issue somehow. In the end, even with still a bit junky data the Pozyx and UWB seemed like the most potential technical solution for me.

Computer

I continued using Unity for authoring the scenes and DearVR as the spatialiser, but I still had to select the platform where the Unity build would be running. I considered several options, including (1) an outboard computer, either Mac or PC, with audio transmitted to the user over radio, (2) a small single-board computer (SBC) carried by the user, particularly Raspberry Pi since it was supported by Pozyx Creator, and (3) an Android or iOS device carried by the user. The options 2 and 3 would further open a possibility for a combination setup with a master computer working as a server with individual client devices carried by the user(s). This would serve multi-user needs where individual devices would get positional information on other users from the master computer.

With the SBC option there were at least two challenges: an SBC would have needed an operating system compatible with Unity, and it should have been equipped with a high-quality audio module with headphone output, making the setup more complicated and bigger in size. I was

also uncertain of their performance and audio latency, and as I had not much experience with SBCs, I abandoned that option.

I had also read about the long, over 200 millisecond audio latencies in Android devices, being totally unacceptable for an AAR experience. In iOS devices the latencies should have been much less, but as was the case with SBCs, I had little experience with developing for mobile devices, so I discarded that option, too. For instance, in in my preliminary tests with an Android build of a Unity scene I managed to output only monophonic audio.

Hence, the outboard computer option seemed most realistic and performant. The choice between operating systems was then determined mainly by the performance of their audio systems. On Windows, the audio playback of Unity uses the *MME* (Microsoft Multimedia Environment) driver protocol with high latencies reaching even 200 ms (*Low-Latency Multichannel Audio in Unity*, 2019). Such latencies potentially cause serious registration errors when the virtual sounds are spatially lagging behind user's head movements. To confirm the high latencies, I conducted a quick and simple test: I made a Unity executable build that played a test sound when a key was pressed. The key press was checked on every frame, so depending on the current frames-per-second (FPS) rate that may have added some extra delay before the sound was triggered. The DSP (digital signal processor) setting in Unity was set to 'Best Latency' since there was no way to accurately set the audio buffer size. Using a condenser microphone, I recorded the sound of the key press and the outputted sound at least three times for each test series. After that, I opened the recorded audio files in the *iZotope RX* audio editing software, and with the help of the spectrogram, I measured the elapsed time between the sound of the key press and the outputted test sound.

The test build in my 2020 *Lenovo Legion 7* laptop with *Windows 10* was running between ca. 500 and 1700 FPS, potentially delaying the audio triggering for less than 2 ms. The measured latencies between the key press and the audio output were 117–154 ms when using built-in speakers, and 147–157 ms with an external audio interface (*RME Fireface UCX*). Switching the DSP setting to 'Good Latency' the corresponding measured latencies were 139–140 and 155–172 ms.

For comparison, I did the test with my old MacBook Pro from 2013 with *macOS Mojave* running between ca. 45 and 70 FPS, adding an average of 15–22 ms delay before triggering the sound. The total end-to-end latencies were 46–71 ms without significant differences between 'Best Latency' and 'Good Latency' nor built-in speakers and external audio interface. Although the test

was very rudimentary, it suggested quite clearly that from these two environments the one using macOS operating system was the only one capable of acceptable playback latencies even with a much less powerful computer. I had not much experience on other operating systems such as Linux, and thus decided not to investigate into them.

Although I was first aiming at developing a single-user experience, I was also contemplating the multi-user option. With a single outboard computer there may have been a challenge with that: at least the Unity version I was using (2017.3.1) did not allow multiple audio outputs, nor it was possible to have multiple audio listeners inside the scene with separate spatialiser plugins. Hence, in a multi-user experience, it would have been impossible to feed individual virtual audio content to each user using a single Unity instance. Perhaps, with multiple Unity builds running simultaneously the problem could have been solved, but I left that exploration for the future.

In addition, I learned that if multiple tags are tracked in a Pozyx Creator environment, the update rate of 60 Hz is divided by the number of tags (*Scaling the Creator System*, 2021), hence increasing the latencies beyond acceptable levels.

To feed audio from the computer to the user's headphones a wireless transmission system was needed. In the *Invisible Voices* demo I had used a pair of wireless headphones, but for this prototype I wanted a solution with less background hiss, a better stereo image, and a longer reach. I also needed unnoticeable latency and good sound quality. Bluetooth headphones would have been one potential solution with their built-in wireless transmission capabilities. However, they tend to introduce a significant latency to the signal which, combined with the latency accumulated earlier in the system, would lead to unacceptable registration errors. Therefore, I decided to use conventional headphones connected to a wireless in-ear monitoring (IEM) system primarily used by performing musicians. In a wireless IEM system, the transmitter sends (stereo) audio typically over a license-free UHF radio band to the receiver unit, or 'belt-pack'. The transmitter can be equipped with external antenna equipment to enlarge the coverage. The belt-pack receiver is small and lightweight with usually a good-quality headphone amplifier. For my test I was using my own, rather affordable *LD System MEI 1000 G2* kit, but even with that the sound properties were good enough for at least the prototype use.

For headphones I decided to use my own pair of *Beyerdynamic DT 990 PRO*, which have an open-back structure and are comfortable to wear and lightweight with transparent sound quality. Their impedance is moderate, 250 ohms, working well with the IEM receiver. With mobile devices with less powerful headphone amplifier such impedance could potentially result

in decreased sound quality, so some other headphone model should be considered for such applications.

Later, I also tested the prototype with noise-cancelling headphones (*Sony WH-1000XM3*) using their mic-through functionality called 'Ambient Sound Mode'. Although the pseudoacoustically translated soundscape of the real world appeared rather authentic, the spatial directionality felt sometimes odd especially with near-field sounds. Also, being closed-back headphones, wearing them felt encapsulating and tiring after a while, and they enhanced bodily sounds more than open-back headphones, all these factors potentially reducing immersion. Current prototype solution

4.2.3. Current prototype

In February 2019 I managed to set up a full working prototype in my workshop. The prototype was built using the following components (Figure 13):

- *Pozyx* UWB tracking system with
    - Four anchors (increased to six in 2021) placed on the walls at different heights
    - Tag attached to the headphones, powered by a small 5 V power bank (Figure 14)
    - Another tag working as the master tag, connected to an *Arduino Uno*
    - *Arduino Uno* microcontroller hosting the master tag
        - Sending positional data to computer (serial data)
- *MacBook Pro* computer with
    - *Unity 2017.3.1* game engine / executable build with *DearVR* spatialiser plugin
- *RME Fireface UCX* audio interface
- *LD System MEI 1000 G2* IEM transmitter and belt-pack receiver
- Beyerdynamic DT 990 PRO open-back headphones

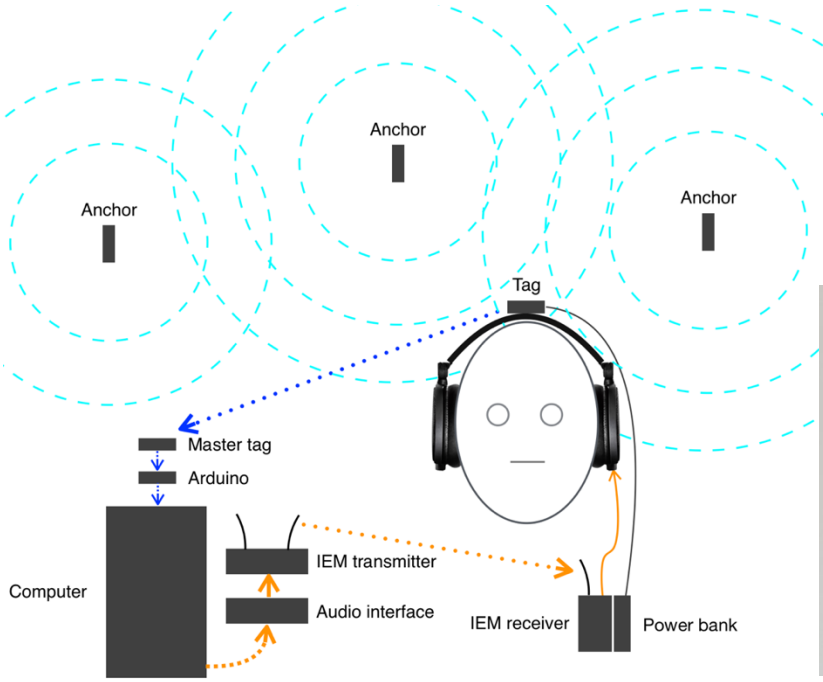

Figure 9: Current prototype setup using Pozyx Creator UWB system for positional tracking and wireless IEM system for audio transmission.

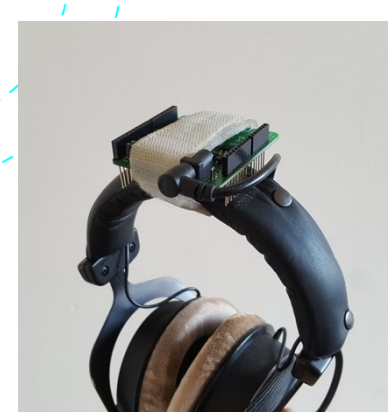

Figure 14: Pozyx tag mounted on headphones.

The power bank and IEM receiver were carried by the user in a waist bag. This prevented any unnecessary weight to be added on the headphones, something that would have potentially reduced the comfortability and thus immersion. Pozyx tag weights only 12 grams and its mounting foam and the power cable some more. However, the Pozyx tag is rather big in size

since it is made primarily for prototyping and connecting directly to Arduino Uno and Raspberry Pi as a shield. For this setup that was not a problem, but if public experiences would be planned, then it may be necessary to consider smaller tags.

Since the locations of the anchors need to be measured as accurately as possible, I used a laser distance meter to measure their heights from floor as well as distances from each other. The shape of my workshop is very irregular, so I used GeoGebra software to triangulate the coordinates of the anchors based on their distances between each other (Figure 15). My estimation is I managed to get to the accuracy of about +/- 2 cm.

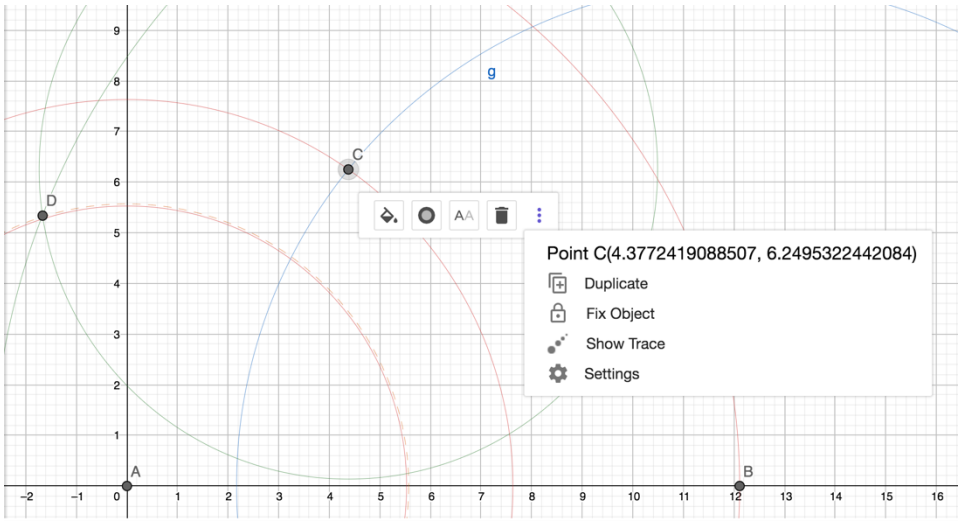

Figure 15: Calculations of anchor positions in GeoGebra.

It was also necessary to map the room with its key furniture and fixtures in Unity in order to place virtual objects to correct places in relation to the real environment (Figure 16). For that I also used the laser meter, but since the need for accuracy was not as high as with the anchor locations, I did the mapping a bit more roughly. With a regular boxed-shaped room the process would have been simpler, consequently resulting in better accuracy.

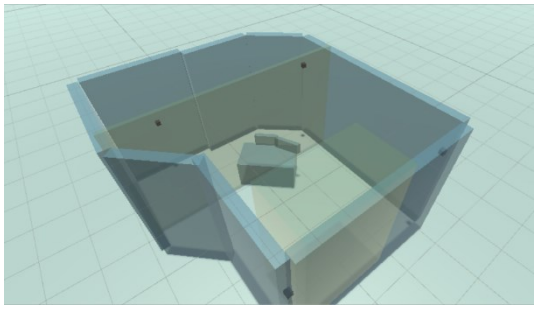

Figure 16: Room geometry and furniture roughly mapped in Unity.

In Unity I made a simple C# script that read the positional data coming from the Arduino/Pozyx receiver as serial data and applied simple smoothing to it. The location data was quite bouncy, and I used a Vector3.SmoothDamp function to average the data with the expense of adding some delay. However, since I expected users moving quite slowly in an experience like this, I did not see that as a problem. For orientation data smoothing a Quaternion.Slerp function was used respectively, however with fast interpolation to reduce registration errors when turning head rapidly.

Unlike in the previous *Invisible Voices* demo, now I also compensated the offset distance between the head-tracker and ears, something Azuma (1997) lists as one possible cause of the registration errors. Originally in my design, the positional data received from the Pozyx tag simply moved and rotated the centre point of the avatar's head. Since that was incorrect behaviour, I now made an own game object for the tracker on top of the head and assigned the rest of the head as its child object. I also set the audio listener approximately 13 cm below the head-tracker, equivalent to the average real-world distance from the tracker to ears.

4.2.4. Sound design

Between February and July 2019, I created most of the demo scenes and tested the setup. In spring 2021 I did some improvements to the scenes. Since I had already many narrative ideas and techniques in mind, I went straight into realising some of those ideas without a very methodological approach. For the sound content I used my own recordings as well as professional sound effects libraries from *Epic Stock Media, Finnolia Productions, Soundsnap, Soundly,* and others. For the most cases I aimed for high-quality, close perspective recordings without any room acoustics to be able to use them as spatialised point-source sounds and add a required amount of room simulation to them. However, if the sound already contained a slight natural room reverberation component, it still worked nicely in most cases since the acoustic properties matched the real-world space.

I edited the files in *Pro Tools*, a digital audio workstation (DAW). I trimmed them, made them looping if necessary, and added some simple high-pass filtering (HPF). However, I did not want to process them much, especially dynamically, in order to keep them as natural as possible. Since DearVR uses stereo files, I bounced the clips as 16 Bit, 48 KHz stereo wave files and imported them into Unity.

For acted scenes I first performed 'placeholder' voices by myself in Finnish. Later, in spring 2021, I hired professional actors to perform the voices in English. They recorded their short lines at their home studios with multiple takes and delivered the files to me via *WeTransfe*r or *Dropbox*. One scene with the character talking in video I acted by myself since I did not have time to arrange and brief an actor, let alone arrange a shooting session.

The music in the *Music Box and Immersive Orchestra* scene was my own composition made earlier for a video game. The instrumentation in the piece is orchestral, so I first made a simple arrangement for a music box, using a virtual instrument in Pro Tools. After that, I came back to the original orchestral version and split the same passage to multiple stems containing either individual instruments or a whole section such as vibraphone, cellos, violas, solo horn, woodwinds, gran cassa, etc. I kept the music clips as 'dry' as possible, i.e., without reverberation, so that I was able to control their perceived virtual space with DearVR.

In Unity, for each sound I created a game object with a DearVR component. To reduce processor stress, I did not let DearVR to simulate reflections by the modelled room, but instead utilised generic room simulation models of DearVR. The illusion that the virtual sounds were in the real-world space was still quite good even though the reflections from the virtual sounds did not match the real-world space. In the *First Page* scene the sounds were supposed to be outdoors and hence in another space than the reality, so naturally a different spatial model was used.

A lot of effort went to adjusting the attenuation curves in order to obtain a desired effect. Although the default roll-off curves in DearVR and Unity may have been based on acoustic models of the real world, they needed to be adjusted to get them appear authentic. Also, it was often needed to manipulate the attenuation so that the sound attenuated faster when moving further than it would have been in real life. That was necessary, for example, with the music box due to its piercing sound quality. Also, because the installation room was quite small, manipulating the roll-off curves made it possible for the user to get 'further' from the sounds that what the actual physical distance was. Since human hearing is not very good in perceiving the distance of a sound source, and relies on sight and other cues when available (Xie, 2013, p. 19), that may have been the reason the trick worked.

Another issue with sound levels was the overall listening volume. In *Sounds of Silence* as well as *Hyperkuulo* the users could adjust the audio volume to their taste, but I felt that in my acoustically transparent experience there would be a risk that the virtual sounds were adjusted either too soft or too loud compared to the real-world sounds, thus breaking the illusion. In

Sound of Things the users did not have a control over the volume. I decided that was my approach, too, and I placed the IEM receiver belt-pack in its bag so that the rotary volume knob does not get turned during the experience. The only concern I had with fixed audio volume was whether some sounds would be experienced being too loud, but at least in my demo scenes all the sounds were either soft or medium loud. The loudest sounds the user would experience would emanate from the real world through the open-back headphones.

Besides pure sound design tasks, I also scripted interactive behaviours in Unity with C# based on user walking into trigger boxes and detecting ray trace collisions for estimating direction of gaze.

4.3. Demo scenes

For the prototyped setup, I made several small demos to 1) test the usability of the identified narrative techniques and 2) assess the system's narrative capabilities. The scenes were realised during the spring 2019 and improved after a two-year break in spring 2021. My original plan was to install the scenes in a public space, the latest idea being a co-working space at the Aalto University campus. However, due to the COVID-19 pandemic and other restrictions at the campus, I finally built the prototype in my new workshop. Unlike the previous one, the room is rectangular but quite small. Poor sound isolation causes occasional auditory disturbances from adjacent spaces. However, the space is otherwise very peaceful and private, thus fitting nicely for the testing purposes.

The demo scenes utilise a number of narrative techniques defined earlier, being an unstructured test whether the techniques work in practice, what issues there may be in using them, and what effects they manage to evoke in the user. Realising the scenes was also a good exercise and experiment on how to create plausible auditory illusions with a 6DoF AAR system.

4.3.1. Knocking on the Door

User hears knocks on the door at the other end of the room (Figure 17)*. A woman's voice starts to talk behind the door: 'Hey, excuse me. Hello? Can you hear me? Is there someone there? I hear you.'*

More knocks on the door. 'I need you to open the door. Can you, please open the door?' If the user stays still, after a while the woman repeats her message with slightly different wordings.

Figure 17: User hearing virtual knocking on the door.

If the user decides to walk away from the door, the voice gets a bit panicked: 'No, please, come back! Hello? Please, don't go. I hear someone out there! I need your help.' If the user approaches the door, a male voice can be heard behind the woman, grunting: 'Is there someone you're talking to?' The woman panics a bit more and lowers her voice: 'Oh shit, let's be quiet. Look, the door is locked, okay. I can't get out. I need your help. I need you to get the key from the table. Should be right behind you. The table behind you. Please, hurry up."

If the user goes to the table for the key, the woman celebrates: "Yes, thank you! You found it! Just, come here quickly! Thank you!" The scene ends.

Techniques used:
- Attachment
- Out-of-reach
- Match
- Input method: Location

This demonstration tests the system's ability to create an auditory illusion of someone being behind the door and evoke a sense of immersion through suspension of disbelief. If the illusion is experienced as plausible, the character's cry for help and reactions to the user's movements could potentially resonate emotionally within the user. Door is a strong symbol in storytelling and often used as a liminal space 'between inciting incident - - - and the protagonist's resolution' (*Doorways in Literature and Film*, 2018). The fact that the user cannot see what is behind the door but relies on auditory information potentially emphasises the excitement and tension of the situation.

Being an interactive scene there is a risk that the user does not react to the voice in any ways and stays still. However, since the purpose of the scene is to test the identified narrative techniques, and not tackle the challenges of interactivity in storytelling per se, one can assume the user to 'play along'. The same applies to the other interactive scenes in this study, too.

From the point of view of the techniques, the knocking and voice are attached to the door, although the sound sources appear as being right *behind* the door rather than *on* the door. However, from the sound designer's perspective the door is a natural real-world object to attach the sounds onto: it functions as an acoustic portal, funnelling sounds to the room. Technically, in the game engine, it would be possible to position the sounds in the space behind the door and let the engine or spatialiser plugin to calculate the propagation and attenuation of the sound waves through the closed door and finally to the user. However, since the perceptional experience would be the same—the sounds emanating from the door—the extra trouble may not be worth in this case.

The sounds match the real world: it is plausible that someone is knocking and talking behind the door, and since the door blocks the view to the assumed person there is no risk of mismatch between the sounds and the primary source of the sounds. The sounds are also out of reach for the user. It is not possible to go around and observe the sounds from different angles.

The interaction is based on the user entering and exiting trigger boxes, hence the input method is location tracking.

The attachment technique relies on the accuracy of the positional tracking, low end-to-end delay and plausible externalisation of the spatial audio render. Therefore, registration errors in these areas can potentially break the illusion.

Also, voice acting is in a key role in creating the feel of realism, and for that reason I asked the voice performance from a professional actor with background in improvisational theatre and, in my opinion, a very realistic acting style.

4.3.2. Virtual Ambience

Normal environmental ambience of the room is augmented with realistic, everyday atmospheric and spot sounds, such as water running from a tap—the tap is closed in real world—, someone tuning a piano in another room behind walls, and renovation work undergoing in the room above.

Techniques used:
- Attachment & Detachment/Acousmêtre
- Match & Mismatch
- Out-of-reach & Within-reach

All of the sounds are attached to a real-world object, a wall, ceiling, water tap. Again, as in the *Knocking on the Door* scene, if the sound sources would be real, the original sources would exist further away than the wall or the ceiling. However, one can argue that the walls and the ceiling are actually producing the sound perceived in the room by resonating and modulating the original sound waves. Hence, attachment of the sounds to these structural elements is reasoned both conceptually and practically.

The sounds in the scene are also matching, similarly as in the Door scene. They are also made as realistic as possible to enhance the plausibility of the illusion. Only the water sound can be considered as a mismatch since the tap is closed and no water is pouring out in the real life. One can even say that the water sound is acousmatic since its source (the water) does not exist, even though the tap does. To analyse further, if we break the water sound into components, the sound of the water running inside the pipes of the faucet is attached, whereas the sound of the water pouring out and splashing on the sink is acousmatic.

Augmented ambiences could be used, for instance, in a historical home museum suggesting the soundscape of the past. However, with an acoustically transparent auditory display, the contemporary real-world sounds such as cars and electronic sounds, could hinder the experience.

4.3.3. The First Page

Figure 18: User looking at the open book.

A book of Hemingway's 'For Whom the Bell Tolls' is placed on a shelf, its first page open. When the user gets close to it and looks at it for a few seconds (Figure 18)*, the room transforms acoustically to the opening scene of the book: Wind is humming on the treetops of the pine forest surrounding the user, birds are singing and a squirrel chirping, a brook burbles close by, and in a distance, water rushes from a dam. Two men start to talk, invisible to the user; one's sound coming from the floor as if he was lying, the other one's sound emanating a bit higher next to him. Rustle of paper can be heard when the lying man unfolds a map. After a moment the scene collapses back to the book, the sound sources moving fast through the room and into the book after which they disappear.*

Techniques used:
- Detachment/Acousmêtre
- Out-of-reach & Within-reach
- Input method: Ray casting

All the sounds are acousmatic, and whereas some are out of reach, some can be walked around such as the birds, squirrel and the two characters. The soundscape begins when the user has been close to the book and stared at it for three seconds: both location and direction-of-gaze from the head-orientation data are used for triggering the scene.

Translating the opening of a novel into a three-dimensional soundscape potentially highlights the author's intentions to accommodate and transport the reader into the storyworld (Herman, 2009, pp. 112–118). However, it seems that not many writers start with such strong situated descriptions of the space and events; in other cases, sonification would be challenging. The purpose of this demo was to test how some of the spatial auditory and interactive techniques of 6DoF AAR could be used in that translation process by using the beginning of Hemingway novel as an example. However, I would estimate it to be more fruitful to create original narratives for 6DoF AAR instead, in order to utilise the unique characteristics of the medium.

In a way the scene fights against the first characteristic of AR as combining real and virtual (Azuma, 1997, p. 356). The virtual auditory narrative and the individual sounds have little or nothing to do with the real physical environment. The scene has no relationship with the space, and in that sense could be installed anywhere else, too. In fact, the furniture and other items in the room potentially disrupt the immersion rather than create interplay with it. However, as a narrative technique such displacement could work for memory and dream sequences, suggesting an internal soundscape or a sonic hologram attached to the environment. Further, through such a juxtaposition the user may start forming her own connections between the real world and the auditory narrative through a somewhat similar mechanism than with 'associative montage' in cinema or metaphors in literature and poetry (Bordwell & Thompson, 1997, pp. 154–155; Kuhn, 1985, p. 219).

4.3.4. Music Box and Immersive Orchestra

User can hear a music box play a song. The box is placed on a table in the middle of the room. If the user walks closer, the music gets louder as it naturally would. When the user after this moves further from the table, an orchestral version of the same music gets mixed with the music box

version, gradually replacing it the further the user moves. The orchestral version is spatialised three-dimensionally around the user as if the user was inside the orchestra. The single sound sources are, however, outside of the room's real walls and thus out-of-reach. The two versions of the piece are running in sync, cross-faded whenever the user moves between the two areas.

Techniques used:
- Change between attached and detached/acousmatic sounds
- Match
- Out-of-reach & Within-reach
- Input method: Location

The scene tests transition from attached sound (music box) to acousmatic sounds (orchestra). The crossfade is based on the user's distance to the music box, and due to the small size of the available space in the room, the crossfade happens quite quickly between the two states.

The scene also tests how temporally synchronised sounds (orchestral stems) around the user can be used to create the illusion of being inside an entity, in this case an orchestra or a one huge musical instrument.

The orchestral sound sources are out-of-reach, but they could have been located within-reach as well. However, since they are not playing back the sounds of single instruments, but rather stems of instrumental sections (e.g., five cellos, three trumpets, etc.), it would have been challenging to maintain the illusion when letting the user to hear multiple instruments emanating from an invisible point source floating in the air.

4.3.5. Influencer's Inner Voice

A computer screen sits on a table in the middle of the room. On the screen the user can see and hear a person talking; he's a social media influencer, talking fluently and apparently enjoying it. If the user walks behind the screen, the talk on the screen gets attenuated and the influencer's inner voice surfaces, explaining about his performance anxiety he suffers from but can control. The inner voice is still emanating from the screen, however its frequency spectrum slightly altered as if affected by the back structure of the screen. When the user walks back to the front of the screen, the screen talk takes over again, and the inner thoughts disappear.

Techniques used:
- Change between matched and mismatched sounds

- Within-reach
- Input method: Location

This scene tests the ability of the medium to create layers of realities on top of each other: the video on the screen is 'real', visible to everyone in the room, and the lip-synced talk is potentially perceived as real for the user wearing the VAD. Then, the inner thoughts suggest another layer of reality, 'true' in a sense that one cannot deny someone's thoughts existing, and 'virtual' in a sense that in real life it would be impossible to hear them. These two realities coexist simultaneously, underlined by the attenuated screen talk being heard in the background of the inner voice. The sound design choice of mixing the two voices together may hinder the plausibility of the experience. On the other hand, the use of inner speech over attenuated real speech is an occasionally used technique in sound design and thus codified into the tradition of sonic narration. Maybe, therefore, the user will not be distracted by that, and the suspension of disbelief is maintained.

4.4. Observations

I was personally rather satisfied with how the demo scenes performed in testing the chosen narrative techniques. My impression is that the way I used the techniques in the demos indicates their usefulness in other narrative contexts, too. Also, the list of techniques, as well as the process to identify them, turned out to be valuable sources of inspiration when designing the scenes. They also worked as a 'reality check' on many occasions. I often stopped and thought: 'What if I applied that technique, what effect would it make? What kind of narrative situations could it serve? Would it be possible to utilise the technique in practice?'

I also see that the techniques are dissimilar and clearly defined enough to be useful. When working on the scenes, it was easy to compare the techniques and choose between them when attempting to gain a certain narrative effect. For example, when designing the *Influencer's Inner Voice* scene, I was first struggling to decide how the inner voice part should be realised. I had a general, vague idea in my mind but could not outline it clearly in practical terms. Once I had conceptualised the techniques, it was easy to go through them and choose the ones that would best convey the effect I was looking for.

A number of techniques were left untested due to the limited time resources, including Spatial offset, Spatial asynchronisation, Additive enhancement, Masking, Manipulation, and Zooming. There will hopefully be another occasion to assess them in practice.

Albeit some technical challenges the prototype turned out to be able to create plausible auditory illusions. Six people visited my workshop during the spring and summer 2019 and informally tested some of the prototyped scenes. Two more people informally tested the improved scenes in April 2021. Based on my own experience as well as preliminary oral feedback from the visitors, it seems that through the use of the tested narrative techniques, the prototype is capable of immersing the user into a storyworld and evoking emotions.

Even though the scenes were not a part of a coherent story but merely technical demonstrations of the narrative techniques, they helped me to understand the capabilities and challenges of the medium. Firstly, the demo scenes exposed several technical complications causing registration errors and thus potentially undermining the desired illusion. One problem was that the location data was jumping to some extent, especially near the edges of the 'play area'. The bouncy location data caused some issues, for instance, sounds moving away from their intended positions. Especially acousmatic sounds were afflicted by this since it became hard to perceive where the exact location of the sound source was. I felt that acousmatic sounds are also narratively rather challenging to use since their relationship with the surrounding reality needs to be justified. With sounds attached to physical real-world objects the jumping location tracking was not that problematic, presumably thanks to the 'magnetising' effect (Chion, 1994). The tracking errors would also have an effect on the input methods, particularly ray casting, due to the trace potentially jumping away from the trigger area.

The tracking challenges raised a question whether it is narratively necessary to track all three dimensions or could the vertical axis be omitted like was done in *Sound of Silence*. It is assumed that 2D location tracking is potentially more reliable than 3D tracking (*What Is the Accuracy of UWB?*, 2020). By nature, people seem to move along horizontal axis more than vertically, and even if sounds were added, for instance, below the ear level, not necessarily many of the users would bend down to take a closer listen to them. In *The First Page* scene the acousmatic characters are positioned low in elevation, one is lying and the other one is taking a knee. If the vertical directionality in the binaural spatialisation worked well enough, the user would perceive the characters being below her ear level even from a stationary position, especially if the user moved her head for more spatial cues. However, as was the case with *Sounds of Things*, properly carried out vertical tracking can enhance the illusion of the sound objects existing in the environment. It might also be inspirational for the storytellers to design scenes that take use of varying vertical positions of sounds.

Besides the tracking errors, it became obvious that the anchors and the UWB system itself induce some practical considerations. For instance, the anchors need to be installed on walls and possibly ceiling, and they require electric supply by either cables or batteries. These installation requirements may be an issue in some places such as historical venues. Also, even though the anchors are not large items, they, together with electric cables or battery units, are still a visual reminder of the technology mediating the experience, potentially hindering immersion.

One problem was the end-to-end system delay which was noticeable especially when turning head: sounds did not align to the new orientation quickly enough. This was very apparent especially with 'within-reach' sounds that could be circled around and observed from different angles. However, by using a Mac computer instead of Windows and optimising the Unity project and DearVR settings for minimal processor impact the latencies were still tolerable when turning head slowly.

An important notion concerns environmental real-world sounds. One of my design principles was to let the user experience as authentic illusion of added auditory reality as possible. To accomplish that I decided to use open-back headphones with are acoustically nearly transparent. That enabled a natural perception of the real-world sounds potentially enhancing the illusion of the real and virtual auditory realities coexisting with each other. However, the acoustic transparency also means that the experience is vulnerable to any unexpected and disruptive real-world sounds. That sets a huge challenge for the narrative design since there is no control over the real-world soundscape and hence the story must sustain any unforeseen external sounds. In my case, I had to interrupt my tests many times when a rock band started to rehearse downstairs, someone started a lively video chat behind the wall, or a dumpster of recycled glass was emptied to the truck outside of the window. With visual-based AR the problem is potentially milder, since, at least in indoors, it is often easy to control the visual stimuli in the room, and if a disrupting visual event appears, the user can often look away and continue the experience. However, sounds penetrate through and around walls, unless the room is specially treated, and there is no way to 'hear-away' and ignore unwanted sounds (Sarter, 2006, p. 441; Kolarik et al., 2016, p. 373). If the story cannot sustain unwanted sounds, then the venue should be auditorily predictable, or the experience should be constructed with a pseudo-acoustic VAD blocking the outside sounds and possibly regenerating or selectively passing through the desired environmental sounds.

One element I ended up being slightly disappointed with was the quality of binaural externalisation. Although the DearVR spatialiser did extremely good job with its generic HRTF's, the effect was still not as perfect as I would have dreamed. Often it also felt that the sounds were too high in elevation. It might be that I had developed a critical ear for the matter and paying more attention to it than someone experiencing the prototype for the first-time. It is left for further research whether users will ignore such imperfections when concentrating on the narrative and interaction, and getting used to the way of representation (Blauert, 1997, p. 374; Lindau & Weinzierl, 2012, p. 804; Chion, 1994, p. 107).

A positive notion was that the low-priced IEM system I was using worked better than I expected, and the background noise level was almost inaudible. In my tests I did not get any radio frequency (RF) interference, something I have experienced multiple times when using the system on music performances. Hence, for public experiences it may be necessary to consider using equipment with better RF qualities or a digital system with error correction and other techniques to avoid disturbances in the signal.

Despite of all the technical challenges, this project has showed that with basic understanding of programming and technology it is rather easy to build a functional 6DoF AAR system capable of plausible acoustic illusions and storytelling. The reason for that is largely the current availability of easy-to-use component kits, authoring tools and other software.

5. DISCUSSION

One can safely argue that 6DoF AAR has potential for immersive storytelling. The related experiences discussed in this thesis, *Sound of Things, Sounds of Silence, Growl Patrol* and *Hyperkuulo*, demonstrate how spatially registered AAR can create almost magical acoustic virtual realities coexisting with the real world, and how this interplay between real and virtual can be used in narratively fascinating ways. Still, at the same time, the medium enables situation awareness due to keeping the visual and other sensory modalities intact.

This thesis suggests that 6DoF AAR has its own characteristic techniques for storytelling and storyworld-creation, and thus it has potential for unique narrative content, not possible to realise with any other medium. Also, this study suggests that by using off-the-shelf components and easily available authoring tools, anyone with knowledge on sound design, programming and storytelling can create gripping immersive 6DoF AAR experiences, although the technological approach should be carefully judged and chosen.

One area of assessment is the most optimal tracking technology for 6DoF AAR. The tested UWB solution has a lot of potential and been successfully used in, for instance, *Sound of Silence* with the *Usomo* system. However, the anchors, cables and other infrastructure required for a UWB setup may be difficult to install in some venues and they can potentially undermine the immersion. Also, in my prototype the registration errors were an issue, although the challenges are inevitably solvable with more development work and appropriate equipment. An alternative technology worth exploring could be image-based inside-out tracking where built-in cameras and machine vision deduce the device's location and orientation. Since there would be no need for fixed beacons and other installations, the approach could be extremely convenient for public spaces and delicate environments while enabling improved immersion without disruptive technology at sight.

The demonstrative scenes in this thesis also raised another question potentially worth investigating further: When a 6DoF AAR experience relies on evoking emotions and maintaining a strong feeling of presence, what happens if the system delays and unreliable tracking start causing registration errors? Will there be a risk that the user gets drawn too violently from the storyworld when the immersion shatters due to technical errors? Or will the emotional commitment to the story carry over any technical faults? Due to my research setting without outside testers the questions were left unanswered, but this may be something worth exploring in the upcoming experiences.

The focus of this study has been greatly in the spatial auditory nature of the medium and its capability of creating auditory illusions. Therefore, the identified narrative techniques reflect that viewpoint. However, there are undoubtedly many other ways to approach the narrative means of 6DoF AAR. One interesting direction would be to create a large-scale story for the medium—possibly using fewer techniques but concentrating on the story, characters, environment, and interactivity—and explore the narrative capabilities in a case study. That would likely bring up multiple issues not yet tackled in this thesis.

To suggest a pool for story ideas, 6DoF AAR could have a lot to offer in challenging the dominant narratives of historical venues and events. Such a story could be set, for instance, in the home museum of the national poet of Finland, Johan Ludvig Runeberg. When putting on the headphones, the visitor could experience the house through the ears of Runeberg's wife, Fredrika. In addition to being a wife and a mother of eight children, she was also a prominent novelist and journalist, which is sometimes forgotten. It would be an intriguing contrast to acoustically overlay Fredrika's busy and productive life with the context of her husband's established narrative, represented by the stately and nowadays rather quiet museum settings. Consequently, with more controversary narratives the effect would probably be much stronger.

Another direction could be concentrating on the multimodality of the experience from the user's perspective; for instance, how the perception and interpretation of the auditory content is influenced by the sensory stimuli of the real-world (see e.g., Nanay, 2012). Also, the problem of acoustic transparency and augmentation of auditory reality would be worth researching more from the storyteller's point of view: since hearing is omnidirectional, and environments are full of distracting sounds, it seems difficult to frame the user's attention in hear-through AAR. A fundamental question might be, how can the user get immersed into a storyworld while constantly being reminded of the real world?

Limitations of the study

The methods in identifying the narrative techniques proposed in this thesis can be reviewed. The analysis behind the identification process was based on a small number of related experiences, and the amount and quality of theoretical concepts studied have been limited. Also, the way I selected the analysed material and how I interpreted them may have been biased. However, even though the identified techniques are, in my opinion, characteristic to 6DoF AAR, they are not unique but for many parts quite universal and applying concepts from other media. Therefore, I conceive the risk of the techniques being unusable very low. That is supported by

the use of demo scenes where some of the techniques were tested: the techniques turned out to be a useful tool set at least to me. At the same time, I am aware that there can be other ways to categorise the techniques, there may be some essential techniques missing, or there may be too much overlap with the techniques of other media.

When evaluating the prototype building process and its capability of testing the identified narrative techniques, I believe that anyone else going through the same process with the same resources would have come up with comparable results. The analysis of the prototype design and development process was rather straightforward, in my opinion, although I do not deny that my personal interests and limitations may have filtered and tilted the reporting. To confirm the observations in this thesis, further studies with different 6DoF AAR setups and scenes would be beneficial.

Without user tests it is impossible to make conclusions on how other people would experience the effectiveness of the techniques and the performance of the prototype. Therefore, this perspective was left out of this study, but would definitely be worth exploring to get a deeper understanding of the medium and its narrative mechanisms and capabilities.

6. CONCLUSION

At the time of writing this, narrative and social audio are trending. Audio books and podcasts are very popular, and in the wake of the audio-chat application *Clubhouse,* audio-only features are being introduced in other social media platforms. Also, perhaps as a counterweight to the hectic everyday life, people seem to appreciate immersion when enjoying art and entertainment. The role of spatial audio is getting attention especially in video games and virtual experiences, and new 3D audio technologies are being developed. Moreover, stories always fascinate people. On that account, the current atmosphere seems encouraging in terms of narrative and immersive AAR, suggesting it could be warmly received and enjoyed by the audience.

This thesis has proposed a set of narrative techniques that would be characteristic to 6DoF AAR. They will hopefully be useful for anyone interested in creating content for the medium. They may also serve as a conversation starter when the narrative language of 6DoF AAR is being shaped and evolved.

This study has also described the process of designing and realising a prototype setup for 6DoF AAR experiences. The observations from the process will hopefully, too, help someone in their efforts in assembling a similar setup and evaluating possible technical approaches. The observations will also offer a viewpoint to some opportunities and challenges in creating narrative content for the medium.

While waiting for someone to come up with a more captivating name for the medium, 6DoF AAR will hopefully find its place amongst the intriguing company of mixed reality experiences with its unique narrative possibilities.

7. REFERENCES

3D Audio Market—Global Industry Analysis 2018—2026. (2018). Retrieved 23 February 2019, from https://www.transparencymarketresearch.com/3d-audio-market.html

3D localization technology | Optical tracking explained | 6DOF. (2019). PS-Tech. Retrieved 27 April 2021, from https://www.ps-tech.com/optical-tracking-explained/

45+ Literary Devices and Terms Every Writer Should Know. (2020, March 20). Reedsy. Retrieved 18 April 2021, from https://blog.reedsy.com/literary-devices/

Ahmad, N., Ghazilla, R. A. R., Khairi, N. M., & Kasi, V. (2013). Reviews on Various Inertial Measurement Unit (IMU) Sensor Applications. *International Journal of Signal Processing Systems*, 256–262. https://doi.org/10.12720/ijsps.1.2.256-262

AirPods Max. (2020). Retrieved 26 April 2021, from Apple. https://www.apple.com/airpods-max/

Akeroyd, M. A. (2006). The psychoacoustics of binaural hearing. *International Journal of Audiology*, *45*(sup1), 25–33. https://doi.org/10.1080/14992020600782626

Albrecht, R. (2016). *Methods and applications of mobile audio augmented reality*. Aalto University. https://aaltodoc.aalto.fi:443/handle/123456789/21198

alps—Acoustic localisation positioning system. (2018). Retrieved 27 April 2021, from https://alps.fi/

Arth, C., Grasset, R., Gruber, L., Langlotz, T., Mulloni, A., & Wagner, D. (2015). The History of Mobile Augmented Reality. http://arxiv.org/abs/1505.01319

Audeze Mobius Headphones. (2021). Audeze LLC. Retrieved 26 April 2021, from https://www.audeze.com/products/mobius

Aural ID—Genelec.com. (2020). Retrieved 8 April 2021, from https://www.genelec.com/aural-id

Azuma, R. T. (1997). A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, *6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

Bederson, B. B. (1995). Audio augmented reality: A prototype automated tour guide. *Conference Companion on Human Factors in Computing Systems - CHI '95*, 210–211. https://doi.org/10.1145/223355.223526

Blauert, J. (1997). Spatial hearing: The psychophysics of human sound localization (Rev. ed). MIT Press.

Boletsis, C., & Chasanidou, D. (2018). Audio Augmented Reality in Public Transport for Exploring Tourist Sites. *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, 721–725. https://doi.org/10.1145/3240167.3240243

Bordwell, D., & Thompson, K. (1997). *Film Art: An Introduction* (5th ed.). The McGraw-Hill Companies.

Bose AR Public Beta Closure FAQ. (2020). Bose Developer Portal. Retrieved 25 March 2021, from https://developer.bose.com/bose-ar-closure-faq

Bose Frames. (2019). Retrieved 15 March 2019, from https://www.bose.com/en_us/products/wearables/frames.html

Busam, B., Ruhkamp, P., Virga, S., Lentes, B., Navab, N., & Hennersperger, C. (2018). Markerless Inside-Out Tracking for Interventional Applications. https://arxiv.org/abs/1804.01708

Buttle, P. (2020). *The Power Behind Video Games: A Look at Game Engines*. Medium. Retrieved 5 March 2019 from https://medium.com/wetheplayers/the-power-behind-video-games-a-look-at-game-engines-2731315086e0

Carnoy, D. (2019). *Bose Frames review: Audio sunglasses that sound surprisingly good*. Retrieved 15 March 2019 from CNET. https://www.cnet.com/reviews/bose-frames-review/

Chion, M. (1994). *Audio-Vision: Sound on Screen*. Columbia University Press.

Daniel, J., Moreau, S., & Nicol, R. (2003). Further Investigations of High-Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging. 18. *AES Convention Paper 5788, (2003 March).* http://www.aes.org/e-lib/browse.cfm?elib=12567

Dardari, D., Closas, P., & Djurić, P. M. (2015). Indoor Tracking: Theory, Methods, and Technologies. *IEEE Transactions on Vehicular Technology*, *64*(4), 1263–1278. https://doi.org/10.1109/TVT.2015.2403868

Doorways in literature and film. (2018). Fringe Arts Bath 2018. Retrieved 21 April 2021, from https://www.fringeartsbath.co.uk/2018-blog/2018/1/20/doorways-in-literature-and-film

Echoes. (2020). ECHOES. Retrieved 26 April 2021, from https://echoes.xyz/

Feng, J., Kim, J., Luu, W., & Palmisano, S. (2019). Method for estimating display lag in the Oculus Rift S and CV1. *SIGGRAPH Asia 2019 Posters*, 1–2. https://doi.org/10.1145/3355056.3364590

Fernandez-Hernandez, I., Vecchione, G., Díaz-Pulido, F., Jeannot, M., Valentaite, G., Blasi, R., Reyes, J., & Simón, J. (2018). Galileo High Accuracy: A Programme and Policy Perspective. *69th International Astronautical Congress (IAC 2018)*. https://www.researchgate.net/publication/328139107_Galileo_High_Accuracy_A_Programme_and_Policy_Perspective

Förterer, H. (2013). *The sound of things*. Retrieved 17 February 2021, from https://www.foerterer.com/sound-of-things.html

Gampe, J. (2009). Interactive Narration within Audio Augmented Realities. In I. A. Iurgel, N. Zagalo, & P. Petta (Eds.), *Interactive Storytelling* (Vol. 5915, pp. 298–303). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10643-9_34

Gamper, H. (2014). *Enabling technologies for audio augmented reality systems.* Aalto University.

Glassner, A. (2017). Interactive Storytelling: Techniques for 21st Century Fiction. CRC Press.

Gordon, J. (2019, August 3). *The first Audio AR beta apps are here | LinkedIn*. Retrieved 26 April 2021, from https://www.linkedin.com/pulse/first-audio-ar-beta-apps-here-john-gordon/

Halverson, J. (2011). *Why Story is Not Narrative | CSC Center for Strategic Communication*. Retrieved 19 April 2021, from http://csc.asu.edu/2011/12/08/why-story-is-not-narrative/

Harju, M. (2019). *Emotive VR*. Retrieved 13 February 2021, from https://matiasharju.com/projects/emotive_vr/emotive_vr.html

Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., & Nironen, H. (2003). *Techniques and Applications of Wearable Augmented Reality Audio*. Audio Engineering Society Convention 114. http://www.aes.org/e-lib/browse.cfm?elib=12495

Herman, D. (2009). *Basic Elements of Narrative*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=437514

Inside-out v Outside-in: How VR tracking works, and how it's going to change. (2017). Wareable. Retrieved 11 April 2021, from https://www.wareable.com/vr/inside-out-vs-outside-in-vr-tracking-343

IRobot Roomba S9+. (n.d.). iRobot. Retrieved 25 April 2021, from https://witt.zone/irobot/roomba-s9plus-dk/

Jacuzzi, G. (2018). 'Augmented Audio': An Overview of the Unique Tools and Features Required for Creating AR Audio Experiences. *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. http://www.aes.org/e-lib/browse.cfm?elib=19688

Kaghat, F.-Z., & Cubaud, P. (2010). *Fluid Interaction in Audio-Guided Museum Visit: Authoring Tool and Visitor Device*. The Eurographics Association. https://doi.org/10.2312/vast/vast10/163-170

Karhulahti, V.-M. (2012). Suspending Virtual Disbelief: A Perspective on Narrative Coherence. In D. Oyarzun, F. Peinado, R. M. Young, A. Elizalde, & G. Méndez (Eds.), *Interactive Storytelling* (pp. 1–17). Springer. https://doi.org/10.1007/978-3-642-34851-8_1

Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Adaptation of the human auditory cortex to changing background noise. *Nature Communications*, *10*(1), 2509. https://doi.org/10.1038/s41467-019-10611-4

Kim, J. (2021, April 20). *Are Bone Conduction Headphones For Real?* Medium. Retrieved 24 April 2021, from https://rethinkreviews.medium.com/are-bone-conduction-headphones-for-real-550676f96d3b

Kim, S. (2015). Bio-inspired engineered sonar systems based on the understanding of bat echolocation. In *Biomimetic Technologies: Principles and Applications*. Elsevier Science & Technology. http://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=2102159

Koivumäki, A. (2018). *Maiseman äänittäminen: Äänimaisematutkimus äänisuunnittelun tukena*. Aalto-yliopiston taiteiden ja suunnittelun korkeakoulu.

Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S., & Pardhan, S. (2016). Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of

sensory loss. *Attention, Perception, & Psychophysics*, *78*(2), 373–395. https://doi.org/10.3758/s13414-015-1015-1

Krekovic, M., Dokmanic, I., & Vetterli, M. (2016). EchoSLAM: Simultaneous localization and mapping with acoustic echoes. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11–15. https://doi.org/10.1109/ICASSP.2016.7471627

Krzyzaniak, M., Frohlich, D., & Jackson, P. J. B. (2019). Six types of audio that DEFY reality! A taxonomy of audio augmented reality with examples. *Proceedings of the 14th International Audio Mostly Conference: A Journey In Sound,* 160–167. https://doi.org/10.1145/3356590.3356615

Kuhn, A. (1985). History of Narrative Codes. In P. Cook (Ed.), *The Cinema Book* (pp. 207–220). BFI.

Kurczak, J., Graham, T. C. N., Joly, C., & Mandryk, R. L. (2011). Hearing is Believing: Evaluating Ambient Audio for Location-based Games. *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, 32:1-32:8. https://doi.org/10.1145/2071423.2071463

Laamanen, V. (2018). Virtual Heritage: Audio Design for Immersive Virtual Environments Using Researched Spatializers. Aalto University.

Larsson, P., Väljamäe, A., Västfjäll, D., Tajadura-Jiménez, A., & Kleiner, M. (2010). Auditory-Induced Presence in Mixed Reality Environments and Related Technology. In E. Dubois, P. Gray, & L. Nigay (Eds.), *The Engineering of Mixed Reality Systems* (pp. 143–163). Springer. https://doi.org/10.1007/978-1-84882-733-2_8

Lindau, A., & Weinzierl, S. (2012). Assessing the Plausibility of Virtual Acoustic Environments. *Acustica United with Acta Acustica*, *98*(5), 804–810. http://dx.doi.org/10.3813/AAA.918562

Liski, J., Väänänen, R., Vesa, S., & Välimäki, V. (2016, August 19). *Adaptive Equalization of Acoustic Transparency in an Augmented-Reality Headset*. Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology. http://www.aes.org/e-lib/inst/browse.cfm?elib=18343

Literary Devices and Terms. (2020). Literary Devices. Retrieved 18 April 2021, from http://literarydevices.net/

Low-Latency Multichannel Audio in Unity. (2019). GitPress.Io. Retrieved 24 April 2021, from https://gitpress.io/@data-tunnel/low-latency-multichannel-audio

Mazuryk, T., & Gervautz, M. (1999). History, Applications, Technology and Future. *VIRTUAL REALITY*, 73.

McErlean, K. (2018). *Interactive Narratives and Transmedia Storytelling: Creating Immersive Stories Across New Media Platforms* (1st ed.). Routledge, Taylor & Francis Group, Focal Press. https://doi.org/10.4324/9781315637570

McGill, M., Brewster, S., McGookin, D., & Wilson, G. (2020). Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376702

Mcmullen, K. A. (2014). The potentials for spatial audio to convey information in virtual environments. *2014 IEEE VR Workshop: Sonic Interaction in Virtual Environments (SIVE)*, 31–34. https://doi.org/10.1109/SIVE.2014.7006287

Microsoft HoloLens. (n.d.). Retrieved 25 April 2021, from https://www.microsoft.com/en-gb/hololens

Microsoft Soundscape. (2018). Microsoft Research. Retrieved 26 April 2021, from https://www.microsoft.com/en-us/research/product/soundscape/

Minute and second of arc. (2021). In *Wikipedia*. Retrieved 14 March 2021, from https://en.wikipedia.org/w/index.php?title=Minute_and_second_of_arc&oldid=1010307616

Myers, C. B. (2011). *The first musical album that's also a location aware iPhone app*. TNW | Apps. Retrieved 26 April 2021, from https://thenextweb.com/news/the-first-musical-album-thats-also-a-location-aware-iphone-app

Nanay, B. (2012). The Multimodal Experience of Art. *The British Journal of Aesthetics*, *52*(4), 353–363. https://doi.org/10.1093/aesthj/ays042

Narrative Techniques in Writing: Definition, Types & Examples - Video & Lesson Transcript. (n.d.). Study.Com. Retrieved 17 April 2021, from https://study.com/academy/lesson/narrative-techniques-in-writing-definition-types-examples.html

NoTours – Augmented Aurality. (2015). Retrieved 27 April 2021, from http://www.notours.org/

OSSIC X: The first 3D audio headphones calibrated to you. (2016). Kickstarter. Retrieved 26 April 2021, from https://www.kickstarter.com/projects/248983394/ossic-x-the-first-3d-audio-headphones-calibrated-t

Positions and Sizes of Cosmic Objects. (n.d.). Retrieved 14 March 2021, from https://lco.global/spacebook/sky/using-angles-describe-positions-and-apparent-sizes-objects/

Proske, U., & Gandevia, S. C. (2012). The Proprioceptive Senses: Their Roles in Signaling Body Shape, Body Position and Movement, and Muscle Force. *Physiol Rev*, *92*, 47.

'Re-imagining the Audio Tour' with the SFMOMA app. (2018). Retrieved 3 October 2019, from https://antennainternational.com/case-studies/sfmoma-re-imagining-the-audio-tour-with-an-app/

Roomscale 101—An Introduction to Roomscale VR. (2017). Retrieved 27 April 2021, from https://blog.vive.com/us/2017/10/25/roomscale-101/

Rouse, R., & Holloway-Attaway, L. (2020). A prehistory of the interactive reader and design principles for storytelling in postdigital culture. *Book 2.0*, *10*(1), 7–42. https://doi.org/10.1386/btwo_00018_1

Rovithis, E., Moustakas, N., Floros, A., & Vogklis, K. (2019). Audio Legends: Investigating Sonic Interaction in an Augmented Reality Audio Game. *Multimodal Technologies and Interaction*, *3*(4), 73. https://doi.org/10.3390/mti3040073

Sandvik, K. (2011). Loïs Tallon & Kevin Walker (eds.): Digital Technologies and The Museum Experience. Handheld Guides and Other Media. New York: AltaMira Press. 2008. *MedieKultur: Journal of Media and Communication Research*, *27*(50), 8 p.-8 p. https://doi.org/10.7146/mediekultur.v27i50.5246

Sarter, N. B. (2006). Multimodal information presentation: Design guidance and research challenges. *International Journal of Industrial Ergonomics*, *36*(5), 439–445. https://doi.org/10.1016/j.ergon.2006.01.007

Scaling the Creator system. (2021). Retrieved 28 April 2021, from https://docs.pozyx.io/creator/latest/beyond-getting-started/scaling-the-creator-system

Schraffenberger, H., & van der Heide, E. (2016). Multimodal augmented reality: The norm rather than the exception. *Proceedings of the 2016 Workshop on Multimodal Virtual and Augmented Reality*, 1–6. https://doi.org/10.1145/3001959.3001960

Sennheiser AMBEO Smart Headset—Mobile binaural recording headset. (2019). Retrieved 26 April 2021, from https://en-ae.sennheiser.com/finalstop

Sinclair, J.-L. (2020). Principles of Game Audio and Sound Design: Sound Design and Audio Implementation for Interactive and Immersive Media. Taylor and Francis. https://doi.org/10.4324/9781315184432

SonicMaps. (2021). Retrieved 26 April 2021, from https://sonicmaps.xyz/

Soundtrails. (2021). Retrieved 26 April 2021, from https://www.soundtrails.com.au/

Spatial. (2021). Retrieved 8 April 2021, from https://www.spatialinc.com/

Story vs. Narrative. (2014). Beemgee. Retrieved 19 April 2021, from https://www.beemgee.com/blog/story-vs-narrative/

Sundareswaran, V., Wang, K., Chen, S., Behringer, R., McGee, J., Tam, C., & Zahorik, P. (2003). *3D Audio Augmented Reality: Implementation and Experiments*. 296. http://dl.acm.org/citation.cfm?id=946248.946841

The English Literary Techniques Toolkit for The HSC. (2018). Matrix Education. Retrieved 18 April 2021, from https://www.matrix.edu.au/essential-guide-english-techniques/the-literary-techniques-toolkit/

Two new Apple HMD Inventions cover Directional Audio Detection and Dust Particle Removal Systems. (2021). Retrieved 25 February 2021, from Patently Apple. https://www.patentlyapple.com/patently-apple/2021/02/two-new-apple-hmd-inventions-cover-directional-audio-detection-and-dust-particle-removal-systems-.html

Ungi, T., Lasso, A., & Fichtinger, G. (2015). Tracked Ultrasound in Navigated Spine Interventions. In *Lecture Notes in Computational Vision and Biomechanics* (Vol. 18, pp. 469–494). https://doi.org/10.1007/978-3-319-12508-4_15

usomo—Unique sonic moments. (2019). Retrieved 8 April 2021, from usomo. https://usomo.de/

Varjo XR-3. (2020). Retrieved 25 April 2021, from https://varjo.com/products/xr-3/

Veltman, J. A., Oving, A. B., & Bronkhorst, A. W. (2004). 3-D Audio in the Fighter Cockpit Improves Task Performance. *The International Journal of Aviation Psychology*, *14*(3), 239–256.

What is GNSS? (2016). Retrieved 10 April 2021, from https://www.gsa.europa.eu/european-gnss/what-gnss

What is the accuracy of UWB? (2020). Retrieved 24 April 2021, from https://pozyx.io/faq/what-is-the-accuracy-of-uwb/

Wu, Y., Tang, F., & Li, H. (2018). Image-based camera localization: An overview. *Visual Computing for Industry, Biomedicine, and Art*, *1*(1), 8. https://doi.org/10.1186/s42492-018-0008-z

Xie, B. (2013). Head-Related Transfer Function and Virtual Auditory Display: Second Edition. J. Ross Publishing.

Yang, J. (Junrui), Holz, C., Ofek, E., & Wilson, A. D. (2019). DreamWalker: Substituting Real-World Walking Experiences with a Virtual Reality. *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 1093–1107. https://doi.org/10.1145/3332165.3347875

Zhang, P., Lu, J., Wang, Y., & Wang, Q. (2017). Cooperative localization in 5G networks: A survey. *ICT Express*, *3*(1), 27–32. https://doi.org/10.1016/j.icte.2017.03.005

Zombies, Run! Wiki. (2021). Retrieved 26 April 2021, from https://zombiesrun.fandom.com/wiki/Zombies,_Run!_Wiki