

UNIVERSITY OF UDINE

DOCTORAL THESIS
XXXVII CYCLE

Exploring Human-Technology Interactions in Audio Mixed Realities

Author:

Alessandro Giuseppe
PRIVITERA

Supervisor:

Prof. Michele GERONAZZO

Co-supervisor:

Prof. Federico FONTANA



*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy*

in

Computer Science and Artificial Intelligence
Department of Mathematics, Computer Science and Physics

2026

Acknowledgements

My deepest gratitude goes to Professors Michele Geronazzo and Federico Fontana, who helped me rediscover my curiosity and interest in science. I am grateful for the way you challenged and inspired me. Beyond your guidance, I would like to thank you for the warm and friendly welcome and for your support in my professional growth, which has significantly broadened my future perspectives.

I would also like to thank Professor Stephen Brewster for his hospitality during my time at the University of Glasgow. My time there was an immersion into a community full of new ideas that greatly shaped my perspective.

To the colleagues I met in Udine, Padova, and Glasgow during these years: thank you for being part of this chapter of my life. The friendships formed in these three cities provided the balance I needed to complete this work. I carry a piece of each place with me.

Contents

Acknowledgements	iii
1 Introduction	1
1.1 Motivation	1
1.2 Aims and Objectives	2
1.3 Thesis Contributions	2
1.4 Thesis Structure	3
2 Background: Egocentric Audio Technologies	5
2.1 Introduction	5
2.2 Egocentric audio perspective	7
2.2.1 Spatial Centrality	8
2.2.2 Binaural Hearing	9
2.3 The Auditory Digital Twin	9
2.3.1 Post-human Participation	9
2.3.2 The Locus of Agency	10
2.3.3 The Entangled Twin and the We-Relationship	10
2.4 The M-I-C-E Methodology	11
2.4.1 Immersion	12
2.4.2 Coherence	13
2.4.3 Entanglement	13
2.4.4 Multiperspective Method	13
3 Immersive Experience and Storytelling	15
3.1 Cultural Heritage Storytelling	16
3.2 Audio Tools and Technological Platforms	19
3.3 Methods	20
3.3.1 Eligibility criteria	21
Search strategy	21
Study selection	21
3.3.2 Data collection process	23
Type of immersion	23
3.3.3 Purpose of each study	24
3.3.4 Computer-aided Qualitative Data Analysis Software	24
3.4 Analysis and Results	25
3.4.1 Sound Reproduction and Spatialisation	27
3.4.2 Interaction with the Virtual Environment	28
Users with special needs	31
3.4.3 Collaborative Experiences	31
3.4.4 Quality of the Experience and Evaluation	32
3.5 General Discussion	33
3.5.1 Sonic Interactions - Notable Mentions	35
3.5.2 The role of personalisation	36

	Personalization in Audio MR Technologies	36
	Storytelling and Personalisation	36
3.6	Research Directions	38
3.6.1	Emotional museum	38
3.6.2	Immersive analytics	39
3.6.3	Archaeoacoustics, virtual musical instruments, historical voices, and personalities	39
3.6.4	Universal fruition, multimodality	40
3.7	Narrami as a case study	40
3.8	Conclusion	41
4	Technological Configurations: Spatialisation in Audio Mixed Reality	43
4.1	Room acoustics: Reverb	44
4.2	User Acoustics: HRTF and personalisation	47
4.3	Volume and EQ matching	48
4.4	Tracking	49
4.4.1	Tracking accuracy	50
	Measurement Setup	50
	Procedure	50
	Results	52
4.5	Discussion	52
5	Immersion and Coherence	53
5.1	Background: authenticity, plausibility and transfer-plausibility	53
5.2	The importance of tracking and distance perception: tracking accuracy	55
5.2.1	Material and Methods	55
	Experimental Question and Conditions	55
	Procedure	57
5.2.2	Results	57
	Plackett-Luce Analysis	57
5.2.3	General Discussion	58
5.2.4	Conclusion	60
5.3	Evaluation on the effect of different rendering factors	61
5.3.1	Audio Quality Evaluation	61
5.3.2	D&D MUSHRA	61
	Setup	62
	Stimuli and Conditions	63
	Procedure	64
5.3.3	Results and Discussion	65
5.3.4	Conclusion	66
5.4	A new method for SAQI evaluation: D&DMUSHRA+	67
5.4.1	Piloting and Screening in VR/AR	68
5.4.2	D&DMUSHRA+: D&D MUSHRA on multiple pages	70
5.4.3	Materials and Methods	72
	Stimuli and Conditions	73
	Procedure	73
5.4.4	Results	75
5.4.5	Discussion	78
	Piloting and Screening with D&DMUSHRA+	78
	Plausibility in Audio AR	79
	Possible Limitations	80

	Reflections on Approximation and Perceptual Tolerance	81
5.5	Evaluation of Co-Immersion	82
5.5.1	Material and Methods	82
5.5.2	Experiment design	83
5.5.3	Results	84
	Preliminary externalization test	84
	Detection performance	84
	Signal Detection Theory analysis	84
	Summary	84
5.5.4	General Discussion	84
6	Entanglement	87
6.1	Entanglement eXperience Map	87
6.1.1	<i>Narrami</i> : Actor-network Map	89
6.1.2	<i>Narrami</i> : EXMap	92
6.2	A Foundational Experiment	94
6.2.1	Immersive and Adaptive Simulation	94
6.2.2	Proxemics	94
6.2.3	Explainability	95
6.2.4	Persuasive Techonologies	95
6.3	A-Priori Interaction Design	95
6.3.1	AI Agent	96
6.3.2	Procedure	97
	Training	98
	Main experiment	98
	Experimental Conditions	98
6.4	From Experiential to Data Space	98
6.5	Intra-Action Analysis: Single Perspectives	102
6.5.1	Researcher's Viewpoint	102
6.5.2	Computer's Viewpoint	103
	Rewards	103
	Agent's Moves	104
6.5.3	Human Participant Viewpoint	104
	Emerging Patterns	106
6.6	Intra-Action Analysis: Systemic View	106
6.6.1	Comparison between Researcher and Computer Viewpoints	107
6.6.2	The Relational Mirror Offered by the Computer Viewpoint	107
6.6.3	Perceived Intentionality	109
6.7	General Discussion	110
6.7.1	Challenges	112
6.8	Conclusion	113
7	Conclusions	115
7.1	Main Thesis Achievements	115
7.2	Future Directions	116

List of Publications

Journal Papers

Alessandro Giuseppe Privitera, Federico Fontana, and Michele Geronazzo. “The Role of Audio in Immersive Storytelling: a Systematic Review in Cultural Heritage”. In: *Multimedia Tools and Applications* 84.16 (June 2024), 16105–16143. ISSN: 1573-7721. DOI: [10.1007/s11042-024-19288-4](https://doi.org/10.1007/s11042-024-19288-4). URL: <http://dx.doi.org/10.1007/s11042-024-19288-4>.

Davide Fantini et al. “Co-immersion in audio augmented virtuality: the case study of a static and approximated late reverberation algorithm”. In: *IEEE transactions on visualization and computer graphics* 29.11 (2023), pp. 4472–4482.

Book Chapters

Alessandro Giuseppe Privitera and Michele Geronazzo. “Designing sonic interactions in intelligent reality with egocentric audio technologies”. In: *The Routledge Handbook of Sound Design*. Focal Press, 2024, pp. 219–251.

Conference Papers

Roberto Barumerli et al. “Spatialized Looming Sounds in Virtual Reality: Reaction Times and Localization Accuracy”. In: *International Conference on Extended Reality*. Springer Nature Switzerland Cham. 2025, pp. 254–265.

Roberto Barumerli et al. “Measuring Motor Planning Using an Affordable Sound-Based Virtual Reality Setup for Accessible Perception-Action Studies”. In: *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2025, pp. 613–618.

Alessandro Giuseppe Privitera, Federico Fontana, and Michele Geronazzo. “On the effect of user tracking on perceived source positions in mobile audio augmented reality”. In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*. 2023, pp. 1–9.

Alessandro Giuseppe Privitera, Federico Fontana, Michele Geronazzo, et al. “Personalization in Audio Storytelling within Virtual and Augmented Reality: State of the Art and Insights”. In: *10th Convention of the European Acoustics Association*. 2023.

Alessandro Giuseppe Privitera, Marco Noro, Michele Geronazzo, et al. “Preliminary evaluation of the auralization of a real indoor environment for augmented reality research”. In: *Conv. Eur. Acoust. Assoc* (2023).

Satoshi Okuno et al. "Effect of Auditory Stimulation to Task and Presence in Selective Blurred Immersive Environment for VR Sickness Reduction". In: *IEEE International Workshop on Sensing, Actuation, Motion Control, and Optimization*. 2023.

Accepted for publication

Florian Ferri, Alessandro Giuseppe Privitera, and Michele Geronazzo. "From Digital Twins to Entangled Twins: A Phenomenological Model for Industry 5.0". In: *Submitted for publication at Zagadnienia Filozoficzne w Nauce (Philosophical Problems in Science)* (2025).

Under Review

Alessandro Giuseppe Privitera, Marco Noro, and Michele Geronazzo. "Piloting and Screening Multiple Factors in Audio Augmented Reality with D&DMUSHRA+: a Case Study on Plausibility". Submitted for publication at *Frontiers in Virtual Reality - Augmented Reality*. 2025.

In Preparation

Alessandro Giuseppe Privitera and Michele Geronazzo. "Traces of Entanglement with an AI Agent in Augmented Reality: The M-I-C-E Methodology". Manuscript in preparation. 2025.

To my family, Ali and Leo

Chapter 1

Introduction

1.1 Motivation

The recent and exponential development of Artificial Intelligence (AI) raises profound questions regarding the impact that emerging technologies, such as virtual and augmented environments (AR/VR, which together form a common ground of Mixed Reality (MR)), have on human users [1, 2]. Despite the technical progress of these systems, the cognitive and behavioural effects of exposure to these technologies remain a subject of intense debate [3].

A critical gap in current research lies in the understanding of the bidirectional nature of this relationship. It is not merely a question of how a user navigates a digital space, but rather how the interaction between the user and the technological interface mutually act on one another. The interface is not a passive tool: even through AI-driven adaptability, it becomes an active participant in the environment [4]. Consequently, it is very important to investigate not only how the user's agency reshapes the system's behaviour, but also to what extent the system influences the user's perception and the decision-making process.

The shift toward viewing the interface as an active participant aligns with the emerging paradigm of Entanglement HCI [5]. This perspective moves beyond the traditional user-centered design, which often treats the human and the technology as separate but interacting entities, considering that the user and the technology are linked in a relation. In this view, agency is not a property held by the user alone, but emerges from the "entanglement" of the human, the algorithmic agents, and the environment. By adopting this lens, this thesis explores how the boundaries between the physical and virtual become blurred, specifically focusing on how these entanglements shape the user's experience in Mixed Reality.

Within this context, personalisation emerges as a pivotal mechanism in modelling interaction [6]. While the theoretical value of adapting experiences to individual users is widely recognized, the specific level of personalization required to be effective—without becoming intrusive or breaking immersion—remains unclear. This highlights the need to identify a correct level where personalisation: *at what point does personalisation serve as a genuine benefit, and at what point does it become a source of friction?* This presents a fertile ground for investigation [7]. Nowadays, it remains unclear whether a deep adaptation of the experience, encompassing environment and interaction, benefits the user or diminishes their sense of agency, for example, by making choices feel predetermined.

This thesis seeks to explore this complex field by establishing a specific focus on **audio MRs**. Auditory interfaces offer a unique pathway to immersion, often bypassing the visual saturation common in modern media, yet they present distinct challenges in terms of spatial presence and cognitive load [8, 9].

To address the complexity of these interactions, this research aims to adopt a multi-perspective approach by integrating the experiences of various actors involved in the MR experience, including the experience designer, the Users, the virtual agents acting on the shared space as mediators or co-constructors of the narrative.

Finally, to ground these theoretical explorations in a concrete social context, this thesis delineates a specific field of application: Immersive Storytelling for Cultural Heritage. By applying these concepts to the preservation and dissemination of culture, we aim to help improve how adaptive, audio-centric MRs can enhance dimensions such as engagement and understanding for diverse audiences.

1.2 Aims and Objectives

The primary goal of this research is to investigate the symbiotic relationship between user agency and intelligent interfaces. To achieve this, the thesis is structured around three core objectives:

Objective 1: To establish a comprehensive framework and working methodology for analyzing the interaction between users and technology. This framework analyses the reciprocal influences human intent, virtual and real environment and intelligent agents controlled by machine learning algorithms.

Objective 2: To validate the proposed framework through an experiment conducted in a real-world, ecological setting. This study will employ a virtual agent, driven by a machine learning algorithm, to interact with the user via an Audio AR. The experiment seeks to move beyond controlled laboratory conditions to understand how these interactions unfold in complex environments.

Objective 3: To delineate a specific field of application within Cultural Heritage where audio serves as a critical mean for engagement. Audio is a key element in fostering user immersion within the virtual experience and functions as a vital channel for conveying environmental information. This objective focuses on how auditory narratives can enhance the experience during cultural heritage experience, e.g. in cultural heritage sites or museums.

1.3 Thesis Contributions

This research aims to advance the state of the art through the following contributions:

Contribution 1: It defines a novel working framework, *M-I-C-E*, which introduces a multi-perspective approach to experimental analysis. This framework integrates the point of view of the key actors operating within the environment providing a more comprehensive understanding of the interaction dynamics.

Contribution 2: It identifies a specific field of application centered on *storytelling within MR for Cultural Heritage*. This contribution explores how audio, narratives and augmented and Virtual Environment (VE)s can be employed to enhance cultural and historical understanding and preservation.

Contribution 3: The research defines a technological platform for audio MR. This platform is designed to easily and quickly implement a personalised real/virtual environment, adapting the experience based on the physical characteristics of the space as well as the anthropometric attributes and the psychological state of the user.

Contribution 4: It presents a set of experiments and studies for the evaluation of interaction in virtual and augmented environments, grounded in the M-I-C-E framework, primarily designed for SIVE. These tools provide metrics and methodologies for assessing the quality and efficacy of interactions from the multiperspective points of view mentioned in Contribution 1.

Contribution 5: It validates the framework through a real-world experiment where a virtual agent interacts directly with a user in an audio augmented environment.

1.4 Thesis Structure

This thesis is organized as follows:

Chapter 2 provides the theoretical background for the M-I-C-E Framework. It defines the methodology starting from the foundational concept of egocentric audio, digital twin, and we-relationship. Furthermore, it delineates the triadic dimensions of Immersion, Coherence, and Entanglement as fundamental dimensions for analyzing user-technology interaction. This chapter is based on publications [10, 11, 12].

Chapter 3 contains a comprehensive review of audio storytelling within cultural heritage. It introduces *Narrami*, a hypothetical case study designed to exemplify interaction dynamics in Audio Augmented Reality (AAR). Critically, this chapter examines the role of personalization in enhancing narrative engagement and user agency. This chapter is based on publication [10, 13]

Chapter 4 details the architecture and implementation of the technological platform developed for this research. This chapter outlines the hardware and software components required to support adaptive AAR experiences, focusing on the systems enabling real-time environmental processing and user tracking. This chapter is based on publication [14, 15]

Chapter 5 offers an in-depth operationalisation of the concepts of Immersion and Coherence. It presents a suite of evaluation tools developed to measure these dimensions, ensuring a consistent methodological approach for assessing the quality of the virtual experience. This chapter is based on publication [10, 16, 15, 14, 17]

Chapter 6 focuses on the dimension of Entanglement. It introduces the Entangled eXperience Map, a novel tool designed to map the dynamics of agency and relationality in MR. Moreover, it applies the M-I-C-E methodology to the analysis of a specific ecological experiment where users interacted with a virtual audio agent in a real-world AAR setting, demonstrating the framework's practical application. This chapter is based on publication [18, 11]

Chapter 7 synthesizes the main findings of this PhD research and discusses their broader implications. It reflects on the efficacy of the proposed framework and outlines potential future investigations.

Although the chapters are presented sequentially, the logical structure of this thesis follows two complementary paths.

Chapters 2 and 3 establish the theoretical and conceptual foundations, introducing the M-I-C-E framework and its application to audio storytelling within Cultural Heritage. These elements converge in Chapter 6, where they are operationalised and validated through an ecological experimental scenario.

In parallel, Chapters 4 and 5 focus on the technological and experimental investigation of Audio MR systems, providing the methodological and empirical basis for evaluating immersion, coherence, and perceptual plausibility.

This dual structure reflects the interdisciplinary nature of the work, where theoretical modelling and experimental validation evolve along partially independent trajectories that ultimately converge in the analysis of entanglement.

Chapter 2

Background: Egocentric Audio Technologies

2.1 Introduction

Section adapted from publication [10].

Sound is a powerful tool that can be used to create immersive and engaging MRs. There are many different ways to use sound in MR. Sound can be used to provide auditory feedback to user actions, spatially locate objects or events within the MR, enhance the sense of presence and immersion, or create a sense of atmosphere or mood, further enhancing the user experience [19]. Overall, the use of sound has the potential to greatly enhance the interactive and immersive nature of the virtual/augmented environment. By incorporating sound into MR, users can have a more realistic experience. For example, in a Virtual Reality (VR) game, the sound of footsteps can help the users feel like they are moving through a physical space. Additionally, by adding ambient sounds, the user's sense of presence and immersion can be enhanced, making the VE more plausible and conforming to the user's expectations. Overall, sound is a vital component in creating a truly immersive and interactive MR experience. MR, as introduced by Milgram and Kishino [20], blends real and virtual worlds along the reality-virtuality continuum. MR can provide an immersive experience in blending real or virtual environments. To achieve a valuable audio MR experience, designing and evaluating interactions properly is necessary, considering various factors such as the technological platform, content, and narrative.

In audio MR (see Fig. 2.1 for an audio-friendly representation of the VR continuum), the degree of isolation from the real environment and the mix of synthetic and real elements that the user experiences can vary. In an audio VE, the user is completely immersed in a synthesised world, while in an audio augmented environment, synthesised audio elements are experienced in the real environment using headphones or other hearing devices with a high degree of audio transparency [21] that can completely or partially limit the external acoustic environment.

In the following, we characterize some classes of audio devices typically used in MR experiences:

- **Bone conductance, projected and supra-aural headphones:** This technology (Fig. 2.1-b), which is frequently incorporated into smart glasses or over-the-ear headphones¹, can project sound directly into the auditory canal without obstructing it. The real audio environment can be heard by the user without any alteration.

¹An historical and iconic example is AKG K1000 headphones.

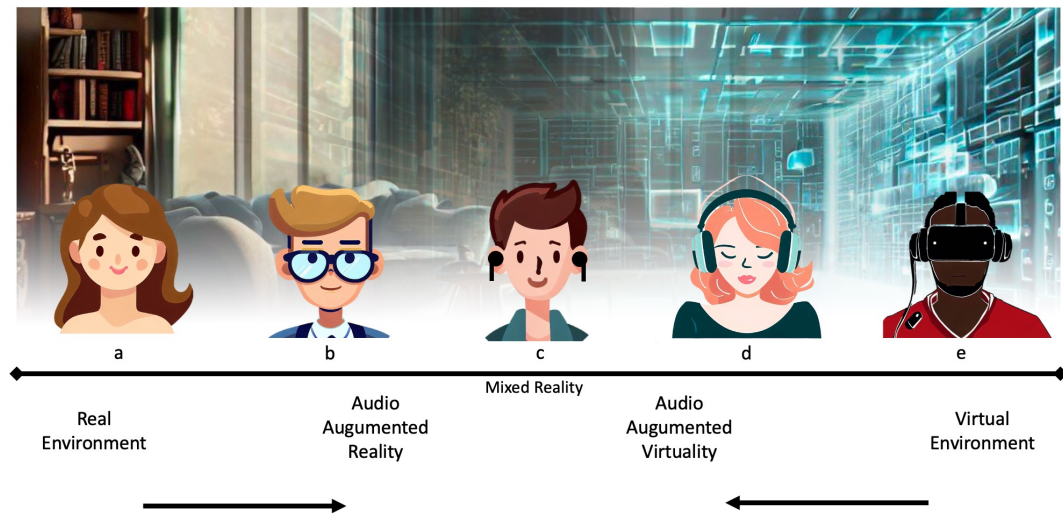


FIGURE 2.1: Milgram and Kishino's Mixed Reality on the Reality-Virtuality Continuum [20] with audio reproduction devices.

- **Earables** (Fig. 2.1-c): Compact, wireless earbuds with different levels of isolation from ambient sound. Audio transparency algorithms that enable ecological listening are implemented in commercial headphones either by employing a design that physically blocks some sound from entering the ear canal (passive) or by employing active components, such as microphones installed in the device, to amplify or reduce environmental audio (active). Although audio transparency permits interactions with the real and virtual environment and other visitors in AAR scenarios, it alters the natural hearing condition, i.e., without hearing devices [22].
- **Headphones**: Headphones (Fig. 2.1-d/e) usually have a high level of isolation from the real environment and can be classified into two types based on their design. Closed-back headphones have a solid back on the earcups, providing high isolation from ambient noise. Open-back headphones allow air to flow freely in and out, providing a more natural soundstage and imaging.

The reproduction devices and audio spatialization techniques are fundamental for convincing audio rendering [23]. They are frequently coupled with spatialization algorithms considering user acoustics in the so-called HRTF [24]. Being in a 3D environment means that a person can use their natural spatial interaction ability. Well-designed spatial interfaces could potentially make interaction and collaboration inside such systems more natural, effective, and engaging. In recent years, there has been a growing interest in the field of sonic interaction design, the study of how sound can be used to create interactive and immersive experiences [25, 26]. This is also attested by the increasing number of available audio-related works, spatial audio frameworks, and SDKs released by commercial companies.²

The integration of AI in MR is triggering shifts beyond materiality, embracing digital dimensions. This entails engaging with dynamic, data-influenced actions

²e.g., Meta Universal HRTF available at <https://developer.oculus.com/blog/improve-spatial-audio-universal-hrtf-meta-quest/>, and Apple's personalized spatial audio (<https://support.apple.com/en-us/HT213318>).

shaped by iterative design-use feedback loops. Nowadays, AI is omnipresent in all parts of MR, including technology through Internet-of-Things (IoT) devices, tracking algorithms, digital narrative, and content, enabling human work in symbiosis with the real and virtual elements moving to an intelligent reality [27]. This new ecosystem should be designed with user needs and preferences in mind using a decentralized perspective, which enriches and gives meaning to a user-centered and egocentric perspective. A higher degree of integration between all the system parts aided by intelligent reality can enhance the user's presence, learning, and engagement inside the cyber-physical system [27, 28].

Ultimately, we can affirm that immersive sonic experiences are inherently dynamic and participatory, actively engaged in a mutual adaptation and personalization process. The motivation behind this chapter is to provide an overview of the research field of Sonic Interactions in Virtual Environments (SIVE [19]), by discussing a theoretical background of the user-technology mutual nature of SIVE.

2.2 Egocentric audio perspective

Technologies for spatial audio rendering can now convey perceptually plausible simulations, and nowadays, there is an increasing consensus that spatial sound contributes to many perceptual and cognitive aspects in VR simulations. Among many different aspects, this is made possible by a high level of personalization in modeling human body acoustics and room acoustics. Therefore, it is essential to provide a terminological characterization that takes into account the prior experience of the users as well as their subjective auditory perception.

In VEs, sonic experience is characterized by an egocentric audio perspective that strongly links the two listening experiences (i.e., in the physical and virtual worlds) [19]. The term egocentric refers to the perceptual reference system for acquiring multi-sensory information in VR technologies and the sense of subjectivity and perceptual/cognitive individuality that shape the self, identity, or consciousness.

SIVE research can be conceptually organized in an inverted pyramid representation with three levels, as shown in Fig. 2.2:

1. The computational aspects of the acoustical-space properties of enabling audio technologies are the focus of immersive audio. This level serves as the apex of the inverted pyramid and the starting point for all SIVE development opportunities.
2. Sonic interaction is human-computer interaction that is carried out through the use of auditory feedback. It includes the study of user interactions to provide abstract meanings and specific indicators of the state of a process or activity;
3. The integration of immersive audio in multimodal VR and Augmented Reality (AR) has an impact on a variety of application domains.

The proposed framework integrates socio-technological entanglements between MR technology and the user, offering the possibility of generating novel insights through adaptive experiences. The egocentric audio perspective plays a pivotal role in engaging with technological intermediaries, facilitating opportunities and tools for reflecting such social interactions in an intelligent reality [6]. The behaviors exhibited by individuals in such an intelligent reality result from the deliberate connection between users and their material existence.

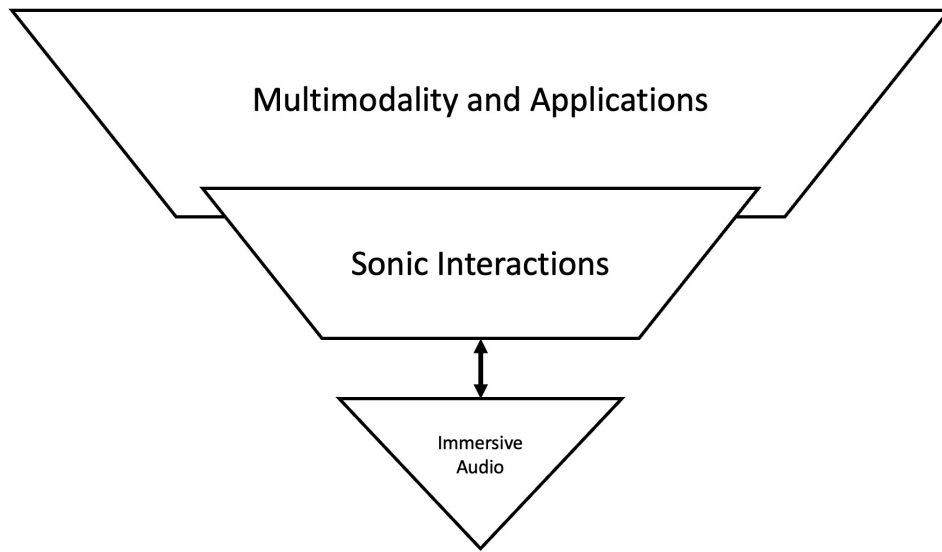


FIGURE 2.2: The SIVE inverse pyramid. Arrows indicate high-level relational hierarchies. Adapted from [6].

Emphasis should be placed on the relationship between the egocentric audio perspective and the research field of egocentric vision, which has more than 20 years of history [29]. Egocentric vision is a subfield of computer vision, including the analysis of images and videos captured by portable cameras, which consider the visual field approximation of First Person Perspective (1PP). From this source of information, spatial and temporal visual features can be extracted to perform various types of recognition tasks, such as detecting objects and activities, and analyzing social interactions. The egocentric audio perspective originated from the same 1PP, and space and time need to be measured around this standpoint for analyzing and synthesising sonic interactions. Furthermore, we emphasize that all hypotheses and assessments in the subjective vision and audition are individually formed around human actors.

2.2.1 Spatial Centrality

Learning and transformation processes occur when the users' attention is directed to external virtual sounds, i.e., "out-of-head" and externalized stimuli. The action space's three-dimensionality is one of the defining characteristics of immersive VEs. Considering this space of transmission, propagation, and reception of virtually simulated sounds, sonic experiences can take on a variety of meanings and present several challenges [6]. Events in multisensory VEs are echoic, producing auditory delays and resonances imprinted by the avatar-VE configurations' spatial arrangements based on the simulated space's acoustic properties. Spatial audio rendering must consider contextual information regarding the spatial positions of sound events and self-produced sounds within the avatar's virtual body to create a sense of proximity and meaningful relationships for the user. The concept of enaction, which states that it is impossible to separate perception and action systematically, shapes the MR experience by considering an embodied, contextually situated perceiver whose sensorimotor processes are tightly coupled to exploratory action.

2.2.2 Binaural Hearing

The sound is acquired through the individual's sensory perception, commonly recognized as 1PP. In particular, externalization can be viewed as a necessary condition for the perception of being immersed in a virtual acoustic space, favoring the so-called place illusion [30]. Ambient reverberation and sensorimotor factors are critical indicators for eliciting a sense of externalisation [31], while HRTF personalization and consistent visual stimuli may, under certain conditions, reinforce the illusion. Consequently, it is essential to investigate dynamic relationships based on specific connections between the evolving states of the user-VE system during VR experiences. Moreover, enormous individual differences in auditory perception necessitate an in-depth examination of several individual factors, such as monaural and binaural HRTF spectral features and temporal adaptation processes.

2.3 The Auditory Digital Twin

The SIVE field finds its most broad-ranging and inclusive expression in the concept of an auditory Digital Twin (DT), a real-time simulation of a cyber-physical system [6]. This simulation could incorporate multiple physical models, probabilistic elements, and temporal data to accurately replicate its real-world counterpart's characteristics while interacting with them. The auditory DT differs from the modern archetype of an avatar [32], which requires a user to control a virtual body that might be noticeably different from the user's physical body. For the auditory DT, considering the egocentric viewpoint of the user is crucial and leads to incorporating personalization techniques on the virtual body when establishing entanglement relations with virtual entities. In any case, the mediation of sound is vital in facilitating the interactions between the user and the diegetic sounds that originate from the avatar's gestures or sound sources within the VE.

The scientific literature attempts to establish different categories of "twins," including "Virtual Twin," "Digital Twin," "Predictive Twin," "Hybrid Twin," "Cognitive Twin," "Contextual Twin," "Experimental Twin," "System Twin" and others (refer to Johansen et al. [33] for a comprehensive list). Moreover, DT could embed AI algorithms in the "Intelligent DT", extending the ability of the digital representation to autonomously alter, control and interact with the virtual and physical world [34], depending on the subjective nature of the experience in which DTs act.

2.3.1 Post-human Participation

Following the recent proposal of a post-phenomenological framework by Verbeek's concept of technological mediation and its extensions [35], it is possible to identify mixed human-technology intentionality within VEs. We can describe the cooperation between humans and technology to reveal a reality that can only be experienced if the technological intention is made accessible to humans, such as by using a reinforcement learning paradigm between a learning agent and its environment in terms of states, actions, and rewards [36]. In particular, Multi-Agent-System (MAS) might access and interact with the physical world, i.e., the user, driving the behavior of Entangled Twins [37]. The self-evolution of MAS has the potential to monitor and predict users' responses and expectations and propose personalized experiences.

The experience participated by a non-self, i.e., auditory DT, is shaped by

1. the past experience and memories of the user and the DT,

2. the physical-acoustic simulation of the body within MR environments,
3. the symbiosis with the technology that induces active and adaptive processes of perceptual relearning.

Placing the physical and virtual worlds on an equal footing produces similar internal representations for the user and her DT, allowing us to promote the transformative role of VR experiences for a human-reality relationship altered after exposure.

2.3.2 The Locus of Agency

According to Frauenberger [5], it is not possible to directly design computers or interaction design. Instead, the focus should be facilitating specific configurations that bring about certain phenomena. Both configurations and phenomena are interconnected in hybrid networks where human and non-human actors engage in causal interactions. The relationships between entities are determined by the “agential cuts,” which establish the boundaries between them [38]. Therefore, the term “agency” denotes a performative mechanism that establishes boundaries and constructs one’s sense of self. The key principle is the shift between *inter-action* between defined objects to *intra-action* within a phenomenon where boundaries between actors are fluidly determined, similar to the Gibsonian ecological theory of perception [39].

In the context of egocentric audio technologies, it is pertinent to consider an intelligent reality as the “locus of agency.” This entails creating a meta-environment that possesses technological and digital characteristics, serving as the custodian, vigilant observer, and vital element for facilitating dialogue and participation among all actors involved. The concept being discussed can be interpreted as an intellectual representation in which extended connections manifest as a dimensional transformation, i.e., analog to digital dimension. It is crucial to affirm that the egocentric perspective remains a cohesive and identity-establishing factor, creating a unique experience around the users and the natural environment that is of value to them.

2.3.3 The Entangled Twin and the We-Relationship

Section adapted from publication [11].

A further development of the DT concept is the Entangled Twin (ET) which emphasise a deeper, more dynamic relationship between human users and digital technology. This idea is the result of interdisciplinary research: on one side, it builds upon a theoretical model developed in the field of Human-Computer Interaction (HCI) [40, 18]; on the other, it incorporates a phenomenological perspective on plural subjectivity arising from the interrelation between the human user and MR technologies. The ET, can be defined as *The user and her/his DT participating in a perceptual and cognitive co-constitution of physical and virtual worlds through adaptive and reflective behaviors of intelligent technologies.* [19]

ETs are conceived as a broader perspective of the DTs, a real-time simulation of a cyber-physical system [41] with the primary aim of replicating and monitoring physical systems within digital environments. On the other hand, ETs give a concrete and perceptible form [42] and a more immersive and interactive relationship. Unlike the typical avatar [32], where a user controls a virtual body that

may differ significantly from their physical self, the ET emphasizes the user's ego-centric perspective. This focus is enhanced by tailoring the experience to the user, particularly when establishing mutual interaction with virtual entities. The user and their DT co-create the physical and virtual worlds through adaptive and reflective behaviors within AR/VRs. This deep collaboration characterises them as ETs. Grounded in phenomenological perspectives [43, 44, 45, 42], Karen Barad's agential realism [46], and Bruno Latour's Actor-Network Theory [47], ETs are conceived as dynamic, adaptive systems in which human users and AI agents co-constitute a plural subjectivity called *We-subject*. This idea refers to a transcendental intersubjective configuration wherein distinct intentionalities enact a common project: the *We-relationship*, a form of intersubjectivity emerges in AR/VR as a shared experiential structure rather, guided by bodily presence and lived coherence, in which virtual and real environments can mutually interact. Drawing on Verbeek's recent post-phenomenological framework, which explores technological mediation and its extensions [35], in AR/VR we can observe a mixed human-technology intentionality within VEs. This collaboration between humans and technology uncovers a reality that becomes accessible only when technological intent is made available to humans. RL paradigms, where a learning agent interacts with its environment through states, actions, and rewards, present a potential method for fostering this type of interaction and collaboration between the two worlds [36]. Specifically, MASs can access and interact with the physical world—the user—thereby influencing the behaviour of these ETs [37]. This perspective challenges traditional notions of agency as an attribute of discrete subjects, emphasising instead that agency is not something possessed, but rather enacted within relational configurations [46]. Accordingly, the ETs do not merely simulate agency; they materialise a distributed agency that is co-constituted within the human-machine mutual interaction. This reading of artificial agency agrees with the position of Floridi [48], who argues that artificial agents, such as avatars or AI systems more broadly, lack intelligence in the human sense and merely perform agency without possessing it intrinsically.

However, it is critical to problematize the reliance on metric space models in such experiments. Within an agential realist framework, space is not a pre-existing container but an experiential topology shaped by intra-action [38]. This perspective also resonates with phenomenological accounts of space [43, 44], where space is not conceived as a neutral, measurable container but as a field of lived experience. Within the ETs framework, the spatial relations enacted by the avatar—though measurable in metric terms—are first and foremost experiential unfolding through the dynamic interplay with the human user of different elements, e.g. proximity, adaptation, and experiential resonance, a shared space of meaning and action.

2.4 The M-I-C-E Methodology

Section adapted from publication [11].

Defining evaluation metrics within and ETs space remains a challenge, as these metrics are heavily dependent on the subjective context in which the ETs operates. In particular, aspects such as the user's perceptual and cognitive state, the level of immersion and engagement within the AR/VR, and the specific context significantly influence how effectiveness and accuracy are assessed. To overcome limitations and better understand and evaluate ETs, a framework based on SIVE is proposed: the Immersion, Coherence, Entanglement (I-C-E) framework. I-C-E framework introduces three intertwined important concepts: *Immersion*, *Coherence*, and *Entanglement*.

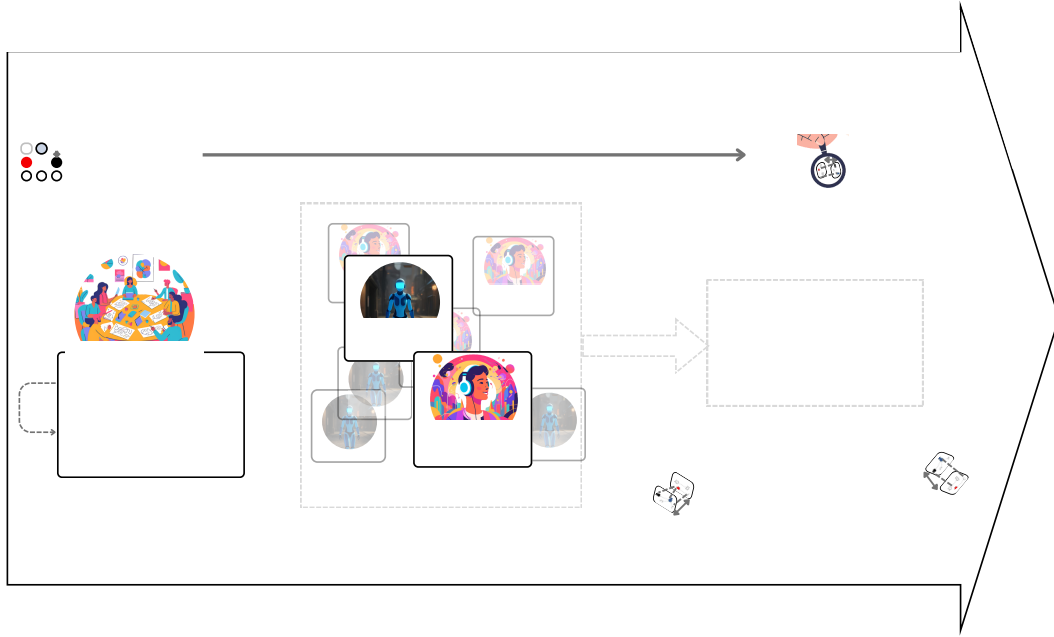


FIGURE 2.3: A schematic representation of the M-I-C-E methodology and its phases and perspectives.

Although the M-I-C-E framework is presented here in a general form, its application in this thesis is specifically grounded in audio-based interactions within Mixed Reality. In this context, sound plays a central role across all three dimensions: as a medium for immersion through spatial audio rendering, as a coherence factor through perceptual plausibility of auditory cues (e.g., localisation, distance, reverberation), and as a key component of entanglement through adaptive and interactive sonic behaviors.

The framework therefore should be interpreted not as modality-agnostic, but as instantiated in this work through egocentric audio technologies and Sonic Interaction in Virtual Environments (SIVE), which provide the primary channel of interaction between users, environment, and intelligent agents.

The operationalisation of the M-I-C-E framework within a concrete Audio MR scenario is presented in Chapter 6, where the proposed concepts are applied to the analysis of an ecological experiment involving a virtual audio agent interacting with human participants.

2.4.1 Immersion

As described by Slater[49], immersion refers to the capacity of a technology to evoke a sense of presence, emphasising the interplay among the system's technological components. Achieving convincing sonic experiences in MR requires carefully designed interactions with both virtual objects and the surrounding cyber-physical environment, supported, e.g., by highly individualised models of human body acoustics and room acoustics [50]. Insufficient or inaccurate design may compromise presence or even lead to issues such as cybersickness.

2.4.2 Coherence

Coherence denotes the perceived plausibility of a hybrid real–virtual environment [51], encompassing digital models’ internal logical consistency, and behavioural alignment (red circle). User-specific configurations (blue circle), captured within the DT, are linked to VE setups such as avatars and other virtual world representations, including the user’s cognitive and sociocultural background, shaped by real-world experiences, thereby enabling a wider range of plausible MR behaviors. An example of coherence in this context is the correct perception of sonic distances, to improve the correct localisation of virtual sounds [52].

2.4.3 Entanglement

Entanglement refers to the dynamic adaptations and active participation of both users and technology, linking Immersion and Coherence [40]. In this context, the concept of the monad, as introduced by Tarde and Latour, is particularly relevant: “*A monad, the relational perspective of each system actor, is not part of a whole but a perspective on all entities taken individually rather than collectively*” [53]. This inherently egocentric perspective encourages the exploration of dynamics within each monadic viewpoint. The term egocentric refers to the perceptual reference system for acquiring multisensory information in MR technologies, anchoring experience in the first-person perspective. From a phenomenological standpoint [43, 44], perception is always situated in the lived body (*Leib*) and structured by conditions of possibility that shape how the world is disclosed to the subject. The experience in VEs, thus, emerges from an egocentric perspective that strongly links the experiences in both physical and virtual worlds [19] through embodied orientation. The use of these I-C-E dimensions to evaluate interaction within the context of ETs can help measurement and assessment. Tools designed to evaluate some of these dimensions are available, especially for coherence and immersion, together with system evaluation metrics, e.g., the performance of the algorithms used. As a result, evaluation often requires a combination of qualitative and quantitative approaches, ranging from interviews and users’ accounts to behavioural data and system performance indicators (e.g.: log files, algorithm evaluation metrics), in order to capture the multifaceted dynamics of the *We-relationship*. The mixed-methods approach underpinning this research serves to complement and extend conventional analytical approaches through the systematic integration of the different perspectives of the actors involved across the design, implementation, and analytical phases of the study.

2.4.4 Multiperspective Method

In an experimental context, several actors interact within and across a series of information and experiential spaces. Each actor contributes uniquely to shaping the unfolding of the experiment and the generation of data.

Figure 2.3 summarises a typical experimental workflow, articulated into two main phases: an a priori interaction design, where design choices and constraints delimit the experiential space, and an a posteriori intra-action analysis, where behavioral and interview information are mirrored within the digital domain. This transition from experiential to data space illustrates how the I-C-E methodology integrates lived interaction with analytical reconstruction, operating not only as a conceptual framework but also as a methodological axis that structures design and evaluation while acknowledging the systemic intra-active nature of the experience.

A Priori Inter-action Design. The idea starts to unfold within the *design team*, responsible for crafting the experiment. This actor has the role of designing the structure, setting conditions, defining the procedure, and recruiting participants. The design process focuses on creating experiences, concentrating on how users will extract meaning from it [54]. For this crucial role, although it should not take an active part during the actual experiment, *design team's* influence is embedded in the VR/AR environment in an attempt to step into the shoes of someone else: the second-person perspective of the researcher shapes the experience of participant's first-person perspective [55] and provides an initial familiarization session. In such a design phase, I-C-E provides a guiding framework to balance technological affordances and human cognitive capacities, ensuring immersive perceptual consistency, a coherent integration, and entanglement through the alignment of adaptive system responses with user expectations. This approach resonates with research through design methodologies that integrate design practice with research inquiry [56].

The Experiential Space. It follows that the *design team* shapes the *experiential space*, including not only the *participants* but also all the non-human actors/agents of the cyber-physical space, primarily the *computer simulation*.

Within this same space, the *computer* acts not merely as a tool but as an active participant, responding to the user through adaptive behaviour and shaping the trajectory of interaction. The participant, in turn, engages with this feedback loop, enacting behaviours and generating responses, giving rise to a *we-relationship*. Accordingly, the I-C-E properties encompass more than perceptual or sensory dimensions: the intertwined analogue-digital environment becomes a locus of agency, where user intentions and computer responses give shape to the ETs. In this perspective, interaction is better understood as intra-action [46], a mutual constitution in which agencies emerge relationally rather than from pre-defined entities.

A Posteriori Intra-action Analysis. From the *experiential space*, a separate but connected *data space* is generated. This space consists of all the quantitative and qualitative data that can be collected for the analysis, following a mixed-methods approach [57], which combines and triangulates numerical measurements with questionnaires and interviews. Specifically, the dataset should include a comprehensive set of data, derived from all the agents, either *global*, i.e., derived from aggregated and predefined categories or definitions, or *local*, defined according to local and specific patterns. These combinations enable a richer interpretation of the results by capturing both the subjective, personalised aspects of interaction and the objectively comparable measures.

During the experiential intra-action, the *data space* is progressively expanded with new data. This space can be a posteriori examined from three different perspectives: the *researcher*, who designs the experiment and later engages with the data space; the *computer*, acting as an experiential mirror of the co-evolving behaviour with the *participants*, who reflect on the lived experience. Together, these perspectives compose the entangled configurations of Entangled Twins (ETs). In this analytical phase, the I-C-E dimensions serve as interpretative lenses for evaluating how ETs experienced the participated environment.

Chapter 3

Immersive Experience and Storytelling

Section adapted from publication [13].

Storytelling and narrative have traditionally been key for cultural heritage preservation throughout human history, embracing diverse contexts ranging from entertainment to culture and knowledge [58]. With the progressive evolution of communication means and multiplication of the supports, stories started to be passed down not only verbally but also with graffiti, paintings, poetry, theatrical plays, musical pieces, and then books, photos, and films until the revolutionary carriers put available by the electronic and finally digital era.

In recent years, the use of technology within digital humanities has become essential for dissemination, education, and research in the cultural heritage field. In particular, MRs are relevant enabling technologies and are nowadays widely used by both academic and corporate research as well as development in multiple application domains ranging from neuroscience to sports training, medicine, education, and entertainment [1]. Cultural heritage can also benefit from MR experiences in a variety of areas, ranging from the reconstruction of cultural sites [59] to the creation of accessible knowledge made possible by technology-mediated direct experience [60, 61]. A *soundscape*, the acoustic real or virtual environment as perceived or experienced and as understood by a person, in context, gives a valuable contribution to tourism and cultural heritage enhancing cultural proposals [62]. In this context, immersive storytelling refers to the use of MR techniques in storytelling, with the goal of eliciting a deeper sense of immersion and presence. Mixed reality offers innovative ways for storytelling, providing immersive experiences that can captivate and emotionally engage audience and facilitate learning ([63]). Moreover, it introduces new affordances that allow more ecological interactive experiences, as suggested by various studies (e.g., [64, 65]). Heightened engagement, emotional connection, and enhanced interaction contribute to increased interest in the subject, thereby enhancing knowledge fruition and dissemination [63]. Audio is a key element in allowing users to become more immersed in the virtual experience and can also be an important channel for information about the environment [66, 67].

Focusing on cultural heritage, disparate studies, and reviews are available on MR storytelling [68, 69], audio [62] and technology for audio AR [70]. However, very few comprehensive contributions that deal with these aspects altogether have been found. Some recent studies put a general focus on audio MR and storytelling [71], but no systematic review is available about the specific topic of cultural heritage, especially including audio as a key aspect.

The insights presented here aim to define a state of the art in the available interactive audio virtual environments for cultural heritage storytelling, and to illustrate

the most useful techniques for enabling audio MR experiences by reviewing the current literature in the field. Additionally, we discuss key elements that create an immersive experience in cultural heritage fruition, as well as how interactivity and personalized, emotional content can be used for educational purposes and knowledge dissemination. Finally, we investigate how audio MR can help storytelling in cultural heritage and the limitations of the current research in discovering and proposing new forms of advanced interactions. From here, we will outline some promising directions. In particular, along with a systematic review of available works, we propose the following research questions:

- Q1. *What enabling technologies and platforms are available for immersive storytelling in audio MR for cultural heritage? These topics will be discussed in Section 3.2*
- Q2. *What are the available works? What insights does the literature research provide about storytelling in audio MR for cultural heritage? The theme will be analysed in section 3.3 and 3.4.*
- Q3. *Which aspects of User eXperience (UX) and interaction are being studied with regard to storytelling in audio MR for cultural heritage research? The reader will find the topics in section 3.4.*
- Q4. *Which implications can be drawn from the literature regarding the design of future research? These matters will be discussed in Sec. 3.5.*

In order to identify a common characterization and to provide some answers to the questions at hand, the chapter is organised as follows: Section 3.2 illustrates tools and technological platforms available from the design and implementation of storytelling in audio MR for cultural heritage. Section 3.3 describes the PRISMA review method and criteria used for the selection of works. Sections 3.4 and 3.5 present analysis results and a discussion based on previously described criteria. Section 3.5.2 tries to define the role of personalisation in the field. Section 3.6 describes possible future research directions and Section 3.7 introduces a fictional case study.

3.1 Cultural Heritage Storytelling

In the context of cultural heritage storytelling, audio rendering, and MR have been introduced and investigated since the seventies. In one of the early theoretical examples, the idea of enhancing the experience with multisensory elements was introduced by Youngblood [72]. The author coined the term *expanded cinema* to describe a form of art that goes beyond traditional cinematic media. This includes special effects, computer art, and multimedia elements. One of the first prototypes for audio augmentation in VR [73] was an automated audio guide, permitting the inclusion in the environment of synthetic audio based on the visitors' position to avoid isolation and enhance social interaction during museum visits. In another seminal study [74], the authors introduced a classification of four categories of museum visitors based on physical navigation, artwork enjoyment, and information browsing for creating an interactive audio guide prototype in AR. The model records the user's physical movement inside the museum to dynamically classify visitors with a non-intrusive approach. Finally, in a pioneering work [75] AréViJava was presented as one of the first platforms for virtual tourism in which an avatar guides tours along virtual

places. In particular, the authors describe the design process for the reconstruction of the Brest harbour site (France) as it was in 1810, enabling a virtual guided tour. During the tour, different viewpoints were suggested by the virtual guide, together with additional audio and video documents accessible through a website.

These early interactive virtual environments established the technical feasibility of avatar-guided heritage experiences, but subsequent research has demonstrated that successful digital narratives require systematic design principles. Some notable efforts to come up with guidelines for effective storytelling recently started: according to The Center for Digital Storytelling in Berkeley¹, digital narratives should be designed following seven specific criteria:

- **Point of View** of the author
- **Gift of voice**, and the register of the narration: colloquial, formal, etc.
- **Dramatic Question**, the intrinsic message conveyed by the story
- **Emotional Content**, the emotions that are transmitted by the narrator
- **Power of Soundtrack**, audio elements in the narration
- **Economy**, the amount of content and information conveyed by the narrative
- **Pacing**, the rhythm of the story, in terms of time storytelling structure

Scientific literature has widely adopted criteria whose primary purpose was education (e.g. [76], for the analysis of autobiographical literacy).

With a specific focus on digital storytelling, Meadows [77] highlights four essential component:

- **Perspective**: the perspective that the author wants to convey with the story consisting of emotions, presentation, and the process of encoding/decoding
- **Narrative**: the story that is narrated by the storytelling in the specific medium
- **Interactivity**: a peculiar characteristic of digital media that can be implemented, e.g., with the design of multiple storylines or choices that the user can do
- **Medium**: storytelling message interpretation can be strongly influenced by the type of medium used [78]

No specific framework for digital storytelling evaluation has been found in the literature, especially in the audio domain, but some efforts have been made in this direction. Sitters et al. [79] explore digital storytelling in the health research domain, suggesting that storytelling should be evaluated for its validity in four different ways. Storytelling material should convey emotions (empathic validity), be credible (intersubjective validity), be sound and just (ethical validity), and have stakeholders play an active part in the design process (participatory validity). This last aspect is also underlined by many other studies, including a framework that considers three different dimensions for the design and evaluation [80]: aesthetics, cognition, and sociality. In sound-driven design, Dalle Monache et al. [67] argue that participatory design should involve stakeholders of different fields in order to develop a deeper understanding of the media and its potential applications. Moreover, we would like

¹<https://www.storycenter.org/>

to emphasize the perspective of Sonic Interaction Design (SID) [81], which considers sound as a primary channel for conveying not only information and meaning but also aesthetics and emotional qualities in interactive contexts. SID goes beyond mere information transmission. It encloses the wholeness of the user experience, enriching it with auditory cues that enhance immersion and engagement. In this context, storytelling can be used to enhance the overall experience, and the SIVE framework aims to help understand how SID principles contribute to crafting compelling narrative experiences. Furthermore, it is important to underline the differences between music and sound. Indeed, sound encompasses a rich spectrum of elements that contribute to the overall immersive experience, including aesthetic qualities and emotional elements. Since music is an art form with specific canons [82], storytelling through music can be considered a special form of storytelling that can be used to enhance the sonic environment or can be enhanced with other sonic or multimodal elements. In this work, we have recognized music's important role in storytelling and included it as a key component. To maintain the inclusive breadth of our work, however, we have presented it as one of several options available. On the other hand, as specified in the two frameworks described above, music can be an integral part of the immersive storytelling experience.

Lugmayr et al. [83] introduce the concept of *serious storytelling*, namely storytelling designed with a specific purpose other than entertainment where "the narration progresses as a sequence of patterns impressive in quality, relates to a serious context, and is a manner of thoughtful process".

Cultural heritage storytelling can be considered a specific case of serious storytelling with some interesting peculiarities. The most important requirements of storytelling for the cultural heritage domain are preserving the correct reconstruction of historical and artistic aspects and conveying a non-ambiguous message. Various works use storytelling techniques to enhance visitor experience in cultural heritage sites or museums, e.g., navigating through a site and following a story. In this kind of situation, personal experiences and emotions can influence information comprehension and interpretation [84] and how users infer and interpret a story [83]. In other words, storytelling can be used to convey cultural heritage notions and concepts for educational purposes effectively. The storytelling creation process itself can also be used to help understand concepts and historical timelines, e.g., in classroom experience [85]. In this case, the storytelling design process should be correctly used to convey cultural heritage knowledge.

Another important aspect in various cultural heritage-related applications, such as tourism and education, is engagement. Storytelling can be a powerful means for transmitting knowledge, but it should be designed in such a way as to correctly fulfill the task, convey the desired message, and avoid errors in the interpretation [68]. In other words, storytelling for cultural heritage can include engaging elements and gamification, however such elements should not be the primary aim. Moreover, Podara et al. [86] suggests that interactive and non-linear storytelling can be useful to elicit engagement, especially in younger people. On the other hand, the amount and type of interactive elements should be carefully calibrated to avoid the risk that users lose interest in the storyline or in the storytelling message [87]. Again, no comprehensive framework can be found to evaluate serious storytelling. However, various tests are available to evaluate specific elements such as engagement, memorability, cognitive load, etc. The method used in the reviewed paper is discussed in Sec. 3.4.4.

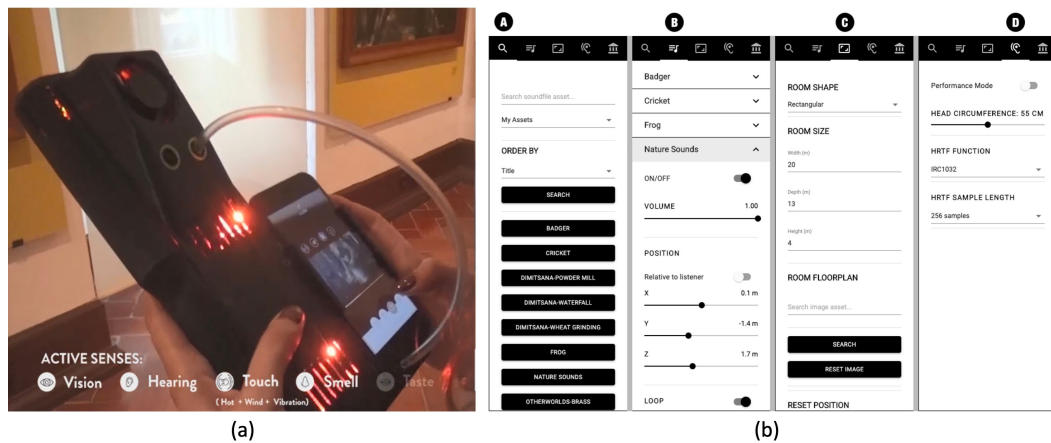


FIGURE 3.1: Examples of technological platforms: a) M5SAR platform, audio guide with the elicitation of all the five senses photo © [88] b) screenshot of PlugSonic user app. photo © [89]

3.2 Audio Tools and Technological Platforms

The analysis of the reviewed papers shows a very diverse situation in terms of frameworks and technologies used. Thus, it is difficult to identify the main trends and durable tools over the years, especially when considering open-source or reusable solutions without license fees. Most of these technologies do not find practical applications in specific cultural heritage contexts but serve as valuable insights into the landscape of available technologies within cultural heritage but may not directly contribute to identifiable trends or prevalent practices. However, many recent immersive storytelling applications were realized using commercial software (Unity3D² or Unreal Engine³) or free SDKs (Vuforia, ARKit, AR Foundation⁴) during the design and development stages, while some examples of custom solutions can be found for specific purposes.

Platforms specifically designed for exploring spaces, providing navigation and spatial audio are described in various studies [90, 91, 92, 93]. In particular, AVIE is a system for directional audio in VR theaters [94], while other audio technologies for spatialisation in theatres are included in Beck's compendium [95] along with a historical overview.

Various technologies are available for creating personalized sonic experiences, using both multi-channel speakers for room audio spatialization (Ambisonics, Dolby Atmos, etc.) and headphones for mobile applications. With the latter technologies, head and body movement tracking along with dynamic binaural audio technology are crucial for auralisation, i.e., the mathematical modeling of physical audio sources in space [96] convincingly localizing a sound source for attentional and navigation purposes [97]. User movement can be tracked by various types of sensing devices. For outdoor position tracking, GPS information is used [98] often combined with inertial measurement units (IMU) [90, 99], and gyroscopes [100] are used for determining the indoor precise position and/or head orientation.

Audio is combined in a multimodal interaction through the five senses in the M5SAR platform (Fig. 3.1-a) [88, 101], by using a custom-built mobile device along

²<https://unity.com/>

³<https://www.unrealengine.com/en-US/>

⁴<https://unity.com/unity/features/arfoundation>

with a mobile app that recognizes artwork in a museum. PlugSonic [89] (Fig. 3.1-b) is a web-based platform for spatial audio storytelling in cultural heritage. Another storytelling design tool is MOGRE [102, 103], a software aimed at the creation of 3D scenarios and stories based on the Ogre⁵ graphics engine for children.

Ec(h)o [104, 105] is an engine for audio material classification and real-time search based on user position tracking and preferences. Proper audio material, classified by using a given ontology, is inserted in a database and presented in real time to the user who interacts with a tangible interface (cube) while moving. Practical guidelines and conceptual frameworks are available on cultural experience design. For example, “Adaptive Augmented Reality” (A2R, [106]) is an AR museum guide architecture that recommends content based on gestural and biometric information to elicit interest and engagement during museum visits. Katz et al. [107] suggest that acoustic space is an important aspect to consider in communicating acoustic heritage and producing audio experiences. Moreover, they introduce *Past Has Ears* (PHE), a hardware/software prototype for the presentation of immersive audio experiences adaptable to multiple platforms. Polaris~ [108] is a wearable platform creating privacy-respectful audiovisual AR experiences to foster artistic and musical expression. Polaris~ comprises an open-source AR headset (Project North Star), a pair of bone-conduction headphones, and software developed in Unity and PureData. Meyer [109] suggests guidelines for story design in a spatial context, based on a literature review, suitable for interactive audio-visual narrative and 360° films. In his work, Fujinawa et al. [110] hypothesized that a comfortable sound field could induce different behaviours in human movement and navigation in a physical space. In a laboratory study, users are left free to move in a room where sound is diffused with different pressure levels by motor-controlled moveable speakers in three different audio stimuli conditions (white noise, Jazz music, no sound). By using questionnaires and position tracking, user preferences and time spent in areas with different loudness levels have been analyzed, overall showing longer staying in more quiet areas when audio stimuli were evaluated as unpleasant. Kenderdine et al. [111] describe the potential application of computational practices within archival and museological domains that could improve the preservation and cataloging of sources, provide new forms of representation and knowledge, and can empower new forms of art. In this direction, *Jazz Luminaries*⁶, and interactive installation that display the connection between jazz, blues and Latin artists using a 3D visualization in a fulldome. Other contributions can be found in interactive musical performances such as *Carillon*⁷ and *Membrana Neopermeable* [112] that uses a virtual environment used by the performers to interact with virtual instruments.

3.3 Methods

While structuring this review, we followed the PRISMA method [62] as a guide for our work. PRISMA is widely used in diverse research areas as medicine [113], psychology [114], and computer science [36]. The data search and paper selection process is described here.

⁵<https://www.ogre3d.org/>

⁶<https://www.kirellbenzi.com/art/jazz-luminaries>

⁷<https://ccrma.stanford.edu/rob/portfolio/carillon/>

	Description
Inclusion Criteria 1	Article and book chapters related to Audio Virtual and Augmented Reality Storytelling in Cultural Heritage
Inclusion Criteria 2	Work from Journal Paper or Conference Proceedings
Inclusion Criteria 3	Paper, Article or Book written in English

TABLE 3.1: Inclusion Criteria

3.3.1 Eligibility criteria

Selected papers or books must include research, reviews, or case studies about the use of MR for cultural heritage storytelling, specifically referring to audio as a part of the dissertation. Table 3.1 lists the inclusion criteria. It is worth noting that the author found no work including all the considered topics, probably due to limitations binding the different fields. See, for example, the work in [115], claiming implementations of a mobile virtual environment through a PDA as a future possibility should technological advancements allow for, or “KidsRoom” [116] that describe an AR interactive experience for storytelling not specifically designed for cultural heritage.

Search strategy

Our research was conducted by performing an extensive search on online publication-related databases: Elsevier Scopus, ACM Digital Library, IEEE Xplore, and Google Scholar. In all these databases, four different clusters of keywords were used to include synonyms and related keywords. Presented below are the four names representing different groupings:

- **Storytelling:** storytelling, narrative, guide, guidance.
- **Audio:** sound, audio, acoustic, acoustics, auditory, sonic, hearing, soundscape, soundscapes, voiceover, narrator.
- **Immersive media:** VR, Augmented Reality, VR, AR, Spatialized, Spatialised, Spatialization, Spatialisation, virtual environment, XR, extended reality, immersive media, metaverse, immersive space, immersion, virtual agent.
- **Cultural heritage:** cultural heritage, museum, history, tourism, culture, archaeology, archaeological, historic, conservation, preservation, tangible heritage, intangible heritage, living heritage.

Keywords were connected by using OR inside clusters AND between clusters. Search was performed in Title, Keyword, and Abstract in Scopus, IEEE, and ACM by using their specific language. Due to Google Scholar’s limitation in search string length (256 characters), a different string format has been used to include a wider selection of papers.

Study selection

Starting from the previously mentioned inclusion criteria, Fig. 3.2 depicts and characterizes the PRISMA phases that considered a total of 574 articles selected from 4

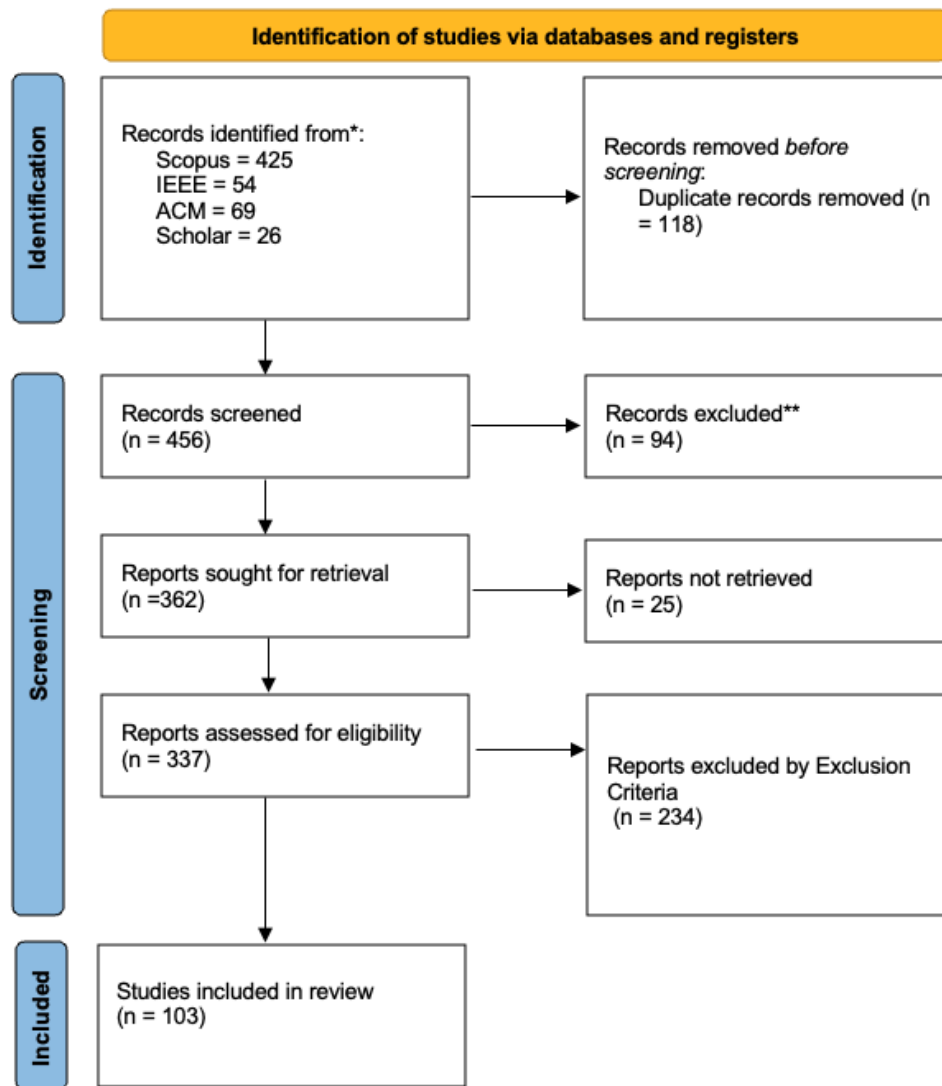


FIGURE 3.2: PRISMA flow diagram.

scientific databases. Duplicate papers (118 in total) were removed, and the remaining 456 papers were screened by title, abstract, and scanning using the exclusion criteria listed in Table 3.2. The selected papers must have a strong research focus on cultural heritage and Storytelling, therefore work with different main topics (e.g., technologies that can also be used for cultural heritage, general sound localization studies, generic AR platforms, solutions for urban navigation, etc.) and works in which one topic was absent have been excluded. Since the audio-sensory modality was particularly important for our systematic review, we introduced an ad-hoc classification:

- **Category 1, Audio First:** audio is the leading modality in the paper.
- **Category 2, Audio and Multimodal Integration:** audio is considered together with other modalities (e.g., haptics, video, etc.), but specific information is given.

	Description
Exclusion Criteria 1	Article or paper not purchasable/unavailable
Exclusion Criteria 2	Articles or papers are classified using the following categories and double-checked by authors. Articles in Category 5 and 6 are excluded from evaluation.
Exclusion Criteria 3	Not about cultural heritage but other aspects are described (e.g., a technology that can also be used for cultural heritage)
Exclusion Criteria 4	No Storytelling is involved
Exclusion Criteria 5	No VR

TABLE 3.2: Exclusion Criteria

- **Category 3**, Audio in Multimodal Comparison: audio experience is considered in comparison with others (e.g., Audio vs. Haptic feedback).
- **Category 4**, Multimodal Experience: audio is present, but it is not possible to ascertain its specific audio contributions.
- **Category 5**, Other experience: audio is present but not considered.
- **Category 6**, No Audio: audio is not present.

All papers with level 5 or 6 were excluded since they were of little or no interest for this review.

3.3.2 Data collection process

Type of immersion

A classification for MR systems has been proposed by Bekele et al. [63], which introduces two different categories for AR systems and three for VR systems. Bekele et al. proposed a general classification based on system flexibility and experience. We exported this classification to auditory information:

- **Outdoor audio AR:** By using GPS or other markerless tracking techniques, people can navigate inside a vast open-air field and obtain location-aware audio information.
- **Indoor audio AR:** Indoor tracking needs precise information about head orientation and user in a limited space. The most used techniques are IMU, Gyroscope, and AR toolkits like ARKit⁸ or Vuforia⁹.
- **Non-Immersive audio VR:** The virtual environment is viewed from a desktop or screen, and audio content is conveyed from a speaker and not spatialized.
- **Semi-Immersive audio VR:** Audio content is conveyed through multichannel or directional speaker systems, usually in rooms with multiple users.

⁸<https://developer.apple.com/augmented-reality/>

⁹<https://developer.vuforia.com/>

- **Immersive audio VR:** A user is fully immersed in an auralized virtual environment or audio content that is realistically integrated inside a physical environment, with a high level of presence.

3.3.3 Purpose of each study

Mixed reality can provide important support to cultural heritage for different purposes. Different surveys have already been published on this topic [117]. Based on these works, we identify the following main purposes for the use of immersive media technologies:

- **Education:** We identify systems designed with the aim of learning and dissemination, such as MR books for children [118] and audiovisual storytelling experiences [119].
- **Exhibit enhancement:** A work about cultural heritage audio augmentation applied to tour visits [120] and an app for augmentation of a physical diorama [121].
- **Exploration:** Applications or methods for navigating or discovering spaces or contents, such as augmented audio guides [98, 122].
- **Reconstruction:** Re-creation of elements of the past inside today's world as the reconstruction of Fort Sant Jean [123] and AR interviews [124].
- **Virtual museums and interactive installations:** A virtual cultural heritage experience such as virtual drama [55].

3.3.4 Computer-aided Qualitative Data Analysis Software

As part of the research process, we conducted an automatic content analysis on the papers we considered. For the automatic content analysis we utilized *Leximancer*¹⁰, a Computer-Assisted Qualitative Data Analysis Software (CAQDAS) that extracts statistical properties from a text to identify a list of terms and concepts [125]. The software identifies highly connected concepts and clusters them into higher-level groups, defined as themes.

Leximancer uses an analysis method that examines a unified body of text—in this case, reviewed articles. The program selects a ranked list of emerging lexical terms based on their frequency and co-occurrence usage. These terms are used for creating a thesaurus, which in turn builds a set of classifiers from the text by iteratively extending the seed word definitions. The result is a set of weighted term classifiers known as concepts. The text is then classified using these concepts at a high resolution—typically every three sentences. This produces a concept index for the text and a concept co-occurrence matrix. An asymmetric co-occurrence matrix is created by calculating the concepts' relative co-occurrence frequencies. This matrix produces a two-dimensional concept map using a proprietary clustering algorithm based on the spring-force model for the many-body problem [126]. The correctness of each concept in this semantic network generates a third hierarchical dimension, which displays the more general parent concepts at the higher levels.

¹⁰<https://www.leximancer.com/>

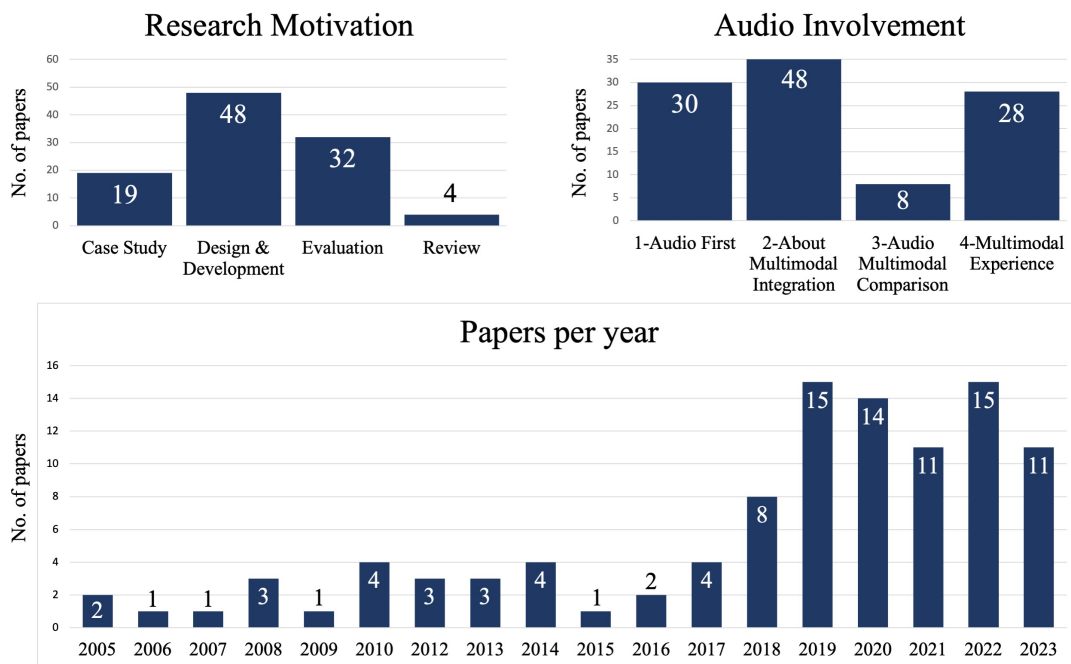


FIGURE 3.3: Count of articles categorized by (top left) research motivation, (top right) audio involvement level, and (bottom) publication year of eligible studies

3.4 Analysis and Results

A total of 103 papers have been considered eligible for the study. Works were read and analyzed to collect useful information about the study and disclose research trends. All contributions, available in Supplementary Materials of [13], have been classified based on the purposes of the study. Among all papers, 63 describe a cultural heritage site application, and 40 are theoretical or applied studies.

One first important consideration concerns the role of audio components in the context of MR storytelling in cultural heritage. Results in Fig. 3.3 show that 30 studies out of 103 have been specifically designed to evaluate auditory and 37 studies consider audio in a combined [98] or comparative [127] multimodal perspective. On the other hand, in 27% of the studies, it is not possible to determine a specific contribution of audio since this component is ambiguously considered in combination with video or other elements. However, the interest in auditory evaluation is increasing since the number of papers in the field has grown in the last few years (see Fig. 3.3). As also noted by Jerald [128], the interest in MR is undoubtedly increasing, also thanks to new technologies and platforms available in the market.

Figure 3.4 shows the main themes identified by CAQDAS analysis, and the main concepts considered:

- **Sound:** concepts regarding audio and spatialisation are included in the theme, e.g., *spatial, sound, position, environment, etc.*
- **Story:** represents storytelling and narrative elements including *narrative, emotional, music, media, history, landscape, world, etc.*

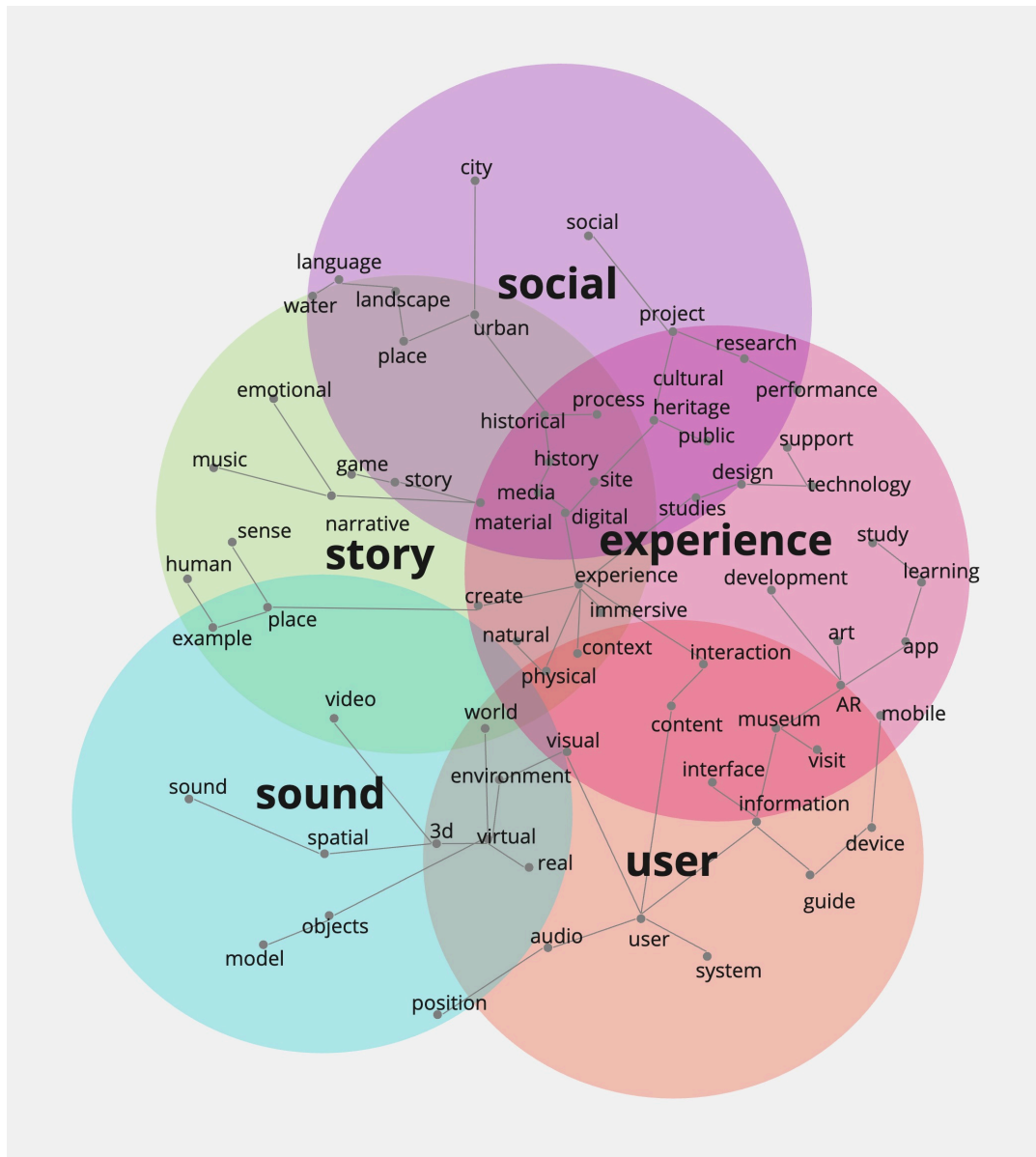


FIGURE 3.4: Topical map of discovered concepts. Clusters are displayed as heated circles.

- **Experience:** it refers to the overall user experience and is very intersected with both content and narrative terms (such as *immersive*, *history*, *natural*, *create*, *cultural heritage*, *digital media*) and technological elements (*technology*, *interaction*, *interface*)
- **Social:** includes concepts related to cultural heritage spaces and collaborative experiences, e.g., *urban*, *public*, *site*, *city*, *landscapes*, *performance*
- **User:** centrality and need of personalisation of user in theSIVE experience are underlined in this theme that includes *user*, *audio*, *interaction*, *guide*, *visit*.

Storytelling is central to the MR experience and requires careful design in order to improve user interest in different aspects (engagement, education, experience time,

etc.) while maintaining the characteristics of effectiveness and correctness of information. Content has to be built around the user, following an egocentric perspective, and should be personalised respecting each person's particular traits and special needs. Technology must help make the experience realistic and plausible inside the virtual environment, including using collaborative elements. The main concepts discovered in CAQDAS analysis are discussed in greater detail in the following subsections.

3.4.1 Sound Reproduction and Spatialisation

Technology plays an important part in conveying a sense of presence in terms of Immersion, Coherence, and Entanglement dimension, corresponding to *Sound* CAQDAS concepts. A plausible audio rendering requires several types of reproduction devices and spatialisation techniques [23], as it is also underlined by CAQDAS analysis, which includes *spatial* and *position* in the domain, as well as *real*, *virtual* and *place* emphasizing the interplay between all experience elements, whether tangible or not. Moreover, the intersection between *sound* and *user* underlines the importance of an egocentric audio perspective. The devices identified in the reviewed works have been divided into three different categories:

- **Headphones/earphones:** personal reproduction devices with two small speakers available on the market in different dimensions, ear fit (in-ear, outer-ear, over-ear), and connectivity (wired or wireless). Audio headphones can reproduce monophonic sound, diffusing the same audio information in both speakers, stereophonic sound accounting for a horizontally localized source, or binaural sound considering listener acoustics. Of particular interest here, stereophony conveys a sense of direction, especially if in conjunction with spatialisation models simulating a physical environment [129]. Moreover, audio headphones can isolate totally or partially from the external acoustic environment (audio transparency [130]).
- **Directional speaker:** a particular type of loudspeaker with a narrow acoustic beam. Directional audio speakers provide personal content while maintaining environmental audio transparency [131], with performance limits for listeners who occupy areas outside a known sweet spot.
- **Multichannel:** a loudspeaker set conveying spatial audio and permitting sound source localisation. Multichannel audio can be composed of similar or different speakers in terms of frequency reproduction (frequency response). In the reviewed papers, some particular multichannel sets have been identified in two works [99, 132]. The former uses four bone conduction speakers installed in a headband to create audio spatialisation and orient the user in a cultural site using GPS and compass sensors. The latter uses a combination of custom-built headphones and room speakers to enable a social experience in a theatre (more information is given later in this chapter).

Information about audio reproduction and spatialisation are available in Fig. 3.5. In the considered studies, the most widely used audio reproduction devices are headphones/earphones, which, as opposed to other solutions like mobile phone integrated speakers, can limit social/collaborative interaction inside museums or cultural sites [133] and modify the presence and immersion level of the system [134].

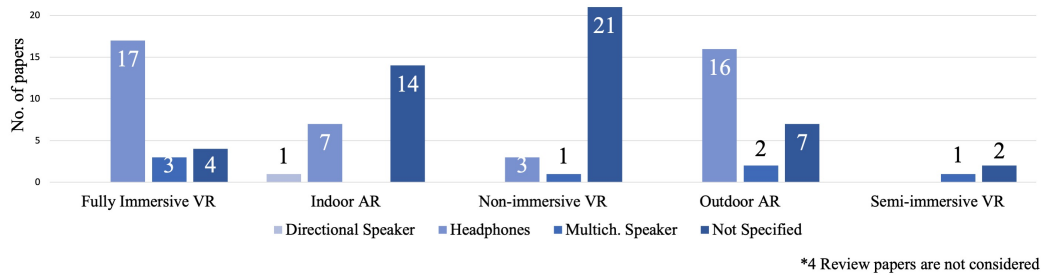


FIGURE 3.5: Count of articles categorized by sound reproduction device and immersion type

However, it is important to highlight that no sound reproduction device is specified in 48 studies (46.7%), even though it could be an important aspect of experience evaluation, especially in studies based on portable devices.

It is worthwhile to note that audio spatialization and algorithm information are still sparsely treated, but interest in the topic is increasing. In fact, only 35 studies involve a technology making use of spatial audio or more than two speakers. Interestingly, 18 such studies have been published recently as in the last two years.

Commercial headphone equalizers based on HRTFs (Unity or Unreal Engine with SteamAudio or FMOD) are used in various works [90, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147]. Other headphones, commercial or open-source headphone-based rendering models, and commercial or open-source libraries have been considered in reviewed papers: OpenAL [148, 149, 150], Audiokinetic Wwise [151], Apple XCode [147], Processing [152] or not specified [153, 100, 154, 155, 156, 157, 158, 159]. Andolina et al. [160] uses a “headset capable of 3D audio”.

In works with a multichannel setup, SensiMAR [161] uses 4-speaker Ambisonics 3D audio format, a method for reproducing audio on a spherical spatial sound-field [162], in outdoor Conimbriga archaeological site to recreate the soundscape of ancient Roman activity. “Trying to get trapped in the past” [163] uses Wave Field Synthesis on a linear loudspeaker array to spatially render a virtual theatrical play. “Pigments” [164] use Dolby Atmos for the sonification of *Pigments of Imagination* audiovisual installation. Finally, a Dolby 5.1 system is used for the sonification of Roald Dahl’s “The Time Machine” story [165].

An interesting approach is used in the installation called CAVE [132, 166], a VR theatrical play on nomadic tribes of Northern Europe, whose audio setup includes a combination of custom-built off-ear headphones providing spatial audio and room speakers. The headphones convey sound effects and dialogues, while room speakers and a subwoofer diffuse soundscape and music.

3.4.2 Interaction with the Virtual Environment

A correct interaction design and evaluation is fundamental for achieving a high level of coherence and immersion in a system, and in fact, the experience theme resulted from an interaction design activity in 83 out of 92 papers. *Experience* in CAQDAS analysis, along with related concepts such as immersion, support, interaction, and interface, emphasize SIVE as a key area. Fig. 3.6 classifies the user interactions with the system grouped by tracking method: user body position and head orientation inside the physical or virtual space, object recognition in rooms or sites, physical actions on tangible interfaces, etc.

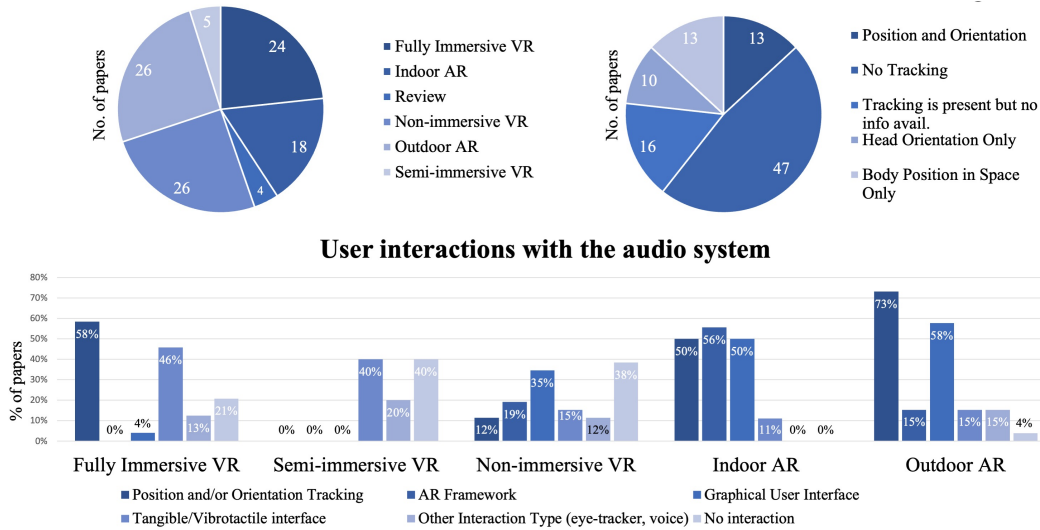


FIGURE 3.6: Distribution of articles by (top left) immersion type, (top right) orientation and position tracking, and (bottom) interaction type

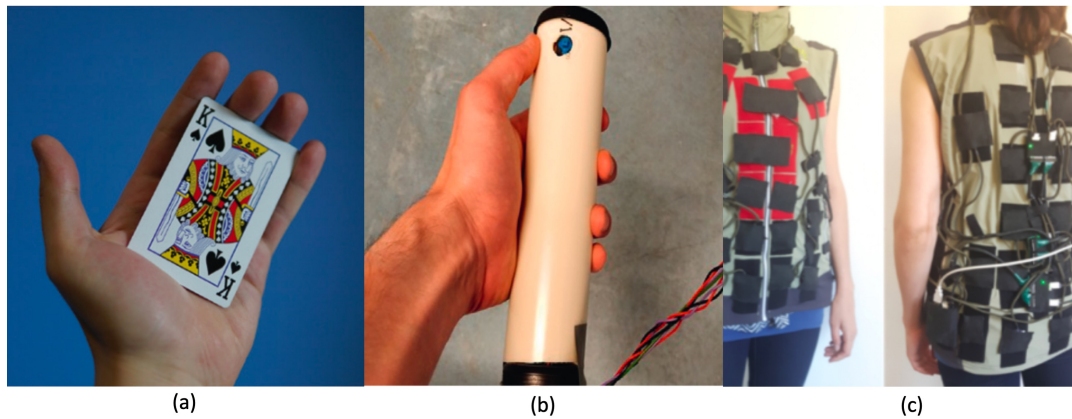


FIGURE 3.7: Examples of tangible devices for interaction: a) AR-controlled playing card. photo ©[167] b) playback control device in[137] c) haptic vest with built-in sensors. photo ©[160]

The most common tracking method (55 papers) selects and personalises audio and video content based on head orientation and position in space, especially in indoor situations. In all outdoor experiences, GPS is used along with compass sensors [99, 100, 122, 168, 169, 149, 150, 170, 171], otherwise tracking is left unspecified [120, 160, 170]. GPS is also used in one interactive artwork described “called Hybrid Gifts” [172] in which a smartphone app permits museum visitors to communicate by geolocated audio messages with peers in order to express and transmit their emotions while looking at paintings in a museum. For a more precise indoor user position estimate, tracking is performed by inertial IMU sensors [90, 173, 174]; otherwise, it is not specified. In VR contexts, user orientation in virtual space is usually tracked using Head-Mounted Displays (HMD) integrating IMU sensors.

Direct interaction with a Graphical User Interface (GUI) on a device screen is widely used (33 papers), especially in AR experiences, often relying on a smartphone as an interaction device, enabling at the same time the use of a wide range of built-in sensors such as IMUs and cameras as well as internet connectivity. It is

important to note that no immersive or semi-immersive experience uses GUI interaction, preferring user movement tracking or tangible controllers.

Computer vision algorithms enabled by AR frameworks (ARKit, ARCore, Vuforia, OpenAL, etc.) are used to recognize objects and elements in space and to interact with an audio virtual environment in 22 works. AR is used to present context- and position-aware touristic information [175, 174, 160, 176, 161, 139] and to navigate in an AR diorama by using a portable device (smartphone or tablet) [121]. Various physical books [118, 177, 178, 179, 178] have been digitally augmented by using devices that recognize book pages and show related content or QR codes [175, 180, 181].

In *War children* [124], stories narrated by WWII eyewitnesses can be watched inside the user's room through a smartphone screen. Huang et al. [182] describe a series of video games for touristic purposes, one of which asks visitors to find a specific image inside a cultural site by using AR to get a game reward.

23 studies use tangible devices, such as console games or VR headset controllers [183, 136, 184, 185, 153, 182, 140, 186, 141, 142, 144, 147], haptic interfaces [187] or custom-built devices. In *ec(h)o* [105], users indicate a particular preference for presented audio content by rotating a wooden cube with coloured faces whose position is detected by a camera. A similar approach is used [167] in a comparison test about movement in a virtual museum, with four different interaction methods on movement in the virtual environment and control condition over an audio guide playback. During the experiment, users can control audio by rotating a playing card (Fig. 3.7-a), whose movement is recognized by using the Vuforia AR framework. A preference for controlling movements directly was identified.

In an interactive installation, Kenderdine [188] presents a custom-built console to interact with, Kortbek et al. [131] uses a three-step staircase with pressure sensors to detect visitors stepping. Moreover, Geronazzo et al. [137] implemented a custom-built pipe controller (Fig. 3.7-b) with a button and a 9-axis IMU in order to control movement in an audio virtual environment. On the other hand, Andolina et al. [160] uses a haptic vest with different mounted sensors (Fig. 3.7-c) and actuators in a navigation task inside a city. The preliminary study compares different interaction/feedback modalities (Visual AR, haptic-audio guidance, and the two combined), showing a preference for the haptic-audio modality. Finally, *The Time Machine* [122, 168, 169] uses vibrotactile feedback together with audio produced by a smartphone for navigating a cultural heritage site. The audio-vibrotactile feedback provides user information about the distance from a specific position of the site.

Other interesting multimodal interactions were described in 8 studies. Two of them use different eye-tracking techniques to detect the user's gaze and present information about the artwork or artistic building the user is looking at in a virtual environment. In particular, Kelling et al. [189] uses eye position estimation based on an IMU sensor of an HMD, Sanchez et al. [152] uses gaze to control the playback of audio elements, such as effect or noises, while reading an children's tales book, and Kwok et al. [190] use eye-tracking glasses. In a fourth study [187], one of the author's interactive installations uses voice and pitch tracking to induce vibration in objects inside a virtual environment based on voice intensity. Voice is also used in an AR experience [139] to receive information based on the narration.

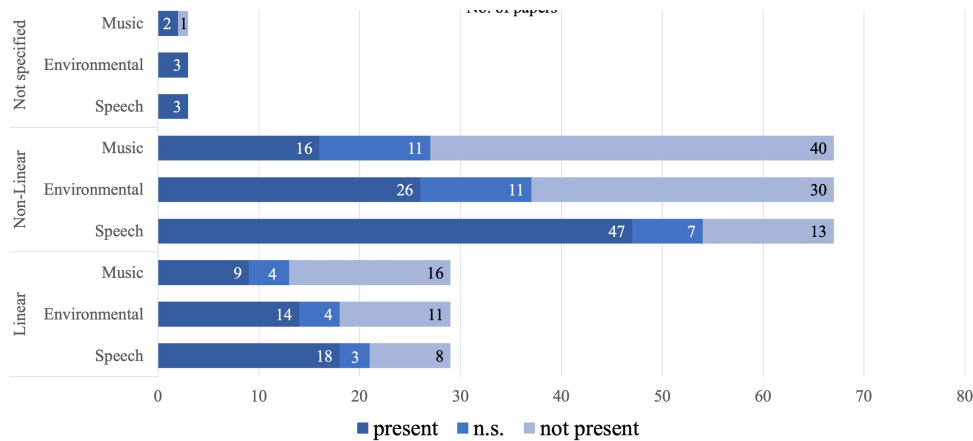


FIGURE 3.8: Distribution of articles by audio material/storytelling structure

Users with special needs

In reviewing papers, special attention has been paid to SIVE accessibility for cultural heritage experience users with special needs. Unfortunately, available solutions often provide limited accessibility for the unnatural or insufficient interaction methods proposed therein [175]. This is also noticeable in CAQDAS analysis that presents no concepts related to this theme (such as *special needs*, *impairment*, *accessibility*, etc.). Notwithstanding, users with disabilities, especially low vision, can benefit from audio technologies, and some contributions can be found in audio storytelling for cultural heritage as well. In the following paragraphs, we would like to mention five research projects that provide meaningful examples of the inclusive role of audio in such a context. *The Time Machine* [122, 168, 169] is a touristic guide that provides vibrotactile feedback to convey distance and directional information about interesting sites, so as to keep focus on the environment and help navigation of visually impaired people and elderly users. *MuSA* [175] is an application specifically designed for people with low vision, which reads information about art pieces in a museum and then renders it through AR. By receiving an image from a smartphone camera pointing to a specific artwork, the application recognizes various elements and then provides speech information as well as augments the image with colour optimizations and zooms on its particulars. *Greta* [159] is a mobile application that provides audio descriptions of films in cinemas specifically designed for people with low vision. The preliminary evaluation study shows that the use of voiceovers improves enjoyment and immersion but also affects engagement. *Trying to Get Trapped in the Past* [163] is a virtual drama whose narrative relies on spatial audio content, with visually impaired people in the sonic interaction designer's mind. In the recent work by Kelly [191], the author describes the early stage design of a sonification aimed at improving the accessibility of Irish cultural heritage sites.

3.4.3 Collaborative Experiences

Collaborative and social experiences are important, especially for improving the level of Immersion and Entanglement with both technology and the user's peers: in different cultural heritage subfields, such as music or multimedia production, collaboration elicitation among users is crucial [192]. CAQDAS analysis underlined that *social* is one of the key concepts found in the reviewed works. It is strictly connected

with *story* theme (*place, history, urban*), which emphasizes that when narratives occur in public places, users can benefit from collaboration, and with *experience* that together with the previous theme suggest the possibility of new *digital media* that can include collaborative narratives. In our review, 10 works consider collaborative and social aspects. The majority of them were ad hoc installations in which users can interact in real/virtual shared spaces with tangible interfaces [188], voice intensity and pitch [187], or movement [148]. In [131], social interaction is fostered by a particular physical room setup in which directional speakers mounted on the ceiling deliver personal information about a physical museum without isolating visitors from environmental sounds. Hättich and Schweizer underline the importance of cinemas in fostering sociality [159]. In *This Land AR* [193], three users can interact with virtual musical instruments implemented in mobile devices to create a collaborative audio performance. Moreover, storytelling can be efficiently designed for eliciting discussion during or after the experience as in *Traces* [194] artistic installation/touristic guide, where different complementary linear narratives are used to enhance discussion among different visitors.

Another remarkable point of view [195] states that the use of audio guides or other mediation devices during a cultural heritage visit has the twofold effect of enhancing learning performance and diminishing social interaction. This effect was particularly evident when wearable devices such as headphones were employed, facilitating the creation of personal virtual environments without any collaboration support.

Previous works suggest that the perception of a virtual space changes with the proposed experience, and can help achieve different specific purposes. A careful design of sonic interactions is a powerful tool for inducing collaboration in a virtual environment and has to be encouraged as a mean of enhancing immersion in shared virtual spaces and providing different communication channels.

The concept of proxemics, defined by psychologists as the set of implicit social rules of interpersonal distance among people, has well-known cognitive effects [196] and can help in describing immersive sonic experiences in shared spaces.

3.4.4 Quality of the Experience and Evaluation

Evaluation of different aspects is a fundamental task for creating an effective, immersive, and engaging experience. Fig. 3.6 shows the immersion type in the considered papers, holding that a wide variety of experimental protocols have been used to evaluate system usability, engagement, immersion, learnability, emotional content, cognitive load, and educational aspects, and most studies rely on questionnaires and semi-structured interviews. From the authors' point of view, the proposed user tests show an overall good acceptance level and engagement, although it is difficult to obtain quantitative information due to differences in the experimental procedure and measured features. CAQDAS analysis does not present any reference to the evaluation of user experience or interaction among the concepts represented, probably because of the heterogeneity of the measurement instruments and traits measured. However, some common traits can be identified in current research.

In order to test User Experience, validated questionnaires are used in many works. The NASA-TLX¹¹ questionnaire is used [122, 190, 169] to evaluate technology in terms of Mental and Physical Demand Performance, Effort, and Frustration and has been proven valuable to test user's technology acceptance levels. System Usability Scale (SUS, [197]) is a 10-item questionnaire and has been used [98, 175, 190,

¹¹ Available at <https://humansystems.arc.nasa.gov/groups/tlx/>

[161] for testing system usability. User Experience Questionnaire¹²(UEQ) is a 26-item questionnaire used [190, 167] to evaluate the attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty of the technology. Both questionnaires have been adopted in order to avoid fatigue in users while obtaining comparable and valid results. Simulation Sickness Questionnaire for Cybersickness (SSQC, [198]) has been used for evaluation in the different study conditions [167]. Narrative Engagement Scale (NES, [199]) is used [159] to evaluate the narrative engagement difference elicited by an audio description of films in people with low vision. In [137], User experience in a spatialized narration task is evaluated in terms of immersion and elicited emotions. Users were asked to navigate in an audio-virtual spatialised soundscape composed of several different virtual rooms while listening to a story by using a tangible controller with different degrees of control over movement, audio spatialisation, and audio playback. In the study, the Immersive Response Questionnaire (IMX¹³) was used to evaluate immersion and the Discrete Emotion Questionnaire (DEQ, [200]) to measure emotional content, along with Heart Rate with a non-intrusive device (a wearable band) showing an increased immersion and emotional content in conditions with spatial audio and controlled playback.

In addition to Geronazzo et al., other studies use biosignals for measuring emotional parameters, particularly arousal. Jurica et al. [201] compare Electro-Dermal Response (EDR) with a custom questionnaire for measuring arousal in soundscape navigation tasks inside a cultural heritage site, highlighting a better accuracy of the questionnaire for this specific task. On the other hand, Mansilla et al. [202] uses EDR to measure the emotional content of *Virtual Acousmatic Storytelling Environment* (VASE) virtual play, showing a correlation between EDR and emotional content increase. Biosignals are confirmed to be useful for emotion measurement and are suggested by the authors of this paper to be promising for use within tests employing auditory stimuli.

Moreover, MR technologies have been proven to enhance performance in the education domain applied to cultural heritage in various studies [203, 195, 179, 119, 121, 118, 204]. In the majority of these studies, a custom questionnaire designed by the researchers involved in the study has been proposed to participants before and after test sessions to evaluate the differences in terms of mechanical and meaningful information memorization for educational purposes.

3.5 General Discussion

Most of the reviewed studies in the cultural heritage domain are related to tourism and museums. In particular, audio guides are augmented with personalized content, non-linear storytelling, user localisation, and positional tracking to (i) propose an enhanced cultural heritage experience fitting the user preferences, (ii) elicit interest and engagement in users, and (iii) foster cultural heritage dissemination. In some applications, tourist information and street directions are shown, helping tourists orient themselves within a cultural site or city. In others, visitors were immersed in audio AR in which synthetic sonic environments aimed at conveying emotional engagement, re-creating historical events, or presenting position-based narratives.

Notwithstanding the vast amount of literature about narrative and storytelling, to our surprise, we found very little information about narrative content and structure. Practically speaking, the mandatory inclusion of all the identified clusters

¹²Available at <https://www.ueq-online.org/>

¹³Available at <https://innermetrix.com/products/assessments/>

(*storytelling, immersive media, audio, cultural heritage*) identified the intersection of all these only. Accordingly, this issue suggests the necessity to design new methods for a combined analysis of immersive and non-linear narrative experiences in interdisciplinary contexts. Since none of the reviewed papers explicitly considered a systematic approach to digital narratives, our work would also provide different perspectives for sonic interaction designers, practitioners, and technologists.

In our opinion, the criteria described in Sec. 3.1 are useful in defining a starting point for advanced sonic interactions in immersive storytelling. In the following, we suggest some promising research directions based on the previous guiding principles.

- **Perspective and Point of View.** To provide an immersive experience, most narratives should be designed using an egocentric audio approach [205], in which sonic interactions have to be defined in terms of personalized listener-virtual environment relations. The term egocentric refers to the perceptual reference system for the acquisition of multi-sensory information in immersive MR environments, as well as the sense of subjectivity and perceptual/cognitive individuality that shape the self, identity, or consciousness of the user. Such a level of personalization should avoid the mediating action of the immersive technology, which might result in a break in presence that can hardly be restored after a pause [206]. These cognitive illusions depend, for example, on the level of hearing training and familiarity with a stimulus/sound environment, and they should be taken into account from an egocentric perspective to create an immersive, coherent, and entangled experience.
- **Gift of voice.** When speech is employed, the narrative is usually held in the third person for informative purposes or city directions and in the first person for conveying emotional content, such as witness experiences or event reconstructions [120, 124]. Popp [136] suggests that the narrator's role can influence the trustworthiness of the conveyed information of the message. In the analyzed studies, no information can be found about the narrator register. However, we suggest that the narration register should be personalized according to the user's psychological state and cultural background to create a more immersive and user-specific experience.
- **Narrative, Dramatic Question, Emotional Content, and Power of Sound-track.** From the perspective of the conveyed message, various works [194, 201, 137, 207] are specifically designed for emotion elicitation. Music is mainly used in experience with artistic purposes and linear storytelling. Non-musical elements, such as speech and sound effects, are more common in reviewed works and are sometimes personalized or dynamically generated based on the user's position or interaction. Again, sonic interactions in virtual environments mediated by physical devices or GUIs help modify the users' internal psychological state during personal or collective situations, resulting in entangled experiences. Emphatic technologies have a pivotal role in such modulation, tracing fluid boundaries between humans and technologies [208], and eliciting internal emotional states following the user's expectations and emotions.
- **Economy, Pacing and Interactivity.** In many works, the total time and pacing of the experience depend on the interest and involvement of the user, who

seems to prefer a higher degree of control in terms of movement and interaction with the virtual environment and storytelling structure. Storytelling personalisation based on user interaction with the sonic or visual virtual environment, as well as non-linear narrative structure, provide promising evidence of eliciting users' interest in cultural heritage and enhancing dissemination and education [209, 178]. Such a complexity might be managed by AI agents, i.e., non-human entities capable of interacting with ecological behaviors [210], that would be able to predict the user's intentions in an intelligent environment for storytelling. Their ability to monitor listeners' behavioural responses could balance users' expectations and cognitive capabilities to adapt and modulate specific interactions and events [211]. More importantly, AI algorithms have the potential to encourage the exploration of different and meaningful paths within a non-linear narrative tailored to the user's needs. 1

- **Medium:** In the reviewed works, the immersive medium in use was clearly specified and the experience was usually designed considering its peculiarities. Based on the reality-virtuality continuum proposed by Milgram and Kishino [20], the level of isolation and the combination of real and synthetic elements can differ in audio mixed realities.

3.5.1 Sonic Interactions - Notable Mentions

In the realm of sonic interaction, tangible and haptic devices are widely employed, especially in non-linear storytelling experiences. Interaction within the virtual environment is used by the system to select a specific storyline and guide the user in the virtual or augmented world. To foster a more natural interaction, several works [167, 188, 131, 137, 160] have adopted custom-built tangible devices that are usually used as the main element during the interactive experience and are usually rated as engaging by users.

Despite the few works reviewed, we would like to mention eye tracking as another promising opportunity to (i) provide information in an ecological manner during interaction with the virtual environment (e.g., [190, 189]), (ii) make the experience accessible in the presence of users with physical impairments ([168]), (iii) augment the experience with multisensory elements such as the environmental sound of tactile feedback (e.g., [152]).

Focusing on interactive installations only, we extracted some interesting insights regarding socialization and collaboration aspects. With the help of particular audio setup configurations such as extra-aural headphones [132] or directional speakers [131], or by using standard headphones with custom sensors [148], researchers recreated a social experience taking place in an augmented shared space. From the technological point of view, it was difficult to retrieve specific information about the audio speaker/headphones setup and spatialisation models, especially in studies involving mobile phones and tablets. For the sake of repeatability in research, more technical audio specifications are thus needed to evaluate immersion and to create design guidelines for cultural heritage augmented audio storytelling platforms. This is even more relevant because studies including cognitive evaluations of emotional content, engagement, interest, usability, learnability, acceptability, cognitive load, etc., lead to better results for audio MR compared to static audio guides and onsite information panels [212, 179]. Educational aspects also benefit from augmented books and virtual AR exhibitions in different contexts: children's narratives, travel

guides, and interactive installations. In particular, children show interest in this specific technology, hence opening room for future research, even though ethical issues about technology addiction must be considered.

3.5.2 The role of personalisation

Section adapted from publication [213].

To personalise audio in MR, it is crucial to coordinate the customization of the technological infrastructure and the storytelling, which should be considered a complex entangled system to achieve an effective and comprehensive evaluation [19].

Personalization in Audio MR Technologies

Personalization of the acoustic environment can significantly enhance the experience in audio MR, particularly in terms of plausibility or the degree of congruence between the simulation and the listener's expectations [52]. Technological advancements should be closely correlated with other aspects, such as interaction design and content personalization, to elicit a coherent entanglement between humans, technological platforms, and interaction features. Considering specific contributions, an interesting approach in this direction is using individual HRTF instead of diffused generic dummy head HRTFs that, in conjunction with reverberation and head tracking, can significantly reduce localization error and enhance externalization. Various works, including [214], discuss computationally efficient methods for HRTF synthesis in real-time, starting from a limited set of anthropometric parameters. A technological platform specifically designed for personalisation will be presented in Chapter 4. Moreover, [16] introduces the idea of audio-augmented virtuality to evaluate the impact of various reverberation algorithms on co-immersion in VEs, highlighting a limited but significant effect of real-time digital reverberators over static spatialized sound.

Storytelling and Personalisation

As underlined by the CAQDAS analysis (*concept story, user and experience*), narrative structure (*narrative, history, landscape*) and the level of sonic interaction between user and audio virtual environment (*sense, experience, process*) are key aspects to consider when dealing with cultural heritage storytelling in mixed realities. The intersection between story and experience themes highlights that stories should be *immersive*, and the *context* is important for the narration. To enhance the sense of presence in MR, especially in the entanglement dimension, storytelling should be designed to be interactive and non-linear in order to allow active participation of both users and narrations eliciting higher engagement and immersion [215]. Non-linearities in narrative can be achieved by using different techniques such as inverting the chronological order of events, creating different parallel storylines, and dynamically modifying the narrative.

The majority of studies use non-linear storytelling (67 out of 92 papers) to let users freely explore the augmented physical environment or the virtual space, conversely, 35 papers follow a linear narrative. No specific information about storytelling architecture can be found in the three remaining papers. An overview of storytelling personalisation features is given in Table 3.3. A particularly relevant work [194] exploits a dual non-linear structure for encouraging a collaborative approach. As soon as they start the visit of the St. Fagan Historical Museum in Wales, users

What	How
Soundscape	User model detected by user position and orientation in space [173] User position and/or head orientation [100] Interaction with the virtual environment and AI techniques [144]
virtual environment and story	User interaction with tangible interface [187, 188, 183] Interaction with voice [188, 183]
Directions	User position and/or head orientation [122, 219, 168, 169, 186]
Storyline	Interaction with HMD headset controller [153, 139, 186, 143, 149, 150, 146] AI content generation [149, 150] Interaction with a tangible device [105]
Content created by users	User interaction with tangible interface [185] Collaborative experience [172, 193]

TABLE 3.3: Content personalisation and personalisation method in non-linear storytelling.

have to choose between two possible partial storylines. Information about the discarded storyline can be obtained only by asking and discussing it with other users. Moreover, artificial intelligence (AI) was employed [153] to adapt storytelling in a VR video game story and, in *Exhibot* [149, 150], to generate storylines about the history of the central square in Heraklion(Greece) based on user position and orientation and third party content services. Surprisingly, the previously cited works were the only contributions using AI techniques, although a large number of works can be found in other interactive and non-linear storytelling contexts. For instance, Riedl et al. [216] present a review of AI techniques in Computer games storytelling, Hernandez et al. [217] discusses the application of AI for eliciting emotions during storytelling, Pisoni et al. [218] introduces AI techniques for accessibility in cultural heritage including interactive storytelling.

Position and head/body orientation are often tracked and used to select a storyline. In this context, content can be pre-determined or personalized in different aspects (point of view in space, content presentation order, dynamically generated elements, and directions) by a user's movement or explicit command. City tour maps are dynamically generated [98, 168, 169] by using the user's navigation path, along with historical and touristic insights about buildings and monuments nearby and at the destination. In *Cave* [132], linear storytelling is spatialised according to the user's specific position inside the room. In the SARIM system [138], sound zones are associated with audio samples coherent with the specific exhibited art piece or historical device.

In terms of audio content, we identify three main high-level categories: (synthetic or real) speech, environmental sounds, and music (Fig. 3.8). Since voice is the most used material, especially in tourist or educational experiences, sound effects are used to create soundscapes and personalized content. In *Ec(h)o* [105], audio

soundscape and content are based on user preferences selected by choosing the orientation of a physical cube. In a work by Fu et al. [100], the audio soundscape is dynamically generated based on the user's position. On the other hand, music is used mainly in theatrical or linear storytelling experiences.

3.6 Research Directions

It is evident from our perspective that entanglement in human-computer interaction is highly desirable, as suggested by Frauenberger [5]. Mid-long-term guidelines should integrate technological aspects and content to pave the way for truly immersive storytelling. In particular, our theoretical framework suggests a binding role of the auditory component in its fluidity of perception-action within virtual environments through an ecological perspective [220]. This means that the listener might enact cultural heritage experiences in exploring the virtual environments [221]. Accordingly, we should consider an embodied, environmentally situated perceiver where sensory and motor processes are enabled by technology and inseparable from exploratory action in a narrative space. It is also important to note that in the field of immersive storytelling for cultural heritage, the new methods introduced by AI may open up new possibilities and design in terms of ease of use, involvement, and accessibility. To better understand the possibilities enabled by a proper design of SIVE, some potential research directions are illustrated below.

3.6.1 Emotional museum

In [222], Perry et al. argues that the state of the art of virtual museums doesn't take full advantage of all available possibilities in terms of interaction between participants, and emotional and personal development. An interesting scenario can be the use of audio cultural heritage storytelling to create a virtual museum or augmented cultural heritage site specifically designed to convey emotional content through the exploration of the environment. This kind of experience permits new forms of knowledge dissemination for educational purposes or entertainment. Specific guidelines for cultural heritage audio storytelling can be a useful starting point for the design of more effective storytelling that conveys emotional historical reconstruction in an effective and engaging way.

Challenges: Some historical and cultural experiences can contain a significant emotional impact. In this context, the SIVE framework can embrace the Research through Design (RtD) approach [223], helping in developing methods to identify the emotions to be expressed and to treat with respect and care for the arousal and respect of the users. Research through Design combines design practice and inquiry to better understand complex scenarios. It involves iteratively developing prototypes that allow the capture of emerging patterns. This approach recognizes that design practices are not just meant for implementation or creativity but also valuable methods for conducting research. Such practices could be a valid methodology for exploring interactions and possibilities with non-human agents, e.g., the virtual narrator [224]. The authors recently applied this research perspective in designing an AR audioguide for museums [18].

3.6.2 Immersive analytics

Especially with the new possibilities made available by Artificial Intelligence, immersive media, such as MR and audio MR can be used as a new methodology for research, knowledge analysis and decision-making [225]. *Immersive analytics* consists of using interaction with the immersive media to support analytical reasoning and decision making, providing multimodal interfaces allowing users to immerse themselves in data [226]. A similar approach is already employed in sonification, which consists in transforming data information through sound and auditory features in order to facilitate communication and interpretation. On the other hand, immersive analytics aim to enhance a bidirectional and entangled interaction between the user and the virtual environment [227].

In the literature, few tools [228] exist for immersive analytics of visual media only. However, cultural heritage research can benefit from the auditory component. Many musical sources are available, especially in fields related to audio and music, such as musicology and musical aesthetics but also by analyzing historical audio sources along with written documents. In this context, audio storytelling can be used twofold. Firstly, it helps researchers orient through the vast amount of audio material available by creating specific story paths, e.g., with particular themes or timelines based on historical facts. Secondly, implementing tools to create cultural heritage audio storytelling and use the design process as a reasoning method for knowledge discovery. Moreover, the use of AI can foster collaboration between the user and the virtual environment in an entangled experience [205].

Challenges: In this type of entangled experience, proper verification and processing of sources are crucial to avoid any misleading storylines that may be generated or introduced by interaction with the virtual environment and AI's hallucinations [229]. Moreover, integrating diverse sources of audio data, as well as designing a convincing interactive experience in MR, can be complex and require robust data management and the integration of multiple 3D user interactions.

3.6.3 Archaeoacoustics, virtual musical instruments, historical voices, and personalities

The analysis of the acoustic of ancient historical places is used by many cultural heritage researchers to better understand different historical aspects of everyday living¹⁴. In a work by Ciaburro et al. [230], acoustics properties (reverberation time) of an Ancient Roman catacomb have been measured and analyzed in order to better understand the reason for the choice of the specific prayer space by the ancient population. Similarly, Warusfel and Emerit [231] used available historical documents to simulate the acoustics of an ancient Egyptian temple of Dendara dedicated to Hathor, the goddess of music, love, and joy in order to investigate the role of sound in worship ceremonies. Moreover, the reconstruction of virtual ancient or modern musical instruments is widely used to study the main traits of instrument sound throughout history [232]. Again, in this scenario, audio storytelling can help reconstruct historical events and ancient spaces in virtual spaces with natural acoustics and sound for research purposes, education and entertainment, e.g., in museums and cultural heritage sites. Historical talks and documents played or read immersed in a virtual/augmented reconstructed audio environment could help understand

¹⁴<https://pro.europeana.eu/post/discover-the-potential-of-3d-and-heritage-acoustics>

the role of space in history and better comprehend the emotions elicited by historical figures.

Challenges: This type of scenario requires advanced simulation and reconstruction techniques that rely on technology that may not always be accessible or sufficiently advanced to represent historical acoustic environments accurately. Especially in VR and AR experiences, this can affect the correct representation and perception of spaces (e.g., positions and distances [65]), limiting or altering the planned experience. Considering the evolution of sound environments and listening habits in historical/architectural reconstructions is essential. For example, in the given location, external noises such as car traffic not present at the reconstructed period, could affect the perception of the experience. Another example is the modification of the tuning of musical instruments, such as that introduced by J.S. Bach in the 17th century [233]. Using the previous configuration, which was conventional for listeners of the time, might sound out of tune to users.

3.6.4 Universal fruition, multimodality

Multimodal interaction is widely used in museums and cultural heritage sites for entertainment and education but often has few or no audio elements [234], an important feature in many experiences designed for fruition by users with special needs [235]. It is worthwhile to notice that audio storytelling has been proven to be an important accessibility medium for people with disabilities. Visually impaired people highly benefit from visiting museums and cultural heritage sites for historical and touristic information access and urban navigation [236]. Moreover, storytelling was employed in connection with sound spatialisation and 3D elements in the therapy and rehabilitation of cognitively impaired children for facilitating child-therapist interaction [237]. Audio storytelling can deeply help users with physical disabilities but also users with cognitive disabilities in cultural heritage knowledge fruition, improving attention and engagement [238].

Challenges: The addition of too many multimodal elements could confuse the user with special needs and make it complicated to maintain attentional focus on the topic [87]. Moreover, ensuring the sustainability of multimodal interaction systems involves continuous maintenance, content updates, and technical support can be challenging. This may be problematic when moving from prototyping to implementation and ultimately to maintenance due to limited budget and organisational and political issues.

3.7 Narrami as a case study

Section adapted from publication [10].

To support and better explain the the previously described research scenarios, we introduce a case study on a purely hypothetical scenario that allows us to consider audio storytelling in cultural heritage egocentric audio technologies and their various facets.

Narrami is a cyber-physical system aimed at museums and exhibitions. A museum visitor, equipped with sensorized headphones and a smartphone, enters an exhibition eager to delve into the world of art. The visitor starts the visit moving toward an art piece is interested in. As the visitor moves through the exhibition,

the system tailors a personalized journey through the museum, focusing on art pieces that match their preferences. By using a reinforcement learning algorithm, the system analyses user movements and behaviors inside the museum. In this scenario, the focus of egocentric audio technologies encourages users to discover new, engaging paths inside the museum. The system offers an immersive audio-AR experience. With the help of nonlinear storytelling [215], the visitor can hear a guiding voice and the authors of art pieces discussing his/her masterpieces inside the room based on his position as in [98]. Using different metrics, the system has the ability to learn the user's interests based on the amount of time visitors listen to the narration before moving, their head rotations towards specific points of interest, or eye fixation detected by an eye tracker. It continuously refines its recommendations and adapts its behavior according to visitors' decisions and reactions during the visit. In particular, similar to [239], the proxemics between the user and a virtual agent, i.e., the narrator's voice in our scenario, can strengthen the subtle guidance of the MAS towards the suggested path. Hence, the virtual narrator's movements in the virtual space will be entangled with user behaviours. To provide a plausible listening experience, *Narrami* uses effective auralization of sound sources with perceptually-motivated algorithms that work in a real-time environment, providing a responsive environment for sonic interactions [25].

3.8 Conclusion

This chapter presented a systematic review of the role of audio in Mixed Reality storytelling within Cultural Heritage, highlighting its contribution to narrative construction, user interaction, and immersive experience design. The analysis revealed that, although audio is often treated as a secondary modality, it plays a crucial role in shaping spatial understanding, emotional engagement, and accessibility.

The reviewed literature also showed a growing interest in spatial audio technologies, multimodal interaction, and personalised storytelling, while still lacking a unified framework capable of integrating these dimensions into a coherent design and evaluation approach.

In this context, the M-I-C-E framework introduced in Chapter 2 provides a structured lens through which these findings can be interpreted. In particular, the reviewed works can be reinterpreted in terms of immersion (through spatial audio rendering), coherence (through perceptual and narrative plausibility), and entanglement (through adaptive and interactive behaviors between users and systems).

The introduction of *Narrami* as a conceptual case study, describe more in depth in Chapter 6, anticipates how these dimensions may converge within a unified Audio MR experience, bridging the gap between theoretical analysis and design practice.

Despite the methodological rigor of the review, some limitations should be acknowledged. The selection process, although systematic, may have excluded relevant works outside the chosen databases or keywords, and the heterogeneity of the included studies limits direct comparability across results.

The insights derived from this chapter inform the technological and methodological developments presented in the following chapters. In particular, Chapter 4 focuses on the implementation of spatial audio configurations, while Chapters

5 and 6 operationalise the M-I-C-E framework through experimental evaluation and real-world interaction scenarios.

Chapter 4

Technological Configurations: Spatialisation in Audio Mixed Reality

Chapter adapted from publication [14].

To support the experimental activities of this work, a dedicated auralization platform was developed and deployed on an Apple iPad using the Unity game engine¹. The platform was designed to be highly versatile, making it adaptable to different acoustic and spatial scenarios, although in this study its implementation focused on modeling a virtual corridor reproducing a real-world environment located at the Department of Management and Engineering of the University of Padua. This corridor was reconstructed in its dimensions (width: 19.70m, depth: 2.32m, height: 3.38 m) and its acoustic features (materials and reverb decay time) using accurate physical measurements, thereby ensuring that the virtual environment accurately mirrored the physical space and could serve as a reliable basis for perceptual testing and analysis. By embedding the model in an augmented virtual VE, the system provided participants with an immersive and interactive auditory experience that tries to ecologically integrate virtual elements within the real world. The implementation of the library is publicly available at: <https://github.com/aleprivi/Plausibility-in-AAR-Experiment>.

A central element of the platform was the reverberation algorithm, which simulated the persistence of sound in an enclosed space due to multiple reflections and diffusions from walls, ceiling, floor, and other surfaces. Reverberation is a critical factor in acoustic perception, as it conveys spatial information about the environment and contributes significantly to the sense of presence in virtual and augmented reality systems. In this work, reverberation was modeled using the SDN, an efficient method for simulating Room Impulse Response (RIR) that ensures both computational efficiency and perceptual plausibility [240]. The SDN achieves this by distributing delay lines and scattering junctions across a network that mimics the geometry of the room, resulting in a physically grounded yet real-time capable approach. The use of SDN allowed the system to capture fine details of reverberant behavior, providing listeners with cues necessary for accurate spatial orientation and environmental awareness.

In addition to reverberation, spatial realism was further enhanced by incorporating HRTF. An HRTF describes how an incoming sound wave is filtered by the listener's anatomical features, including the head, torso, and pinnae, before reaching the eardrum [241]. These filters encode the direction-dependent alterations of the sound spectrum, thereby allowing binaural systems to reproduce

¹Available at <https://unity.com/>

three-dimensional auditory perception [242]. Moreover, generic HRTFs often fail to capture individual anatomical differences, which can lead to errors in localization, front-back confusion, and reduced spatial presence [243]. To address these limitations, this work adopts a personalization technique [214], which adapts the HRTFs to the individual listener's anatomical features, thereby improving spatial accuracy and immersion. In this work, a *dynamic* HRTF implementation was employed to continuously adapt the orientation of the virtual user according to the real user's head and body position in real time. This dynamic updating is fundamental for creating a convincing sense of immersion. Motion and orientation data were collected from the iPad sensors and Apple AirPods Pro, and vendor-provided libraries ensured accurate head-tracking performance [244]. By coupling sensor-based tracking with personalized dynamic HRTF processing, the platform achieved a responsive binaural rendering pipeline where perceived sound directions shifted naturally with the user's movements.

To further refine auditory fidelity, the platform included a stage of headphone equalization compensation. This process corrects the specific frequency response deviations of the earphones, ensuring that the reproduced sound matches the intended stimulus and equalisation applied by intrinsic device-specific resonances or limitations [245] is reduced. Without compensation, spectral distortions could interfere with spatial perception and impair the ecological validity of the experiment. The earphones adopted during the experimental sessions also provided *acoustic transparency*, allowing selected portions of the external acoustic environment to pass through, preventing complete auditory isolation and supporting more naturalistic conditions [246]. Transparency helps participants to avoid a different perception between the real and mediated environments. An important aspect to highlight in the context of AAR concerns the issue of acoustic transparency in headphones and headsets, which can significantly influence perceptual plausibility. These devices inevitably affect the way external sounds are perceived, as they modify the acoustic signal before it reaches the eardrum, altering both its intensity and spectral characteristics. As a result, listeners may experience an unnatural auditory condition that compromises the authenticity of environmental sounds and consequently impacts the overall perception of the acoustic scene [247].

Finally, the auditory stimuli were carefully calibrated in terms of playback volume. Calibration involved matching the virtual sound levels to real-world reference values reported in the literature. This step was essential to guarantee perceptual consistency and to avoid discrepancies between the perceived loudness of virtual stimuli and their real-world counterparts. By aligning the volume to documented acoustic references, the platform ensured that participants experienced an ecologically valid sound environment, thereby reinforcing the reliability of the experimental results. In the following sections, the fundamental blocks of the framework are described in detail, highlighting their individual contributions and integration within the overall system.

4.1 Room acoustics: Reverb

The SDN algorithm integrates delay networks and acoustic artificial reverberators, yielding a physically and perceptually accurate acoustic simulation [240, 214]. The SDN algorithm employs the image-source method [248] to render the direct sound path (source-to-listener) and 1st-order reflections in both time and

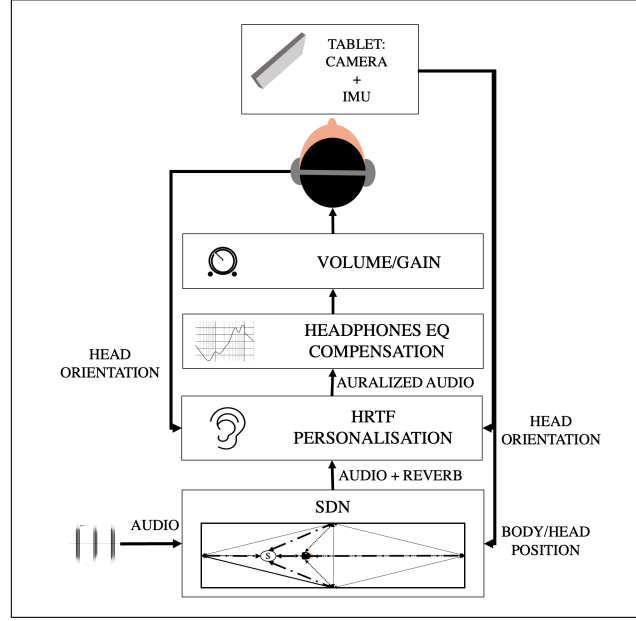


FIGURE 4.1: Schematic representation of the auralisation framework.

space. The following equation can express the computed RIR:

$$R(z) = \frac{1}{K} \mathbf{k}_M^T(z) \left[\bar{\mathbf{A}}^T \bar{\mathbf{H}}^{-1}(z) - \mathbf{P} \mathbf{D}_f(z) \right]^1 \mathbf{k}_s(z) + \bar{g} z^{-D_{s,M}} \quad (4.1)$$

where $\mathbf{k}_M^T(z) = \mathbf{\Gamma}_M^T \mathbf{D}_M(z) \mathbf{G}_M \mathbf{W}$ is the combination of the block diagonal weight matrix for pressure extraction, the source directivity vector, the delay matrix, and the attenuation matrix; $\mathbf{k}_s(z) = \mathbf{U} \mathbf{G}_S(z) \mathbf{D}_S(z) \mathbf{\Gamma}_S$ is the block diagonal matrix that assigns source signals to input wave variables and listener. $\bar{\mathbf{H}}(z)$ is the wall absorption matrix, $\mathbf{D}_f$ is the inter-node delay matrix, $\bar{\mathbf{A}}$ is the scattering matrix, and $\mathbf{P}$ is a network topology-dependent permutation matrix. Moreover, HRTF filters can be easily incorporated to produce a fully customized binaural room impulse response.

To correctly reproduce room acoustics, the RT30 decay time of the room was measured with a Sound Level Meter (Brüel & Kjær 7830) with the setup configuration displayed in Fig. 4.2-a. The RT30 of the virtual room was matched by visual inspection, varying two SDN parameters:

1. absorption material: corresponding to a 4th order FIR filter, $H(z)$ designed using [241] and selected by considering wall materials of the real room,
2. reflection coefficients: a manual gain that can be used to adapt the theoretical filter to the real material, independently for each wall.

Table 4.1 displays the final values. The average RT30 value is 1.38 s in the real corridor 1.27 s in the VE. Fig. 4.3 shows the RT30 decay time of the real and virtual room for each frequency. Notwithstanding those real and virtual environments present similar average decay times, values per band present differences that might influence spatial perception. It is important to note that while many methods have been suggested in literature to compute real room acoustics and absorption filter, the task is still considered complex to solve [249]. The observed

Wall	Material	Refl. Coeff.
Back,Front, Left	gypsum	0.9
Right	glass	0.9
Floor	concrete	0.98
Ceiling	concrete	0.99

TABLE 4.1: Absorption materials and reflection coefficients in SDN simulation.

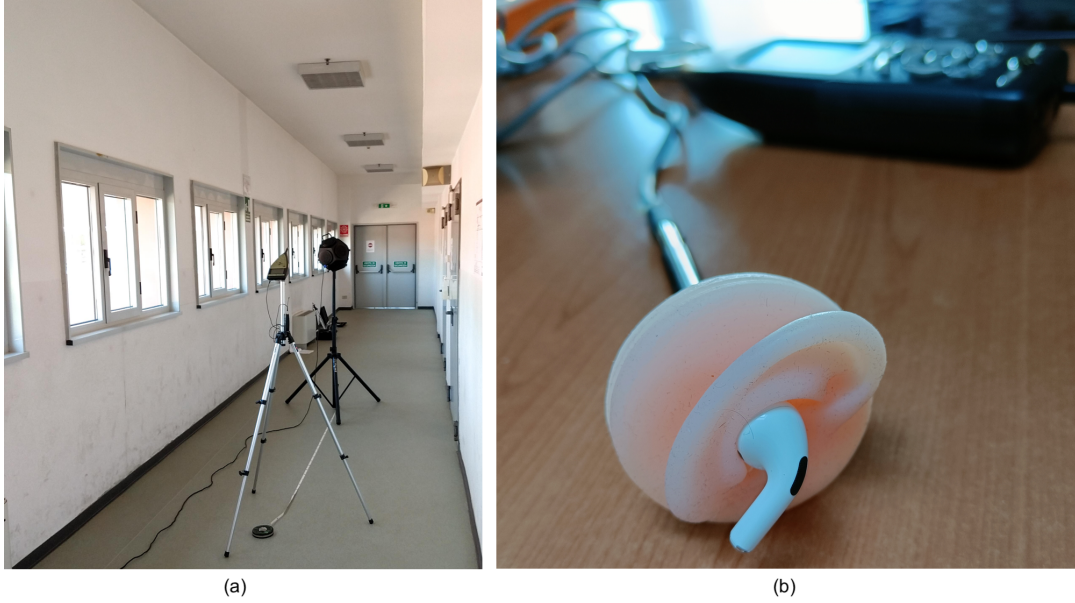


FIGURE 4.2: Calibration setup for (a) the real room's RT30, and (b) the headphone playback loudness.

discrepancies between measured and simulated reverberation responses, particularly visible in Fig. 4.3, may also partially derive from the limited modal density of the adopted SDN configuration.

A relatively compact feedback network may introduce spectral irregularities and reduced echo density, potentially contributing to perceptual artifacts such as metallic reverberation, as reported by some participants during the evaluation. In the present implementation, a limited number of delay lines was intentionally adopted to maintain computational efficiency and ensure real-time performance within the audio MR framework. This design choice reflects a trade-off between acoustic realism and system responsiveness, particularly relevant for mobile and interactive applications.

Future implementations may benefit from larger scattering matrices or hybrid late-reverberation architectures to improve temporal diffusion and spectral smoothness.

In the SDN implementation used in this work, during the Unity game cycle (*void Update*), the algorithm detects the distance and direction from the virtual audio sources to every wall in the virtual room in order to correctly compute audio delays of reflected sounds and apply the correct absorption filter.

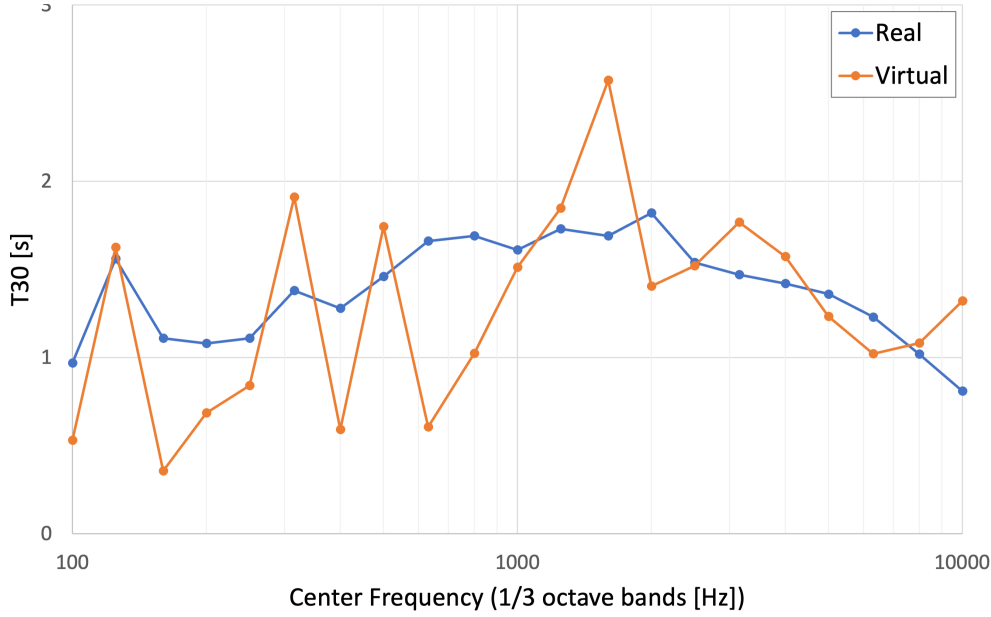


FIGURE 4.3: RT30 per 1/3 octave bands for real vs. virtual room.

4.2 User Acoustics: HRTF and personalisation

Using individual HRTFs, instead of the generic dummy head HRTFs, combined with reverberation and head tracking, can greatly reduce localization error and improve externalization [214]. However, measuring individual HRTFs is time-consuming and requires specialized facilities. Our study employs mixed structural modeling proposed and evaluated by Geronazzo et al. [214] to balance optimization and efficiency in HRTF representation. This model requires a minimal set of anthropometric data (Fig. 4.4), including the ear contour (C1) that outlines the helix, head width (X1), and head depth (X3). Using X1 and X3, low-frequency content (up to 1 kHz) can be modeled using a spherical head model with ear displacement [250]. Assuming diametrically opposed ears, the spherical Interaural Time Difference (ITD), one of the primary cues available to the auditory system for determining the spatial location of sound sources [251], in the horizontal plane is estimated as:

$$ITD(\gamma, a) = \begin{cases} \frac{a}{c} (\sin \gamma + \gamma) & \text{if } 0 \leq |\gamma| < \frac{\pi}{2} \\ \frac{a}{c} (\pi - \gamma + \sin \gamma) & \text{if } \frac{\pi}{2} \leq |\gamma| < \pi \end{cases} \quad (4.2)$$

where c is the speed of sound, the head radius a is given by $a = (0.41 \frac{X1}{2} + 0.22 \frac{X3}{2} + 3.7)$, and γ is the angle between the source vector $\vec{S}$ and the ear vector $\vec{e}$ with origin in the center of the sphere that is displaced relative to the interaural axis to $x_0 = [0, e_b, e_d]$ with e_d/e_b the downwards/frontwards ear shift, by using:

$$\gamma = \cos^{-1} \left[\frac{(\vec{S} - \vec{x}_0) \cdot (\vec{e} - \vec{x}_0)}{\|\vec{S} - \vec{x}_0\| \|\vec{e} - \vec{x}_0\|} \right]. \quad (4.3)$$

On the other hand, high-frequency spectral content is selected from the best-fit HRTF included in CIPIC [252] dataset according to anthropometric characteristics. For each participant, $n = 15$ ear contours C1 and $k = 20$ ear canal positions were manually annotated by using a Matlab script available at <https://github.com/geronazzo/hrtf-personalisation>:

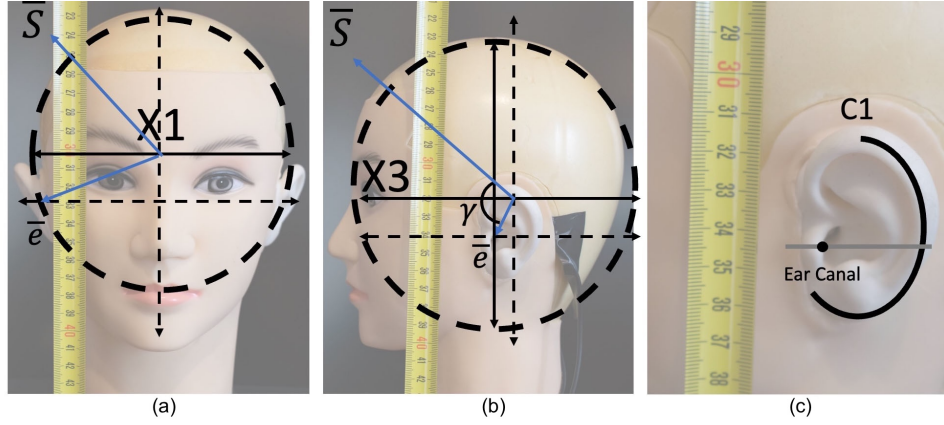


FIGURE 4.4: Head's pictures as taken for the experiment's participants to personalize the HRTF with the measuring tape used as a reference. The spherical model parameters are shown: (a) head width $X1$, (b) head depth $X3$, (c) pinna contour $C1$, and ear canal.

[//github.com/msmhrtf/personalization2020_app](https://github.com/msmhrtf/personalization2020_app) and three photos of the subject (front and side views and ear closeup) with a measuring tape as a reference. Based on these annotations, we estimated pinna notch frequencies f_0 for each elevation angle ϕ [253]. Such notch frequencies f_0 are compared with the corresponding frequencies F_0 extracted from the measured HRTFs from CIPIC employing a mismatch metric:

$$m_{(k,n)} = \frac{1}{N_\phi} \sum_{\phi} \frac{|f_0^{(k,n)}(\phi) - F_0(\phi)|}{F_0(\phi)}, \quad (4.4)$$

given N_ϕ the number of elevation angles. The best matching HRTF from the CIPIC database regarding Eq. 4.4 was selected. Low and high-frequency contents are combined with a linear crossfade at 1Khz.

In the SDN implementation the built-in audio thread is accessed by *void OnAudioFilterRead*, and dynamic HRTF filters are applied to each reflection based on user position and orientation.

4.3 Volume and EQ matching

To correctly perceive the source position inside the virtual space and thus obtain a high level of realism in the scene, the sound sample should have a Sound Pressure Level (SPL) equal to a natural listening condition without headphones. When possible stimulus SPL, in all conditions, was matched in order to obtain the same Sound Level of a real reference. For the volume calibration, the virtual corridor was created with the physical characteristics obtained by the room calibration procedure. Inside the VE, a virtual source was positioned, and a fake ear (Fig. 4.2-b) wearing AirPods simulated the headphone placement over a human ear. By using a sound level meter, the volume was calibrated by adjusting the gain in the simulation. To reduce the coloration caused by headphones audio impulses was also calibrated using an equalization curve consisting in the inverse of the earpods transfer function shown, in Fig. 4.5. Impulse file available at <https://github.com/jaakkopasanen/AutoEq>.

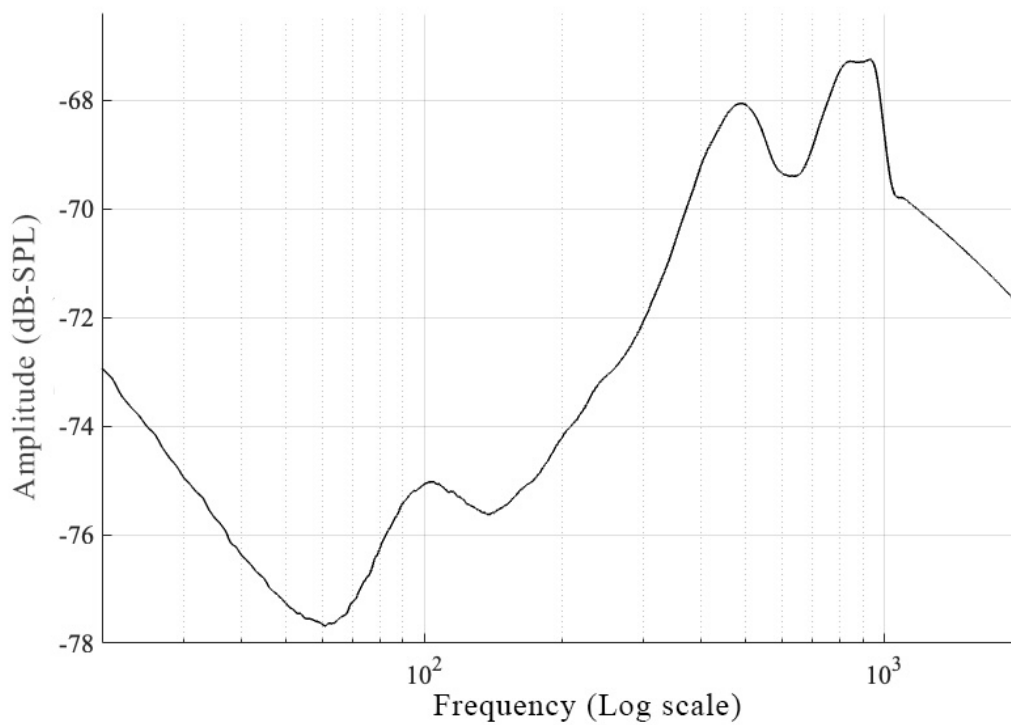


FIGURE 4.5: Headphone compensation.

4.4 Tracking

Section adapted from publication [15].

The proposed evaluation considers different tracking methods to find the most reliable in terms of the user's perceived source position in space.

A correct spatial audio reproduction requires tracking the user's position in space and head orientation. In the adopted platform, we implemented different user tracking and head orientation method made available by ARFoundation², an AR Middleware inside Unity that permits the extraction of those relevant features that allow a correct AR interaction, e.g., device and objects position estimation inside a room, face tracking, object occlusion detection, etc. By integrating API available for different proprietary platforms (ARKit, ARCore, Magic Leap, WindowsXR), ARFoundation can be used on different devices and Operating systems. For the purpose of user indoor tracking, the study employed three different methods:

- **iPad Tracking (T):** ARFramework permits automatically estimating the position of the iPad space. At the start of an experiment session, the virtual position of the device is set to world origin and updated when moved, with respect to the origin. In this tracking condition, the user head position is set as iPad position.
- **Face Tracking (F):** using the Front camera, the head center can be estimated with respect to the iPad position. Utilizing both cameras makes it possible to

²<https://unity.com/unity/features/arfoundation>

simultaneously estimate the position of the user's head and iPad. The combination of the two sets of coordinates can provide the position of the head with respect to the reference origin.

- **iPad Tracking with Height Displacement (H):** head position is set to iPad estimated position with a displacement of 50 cm in height (y-axis), in order to simulate head position with respect to iPad.

Furthermore, the head rotation has been estimated with the following three methods:

- **Earbuds (E):** Apple AirPods pro contains IMU sensors that can be accessed by Unity by using a third-party script³ and can be used to track head rotation, with respect to a starting point. The manufacturer doesn't provide any official technical specifications about the sensor used.
- **Face Tracking (F):** similarly to head position tracking, ARFoundation and the front camera have been used to estimate head rotation with respect to the head center.
- **iPad Tracking (T):** head rotation has been set as iPad rotation with respect to the reference origin.

Depending on the implemented experience, a different tracking method was selected.

4.4.1 Tracking accuracy

To objectively evaluate tracking accuracy, measurements have been acquired inside the same room, lit by natural light from large windows on one of the longer sides, replacing human participants with a dummy head in controlled conditions.

Measurement Setup

The iPad had been positioned on a bookstand allowing different placements inside the room with an inclination of 30° to simulate iPad user handling. Camera height was set at 1.30 m and positioned at the center of the bookstand since the estimated device position was referred to the camera position. To simulate a human head, a dummy head has been mounted on a tripod at 1.80 m. On the floor, a 1.5 m × 1.5 m grid was marked with 16 points, each with a distance of 50 cm from the nearest. Position tracking T and F, described in section 4.4 have been implemented in a Unity application for detecting both iPad and user head center, that record the device position in space (x, y, z coordinates, rotation, and azimuth with respect to initial position) and head position (x, y, z coordinates) relative to the device coordinates. According to the Unity reference system, the floor is parallel to plane x-z while the height to y-axis.

Procedure

Starting from the corner points, the experimenter positioned the iPad stand on the first grid point, orienting the rear camera system to the nearest long wall. In front of the other camera, the dummy head was put in the grid point at a 0.5 m

³Available at <https://github.com/anastasiadevana/HeadphoneMotion>.

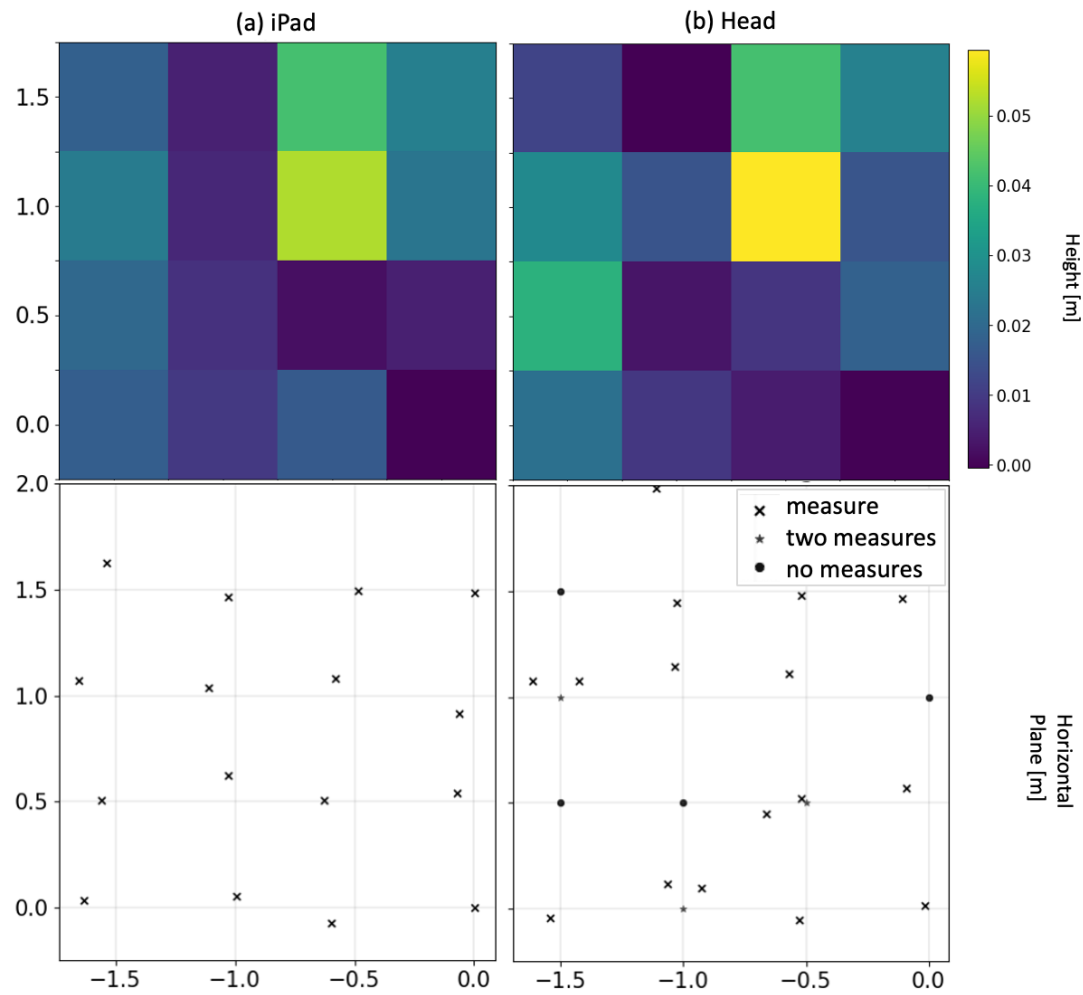


FIGURE 4.6: Position tracking accuracy: measured position of (a) iPad and (b) dummy head. Position tracking error (top). Mean measured points in the horizontal plane for each position (bottom).

distance from the iPad. After checking head recognition by the Face Tracking algorithm, both iPad and dummy head coordinates were recorded for 20 s at the system sample rate (60 Hz). Once the recording stopped, the iPad stand was moved to the nearest grid point at its left, its orientation varied by 90° , and the dummy head was accordingly placed. If not enough space was available to put the dummy head in front of the iPad, it was rotated by 90° until in order to reach a suitable position for the dummy head.

Results

Figure 4.6 shows iPad and face position tracking measurement errors. Since the grid was not completely covered by head measurements, points marked with a $\star$ have two measurements, and points marked with a $\bullet$ have no measurements. iPad position tracking mean error was overall 8.99 ± 3.92 cm (-3.71 ± 5.66 cm width, 2.15 ± 6.24 cm depth, 2.33 ± 1.26 cm height). Head position tracking mean error was 9.71 ± 4.34 cm (3.01 ± 7.08 cm width, 2.15 ± 6.69 cm depth, 2.33 ± 1.26 cm height). A similar test in [254] showed that commercial devices specifically built for VR, such as Oculus Quest 2 and HTC Vive Pro, have a lower error in position tracking ($<1.0 \pm 2.0$ cm). The iPad and AirPods Pro solution for tracking has been preferred here to avoid the not negligible acoustic impact on localization of the real sound sources by VR/AR HMDs [255].

4.5 Discussion

This chapter presented the design and implementation of an audio augmented reality auralization platform, integrating spatial audio rendering, HRTF personalization, dynamic reverberation, and user tracking. While these components have been described from a technological perspective, their role should be interpreted in relation to the broader objectives of this thesis.

In particular, the platform is designed as an enabling infrastructure for investigating user–technology interaction within Audio MR scenarios. The design choices presented here reflect a balance between acoustic accuracy, computational efficiency, and ecological validity, supporting real-time interaction in mobile and context-aware environments.

Although the experimental motivations are fully articulated in the following chapters, it is important to clarify that the technological configurations introduced in this chapter directly support the evaluation of Immersion, Coherence, and Entanglement as defined in the M-I-C-E framework. Specifically, spatialisation and reverberation contribute to immersion, tracking accuracy and audio consistency support coherence, while adaptive system behaviors enable forms of entanglement between users and virtual agents.

From this perspective, the apparent separation between system implementation and experimental investigation reflects a methodological choice: first establishing a controllable and reproducible technological baseline, and subsequently employing it to explore interaction dynamics in ecological conditions.

The following chapters build upon this platform to investigate how variations in rendering, tracking, and interaction design influence user perception and behavior. Chapter 5 focuses on the evaluation of immersion and coherence through controlled experiments, while Chapter 6 extends the analysis to entanglement within a real-world interactive scenario.

Chapter 5

Immersion and Coherence

As defined in Chapter 2, Immersion is intrinsically linked to the technological mediator and its capacity to evoke a sensation of presence, primarily concerning the relationship between the "entangled twins" within a specific technological setup. Coherence is the determinant of the overall credibility of the hybrid analog-digital scenario, relying on internal logical consistency and behavioral plausibility to sustain the illusion of the virtual world.

Immersion and coherence are strictly interconnected dimensions [51] and can be evaluated considering many other related dimensions, such as the system's enactive potential and the user's reconfigurability within the DT.

5.1 Background: authenticity, plausibility and transfer-plausibility

An important aspect related to Immersion and Coherence is the realism of the MR experience. Several criteria have been proposed to evaluate whether a simulated sound source is perceived as realistic. Among them, *authenticity* [256] requires the perceptual identity of the simulation with a real external reference. A virtual sound source is authentic if indistinguishable from a corresponding real one in a direct comparison. Authenticity is the most strict criterion since even small differences can be noticed by listeners. Brinkmann et al. [23] investigated the authenticity of individual dynamic binaural simulations in an ABX test. Real stimuli were reproduced with loudspeakers, while virtual sources were simulated through headphones. Results varied based on stimulus—speech was less authentic than noise—and environment, but not source position.

Only a few AAR applications require the simulation's authenticity, thus less demanding evaluation criteria exist. One of them is *plausibility*, i.e. the simulation's agreement with the listener's expectation. Plausibility relies on the listener's internal representation of reality, without an external reference as a comparison. Lindau and Weinzierl [257] proposed an experimental procedure to assess plausibility in VE s by means of a 2AFC test with "real" and "simulation" as choices. Participants listened to random trials of real stimuli and non-individual dynamic binaural simulations of them. Using SDT analysis, authors found a slight sensory difference between simulation and reality, though it was insufficient to represent a meaningful effect. A similar experiment was conducted by Pike et al. [258] to evaluate the plausibility of simulated stimuli with SDT.

Neidhardt et al. [259] investigated the influence of different BRIR simplification approaches on the plausibility in an interactive approaching motion towards a virtual loudspeaker rendered with dynamic binaural synthesis. Neidhardt and Zerlik [260] conducted a two-part plausibility experiment in 6 DoF. In the first

part, where all stimuli were simulated, inexperienced listeners tend to accept virtual simulation as real, while experienced listeners' responses were equally distributed. In the second part, similar to the experiment of Lindau and Weinzierl [257], results varied for inexperienced listeners, while experienced ones could reliably detect the simulation. Several more studies investigated the plausibility of simplified binaural rendering approaches [261]. Some of them employed VR environments [262] and analyzed the role of visual cues, too [263, 264]. Also, the authenticity and plausibility of different HRTF recording methods were assessed [265].

Between the strict authenticity based on direct comparison and the less demanding plausibility relying only on the listener's expectation, there is another common scenario in AAR applications. In this scenario, real and virtual sound sources simultaneously reproduce different but similar stimuli. Thus, such a scenario allows a comparison of the virtual sound source with real ones but it prevents the direct comparison with the real counterpart. Two worth citing criteria for this scenario are *co-immersion* [266] and *transfer-plausibility* [267]. In a scene with existing sources, both these criteria evaluate the immersion of a simulated source without being perceived as such. Co-immersion and transfer-plausibility differ in their application. While transfer-plausibility is defined only for pure AAR scenes, co-immersion also encompasses fully artificial scenes where a sound source should be immersed among existing virtual sources by matching their acoustic features. Wirlor et al. [267] evaluated the transfer-plausibility in a room with up to eight loudspeakers. None or one of them was simulated with dynamic non-individualized binaural synthesis.

A fundamental aspect of plausibility is the degree of coherence in the sonic interaction design for each user within an augmented environment. The concept of *proxemics*, i.e., how human interactions and behaviors shape the perceived space, is particularly interesting within theSIVE framework. This is relevant especially when the co-presence with other real or virtual objects or agents influences the experience [268]. Williamson et al. compared two audio designs and two displays with different social signal affordances in a virtual meeting experience using Mozilla Hubs¹. They claimed that *digital proxemics*, that is, proxemics in a VE, depends on the technology used and the interaction design choices. In such a context, user tracking is crucial for a correct distance representation and perception and can strongly influence a coherent sonic experience.

Finally, it is worthwhile to mention a peculiar issue in AAR that can heavily affect plausibility: the degree of acoustic transparency of headphones or headsets in general. These devices in fact can alter the perception of environmental sounds by modifying the acoustic signal in terms of intensity and color before reaching the eardrum. This unnatural listening condition can consequently alter the overall environment perception, making real audio sources far to be authentic from an acoustic standpoint [247]. In [22], Schlieper et al. measured the HRTF of a KEMAR head wearing headphones, using 34 different commercial headphone models of different types, including over-ear, on-ear, in-ear, and earplug headphones, with and without active noise control and hear-through functionality. The study demonstrated that every headphone model and configuration introduces a certain amount of distortion in both equalization and magnitude compared to the reference HRTF, measured when the KEMAR mannequin wore no headphones.

¹<https://hubs.mozilla.com/>

5.2 The importance of tracking and distance perception: tracking accuracy

Section adapted from publication [15].

Correctly tracking the user's position and orientation in real space is a critical building block of AR applications, essential for creating a point of reference for a realistic user experience [269]. Nowadays, many commercial mobile devices such as smartphones and tablets contain the hardware/software technology for tracking the user position and orientation in space. However, such proprietary technologies often provide limited access to their control parameters. For this reason, adopting these devices requires trade among factors such as accuracy, cost, portability, and acceptability [270].

Available tracking technologies include IMU sensors, vision-based methods, marker or markerless approaches, and hybrid approaches called Simultaneous Localisation and Mapping (SLAM), combining camera information with inertial sensors [269]. In particular, SLAM is used by the vast majority of commercial devices, such as Head-Mounted Displays (HMD), smartphones, or wearable devices that combine camera information with inertial sensor data. Moreover, user tracking can be implemented by employing eye-tracking to provide gaze-guided audio-visual information in a controlled laboratory environment [271] or GPS along with compass sensors in large spaces such as parks or city centers [269].

Various studies focused on the role of audio-visual cues in indoor [272] and outdoor scenarios. However, to the authors' knowledge, no study so far evaluated how different tracking methods affect the realism of sonic experiences. By adopting the tracking implementations described in 4.4, in this section we focus on how different user tracking and head orientation methods affect the perceived position of a virtual object, finding a place within a synthetic indoor environment reconstruction that is reproduced using AAR technologies available in the consumer market.

5.2.1 Material and Methods

The tracking comparison was conducted on 20 participants (19 male, 1 woman, age 20-35), recruited inside the campus of the Department of Engineering and Management of the University of Padova, with no specific experience with binaural sound nor audio-related studies, self-reporting normal-hearing capabilities. No clinical screening of hearing capabilities had been performed. The experiment took place in the corridor described in Ch.4, using the methods implemented in the previously introduced technological platform (refer to Sec. 4.4). Figure 5.1 shows a schematic view of the evaluation scenario.

Experimental Question and Conditions

Our study applied the *Multiple Stimulus Ranking* method used in [273] to evaluate the different tracking approaches. During the experiment, participants could listen to a test sound, consisting of a train of three 40 ms Gaussian noise bursts with 30 ms of silence between each burst spatialized by using SDN reverb whose parameters was calibrated on test room combined with personalized HRTF, heard through AirPods. Table 5.1 summarizes the combinations of position and orientation approaches that define 3×3 tracking solutions to be tested. Tracking methods had been selected to be either out-of-the-box solutions or algorithms easily

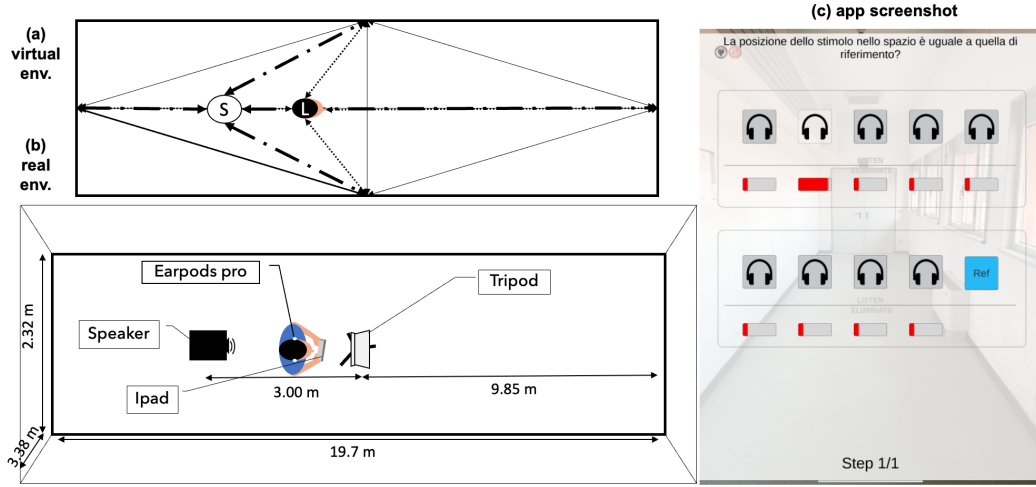


FIGURE 5.1: (a) Scenario of the virtual environment with virtual sound source directions computed by using the SDN algorithm. (b) Scenario of the real environment. (c) Screenshot of experimental app GUI.

Name	Position	Orientation
FE	face tracking	earpods
HE	ipad tracking + height displ.	earpods
TE	iPad tracking	earpods
FF	face tracking	face tracking
HF	ipad tracking + height displ.	face tracking
TF	iPad Position	face tracking
FT	face tracking	iPad tracking
HT	ipad tracking + height displ.	iPad tracking
TT	iPad tracking	iPad tracking

TABLE 5.1: Experimental conditions.

replicable in similar mobile devices. A reference sound with the same nominal position as the virtual sounds could be played back on demand through a speaker in the room. All 9 experimental conditions and the reference sound were administrated by a custom GUI in an iPad (see Fig. 5.1.c), and could be selected by the participant touching the corresponding button. By moving the slider below the button to the right, the participant could remove the specific condition from the pool of available conditions: in that case, the button turned white, and the participant couldn't interact anymore with it. The same device was also used for tracking. During the evaluation, the participant had to consider the following question:

"Is the spatial position of the virtual stimulus the same as the reference"

and was asked to eliminate all the available conditions starting from the condition with the greater difference (worst case) to the one with the smaller (best case), based on the quality of interaction with the AR environment. All the experimental conditions were evaluated in two different positions: with the loudspeaker,

both real and virtual, at 1.70 m height (*sight level*) and at 0 m (*ground level*). At the end of the experimental task, each participant determined his/her individual rank of preferences in tracking conditions.

The setup depicted in Fig. 5.1.b consisted of a bookstand positioned in the center of the corridor and a speaker at 3 m from the bookstand (1.7 m height) connected to an Edirol UA-4FX audio interface. A custom-built web page and a PHP script played back reference sound through a sound card with a mono audio speaker connected. The sound pressure level was calibrated to be -70 dbA with a Sound Level Meter.

Procedure

The participant entered the room, sat down, and filled out the Informed Consent. Three photos of the subject (front and side of the head, ear closeup) were taken and analyzed by the experimenter to create the custom HRTF.² At the end of the procedure, the HRTF files had been transferred to the iPad app. The experiment have been explained to the participant, which wore AirPods and read the instruction on the iPad screen which was put on a stand. In order to correctly evaluate each condition, has been asked to the participant to:

1. navigate the room leaving the iPad on the stand,
2. take the iPad in hand, moving in every direction,
3. crouching

The instructions also explained that, if the iPad loose user tracking the screen became black and the interface unusable. Once the tracking was restored the interface returned to usable. After a short tutorial about the use of the GUI, the experiment started. The whole experimental procedure, which lasted around 20 minutes depending on the participant, was divided into two pages. All experimental conditions were presented at eye level on the first page, and at ground level on the second page. Even if it could be paused and restarted at any time, all participants carried out the experiment without any interruption. Finally, after completing the ranking task participant was asked to fill NASA-TLX [274] questionnaire to detect difficulties and tendencies in the execution of the experimental session.

5.2.2 Results

Plackett-Luce Analysis

Data have been analyzed by fitting the elimination ranking vector with the Plackett-Luce (PL) model, by using R *PlackettLuce* package. The model is built upon Luce's axiom of choice, which states that the odds of selecting one item over another are independent of the set of items from which the selection is made. Having a set of J items $S = \{i_1, i_2, \dots, i_j\}$, the probability of selecting item j from S is given by:

$$P(i_j|S) = \frac{\alpha_j}{\sum_{i \in S} \alpha_i}$$

²For privacy issues: after the experimental session all subject photos and anthropometric data had been deleted.

where α_i represents the worth i . Considering a ranking of j items as a sequence of choices, from the top to the last ranked, the pool of items from which selections are made decreases as the experiment proceeds, and the probability of observing a specific ranking can be assessed by:

$$P(\pi) = \prod_{j=1}^j \frac{\alpha_{i_j}}{\sum_{i \in S_j} \alpha_i}$$

To estimate the parameters of the item worth, the PlackettLuce package utilizes maximum likelihood estimation, analyzing a collection of ranking vectors.

Figure 5.2.a shows the PL-model estimated parameters for the experiment. The worth parameter is displayed in a log scale referring to the condition TT (iPad for both head position and orientation), whose ranking has been set to 0. Statistical significant differences were found between TE and HF ($z=2.543$, $p<0.05$), FT ($z=2.437$, $p<0.05$), FF ($z=1.743$, $p<0.1$), and TF ($z=1.658$, $p<0.1$), FF and FT ($z=1.958$, $p<0.1$), FT and HT ($z=1.292$, $p<0.1$).

Separating sound source at the sight and ground level (Figure 5.2.b), statistically significant differences were found between, HF and HE ($z=2.466$, $p<0.05$), TE ($z=2.272$, $p<0.05$), TF ($z=2.281$, $p<0.05$) and HT ($z=1.964$, $p<0.05$) at eye level, between HE and TE ($z=2.236$, $p<0.05$), HT ($z=2.213$, $p<0.05$) and FE ($z=1.681$, $p<0.1$), TE and HF ($z=1.902$, $p<0.1$), HF and HT ($z=1.942$, $p<0.1$) at ground level.

NASA-TLX test³ reported the following mean ratings:

- Mental Demand: 66.8 ± 19.3
- Physical Demand: 27.5 ± 19.8
- Temporal Demand: 37.5 ± 25.5
- Performance 41.0 ± 20.7
- Effort: 57.5 ± 20.4
- Frustration: 42.0 ± 24.7

Denoting high scores on Mental Demand (66.8 ± 19.3) and Effort (57.5 ± 20.4) according to the values reported in [275].

5.2.3 General Discussion

Overall, conditions H and F about position tracking generally received the highest scores concerning the perception of the correct source position. On the other hand, orientation tracking by the AirPods (E) was preferred the least. In particular, HF condition was preferred. At the same time, head rotation estimated by F was considered generally adequate. iPad rotation tracking (T), which is less reliable in detecting head rotation, does not result in very poor performance. In general, the tested conditions reported a limited number of statistically significant tracking approaches, showing that it was difficult for the participants to clearly identify a ranking, even though HF shows a tendency. The NASA-TLX results confirmed that participants found the task highly mentally demanding. A post-questionnaire confirmed their difficulties in distinguishing different conditions with few exceptions.

³<https://humansystems.arc.nasa.gov/groups/TLX/>

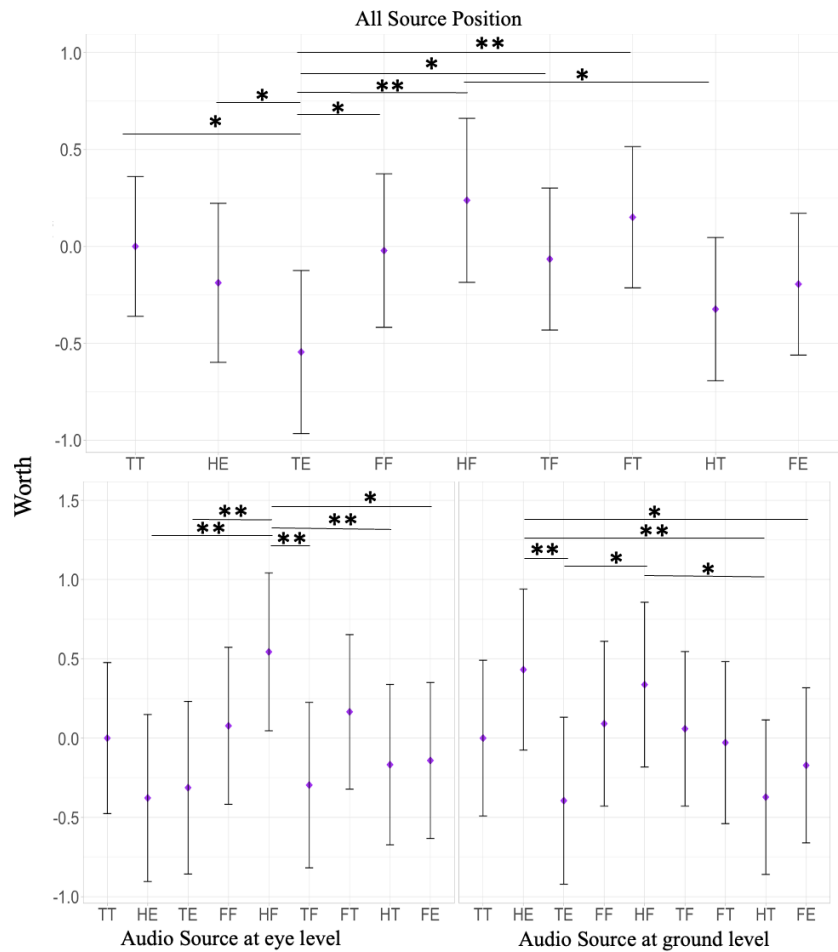


FIGURE 5.2: Rank analysis of available condition on overall evaluation (top), with the sound source at 1.70 m(bottom left) and floor level (bottom right). Asterisks and bars indicate a significant difference (*: $p < .05$, **: $p < .01$, ***: $p < .001$ at post-hoc test)

At eye level, HF resulted in being the preferred condition. Concerning position tracking, F and T received low scores. One can interpret such an outcome as a user preference for more stable tracking despite tracking accuracy. Concerning orientation, earpods (conditions HE, TE, FE) are generally the least reliable for the task. The tablet is generally considered reliable in describing orientation, except for HT.

At ground level, few participants distinguished positional and orientation tracking. Concerning orientation, AirPods (E) were considered the best in position representation using HE condition but the worst in all other cases. HF condition also performs very well with audio sources at feet level. It is important to note that a problematic perception of elevation is a well-known problem of auralization through headphones in VR/AR [276, 277, 278].

Surprisingly, conditions with the H position-tracking method generally received the highest rank. Although this kind of estimation contains a certain amount of localization error in the horizontal plane caused by arm movements and position, it held few variabilities of the position was small when participants were standing still. Section 4.4.1 shows that tracking error for both T/H

and F conditions, while maintaining low variability for small movements, is acceptable for the distances considered (starting position 3 m far from the source), yet it becomes inadequate when a user approaches the audio source. Regarding orientation, F and T have comparable acceptable performances, and participants generally judged E to have poor performances. It is supposed that the current AirPods model introduces jitter and high latency due to Bluetooth communication [279] over time without adequate compensation on the tracking. This peculiar result suggests that participants, once left free to move in every direction inside the room, show an adaptation according to the system resulting in less movement. Moreover, most common HRTF models and measurements for dynamic binaural synthesis, even though personalization permits enhancing sound localization [214], rely on HRTF measurements that do not contain auditory cues for separate and relative head and torso movements [280]. This can lead to a mismatch between the perceived position of the virtual source compared to the real reference for which the user body itself naturally defines the HRTF. Furthermore, ergonomically, having the iPad in their hands forced a coupled movement of the head and torso, making the HRTF synthesis method perceptually coherent with an unnatural behavior. For this reason, participants prefer conditions that avoid sensory-motor mismatch to maintain coherence between actual movements and virtual sound source position. System constraints, such as tablet handling, errors in tracking, and the HRTF synthesis method interact with user perception promoting a tendency in accommodation to select the best possible movement and behavior for each specific tracking condition.

Following the SIVE framework, using a system with more degrees of freedom, including tracking accuracy and audio simulation, may improve the perception of source position. Accordingly, a more robust tracking method should be used in conjunction with a more immersive HRTF synthesis. In this case, the entanglement between the system and the user promotes a mutual adaptation between both actors.

To conclude, it is vital to recall a relevant challenge of available consumer electronic systems like the one used in this study: as stated in Sec- 5.1, AirPods Pro filters the acoustics of physical spaces, partially resulting in a distorted perception of the real audio environment and stimulus. The hear-through effect might cause a potential bias in users' preferences while considering different devices. McGill et al.[281] already reported the influence of acoustic transparency on presence, realism, and externalization.

5.2.4 Conclusion

In this pilot study, we implemented an interface and applied it to an AAR evaluation experiment. The adopted scenario was developed in a non-laboratory environment using mobile AAR, which can be adapted to various contexts. Personalization and calibration of the VE help render a more plausible AAR, even though the number of subjects should be increased in future tests to confirm such a tendency. On the other hand, the specific contribution of each part of the system should be carefully evaluated in order to simplify the design of the VE. Moreover, using a noisy environment led participants to prefer higher-volume conditions. According to the participants, the GUI was easy to use. However, the pilot test revealed some issues that need to be fixed for subsequent testing: (i) how the participants interact with the GUI should be adjusted to force them to use the real reference more frequently; (ii) additionally, user tracking should be carefully

improved to obtain a more precise user position and orientation in the VE, while maintaining a trade-off between portability and accuracy. Even if these important factors significantly reduced the interpretability of the final outcome of the ranking test, this preliminary assessment allowed us to identify valuable insights for an ecological evaluation of mobile AAR scenarios.

5.3 Evaluation on the effect of different rendering factors

Section adapted from publication [14].

One of the most important aspects in realistic AAR is *auralisation*, the audio rendering process involving models that combines listener acoustics, i.e., HRTF, and environmental acoustics, namely RIR. Effective auralisation of sound sources is fundamental to providing a plausible listening experience, but the flexibility of the rendering process is crucial for a real-time environment [25]. In this section, we developed the drag-and-drop MUSHRA [282] in a mobile application to evaluate the effect on an AAR scenario of acoustic personalization and calibration in terms of perceived *plausibility*. Using the technological platform described in Chapter 4, the objective of this test is twofold: to preliminarily assess the impact of different rendering factors in an AAR setup and to verify the efficacy and usability of the procedure.

5.3.1 Audio Quality Evaluation

This test evaluates different degrees of personalization and calibration of the acoustical space. It involved 16 participants (all male, age 20-35), self-reporting normal-hearing capabilities.

5.3.2 D&D MUSHRA

To collect user preference, we implemented the evaluation method suggested by [282, 283] called *drag-and-drop MUSHRA* (Multiple Stimuli with Hidden Reference and Anchor, D&D MUSHRA), a modified MUSHRA test⁴ aimed to simplify the rating interface, especially using touch screens. MUSHRA is a standard methodology for audio stimulus comparison tests in terms of perceived quality, usually employed in evaluating lossy audio compression algorithms. The suggested modification introduces a unified interface (Fig. 5.3) for playback and rating using drag-and-drop actions, resulting in a reduced time to complete the ranking task [283]. The stimulus playback is triggered by the user pressing the corresponding button with a generic icon and no stimulus numbering. The rating is achieved by moving the button on a xy-grid, where the horizontal axis modifies the rating scale. The vertical axis can help users visually organize their ratings without affecting stimulus scoring. In the upper right corner of the GUI, a light blue button is available to play reference stimulus on demand.

The D&D MUSHRA could be easily adapted for testing audio quality taxonomies such as Spatial Audio Quality Inventory (SAQI) [284]. SAQI has been developed as a comprehensive taxonomy of attributes that are crucial for understanding how audio contributes to the overall user experience in VR/AR environments. This inventory categorises and defines various perceptual qualities related

⁴ITU-R recommendation BS.1534-3 available at <https://www.itu.int/rec/R-REC-BS.1534-3-201510-I>

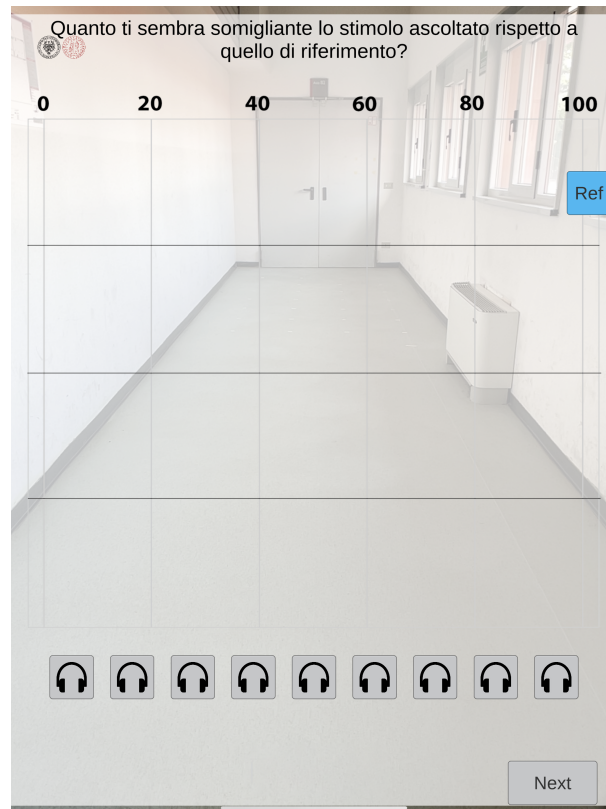


FIGURE 5.3: The D&D MUSHRA User Interface.

to spatial audio, such as localisation accuracy and externalisation, among others. Recent work such as [285] used SAQI of perceptual evaluation frameworks in VR/AR: by comparing binaural auralisations with a real loudspeaker under different room-acoustic simulations, HRTF datasets, and head-mounted display conditions, their study demonstrates how multiple rendering factors jointly influence users' room-acoustic perception in an audiovisual VR scenario. By offering a structured approach to evaluating these attributes, SAQI allows researchers and developers to identify the strengths and weaknesses of audio systems in a detailed and nuanced manner. This is particularly important in VR/AR, where the interaction between auditory and visual stimuli plays a significant role in creating a convincing and immersive experience. Ratings obtained from the D&D MUSHRA interface were comparable to those obtained with the classic interface in terms of reliability and discrimination ability. To the authors' knowledge, this study implements the D&D MUSHRA interface in an experimental setup for the first time.

Setup

The setup consisted of a loudspeaker Genelec 8030C (Woofer 5", Tweeter 3/4", frequency response: 47 Hz -25 kHz, positioned at 1.7 m from the floor) connected to an Edirol UA-4FX audio interface. In Unity, a virtual corridor was created with the same dimension as the real experimental room, spatially aligned with the real room during the experiment using the AR framework. The virtual user avatar was positioned at the same position as the participant, at 9.85 m from the front of the room, with the windows on the right. The virtual audio source was positioned behind the participant in the same position as the real loudspeaker,

leaving the VE hidden on the iPad screen. As depicted in Fig. 4.1, an Apple iPad Pro (12.9-inch 5th gen., iPadOS 15.4) and Apple AirPods Pro (1st gen.) tracked the participant's body. iPad front and rear cameras tracked head and body position in the physical space, while the AirPods Pro IMU sensor tracked head orientation only. To create a more ecological audio environment, though mediated by the use of earphones, AirPods *Audio Transparency Mode* was used.

Stimuli and Conditions

During the pilot experiment, participants could listen to a test sound consisting of a train of three 40 ms Gaussian noise bursts with 30 ms of silence between each burst. A reference stimulus with the same nominal position as the virtual sounds could be played back (volume matched at 70 db(A)) on demand through the loudspeaker in the room. Test audio was spatialized using nine different arrangements of the following components:

- **HRTF:** Include customized HRTF, a generic KEMAR head, or no HRTF.
- **Artificial reverberation:** A second room (20 m x 20 m x 4 m) was modeled in some experimental conditions in addition to the previously described experimental corridor. We applied one of these two reverb conditions or no reverb to the conditions.
- **Headphone equalization compensation:** audio impulses can be calibrated to reduce the coloration caused by headphones⁵. The equalization curve, the inverse of the earpods transfer function, applied is shown in Fig. 4.5.
- **Stimulus volume:** To achieve a comparable Sound Pressure Level (SPL) with the reference, the audio stimulus SPL was calibrated before the experiment. A custom setup was developed using a virtual source positioned within the VE, played through a fake silicone ear wearing AirPods, to simulate the placement of headphones over a human ear. Using a sound level meter, the gain in the simulation was adjusted to calibrate the volume according to the reference loudspeaker. To test the effect of volume on perceived stimuli, two different SPL were used: 70 db(A) corresponding to the real reference stimulus SPL and 75 db(A), i.e., a +5 db(A) gain. Volumes were measured at 1.21 m(4 ft) from the source.

The final experimental configurations are described in Table 5.2. Considered conditions include a fully customized setup (A), and a hidden reference with no spatialization (HR). To keep the experiment time short and not tiring participants, the other condition was chosen to have a representative set of all available arrangements, keeping the test set small:

- B: to evaluate the isolated effect of reverb, no HRTF spatialization has been applied along with a single channel (mono) corridor reverb and calibrated volume; no headphone calibration was employed;
- C: to evaluate the isolated effect of personalized HRTF, no reverb spatialization has been applied in a fully calibrated environment;
- D: fully customized environment with unmatched volume, to test the effect of volume on plausibility;

⁵Impulse file available at <https://github.com/jaakkopasanen/AutoEq>

	HRTF	Reverb	Vol. - db(A)	Hp Eq
HR	<i>none</i>	<i>none</i>	+70	<i>none</i>
A	personal	calibrated	+70	comp.
B	<i>none</i>	calibrated	+70	<i>none</i>
C	personal	<i>none</i>	+70	comp.
D	personal	calibrated	+75	comp.
E	kemar	large	+75	<i>none</i>
F	personal	large	+70	comp.
G	kemar	calibrated	+70	comp.
H	personal	calibrated	+75	<i>none</i>

TABLE 5.2: Configurations available in the pilot test. HR is a hidden reference condition

- E: complete auralisation with no customization and calibration with unmatched volume to compare a fully customized environment with a generic auralisation;
- F: to evaluate the effect of reverb calibration, complete customization with no reverb calibration has been applied;
- G: to evaluate the effect of HRTF personalization, complete calibration has been applied with no HRTF personalization;
- H: HRTF personalization and reverb calibration; unmatched volume and no headphone calibration, to evaluate the effect of personalization without calibration

Procedure

The participant entered the room, sat down, and completed the Informed Consent. The experimenter took and analyzed three photos of the subject (front and side of the head, ear closeup) to create the custom HRTF.⁶ The HRTF files were transferred to the iPad app at the end of the procedure. The pilot experiment was explained to the participant, who wore AirPods and read the instruction on the iPad screen, which was put on a stand. To avoid front-back confusion issues [286, 287], the participant has been asked to navigate the room without turning toward the virtual speaker. The instructions also explained that if the iPad loose user tracking, the screen became black and the interface unusable. Once the tracking was restored, the interface returned to usable.

Using an iPad, all experimental conditions and the reference sound were administered by the developed D&D MUSHRA GUI previously described, which can be selected by touching the corresponding button. During the evaluation, the participant had to consider the following question:

"How similar is the virtual stimulus to the reference?".

After a short tutorial about the usage of the GUI, the experiment started. To avoid accustoming the subject to the stimulus. in the tutorial phase non-spatialized

⁶For privacy issues: after the experimental session, all subject photos and anthropometric data were deleted.

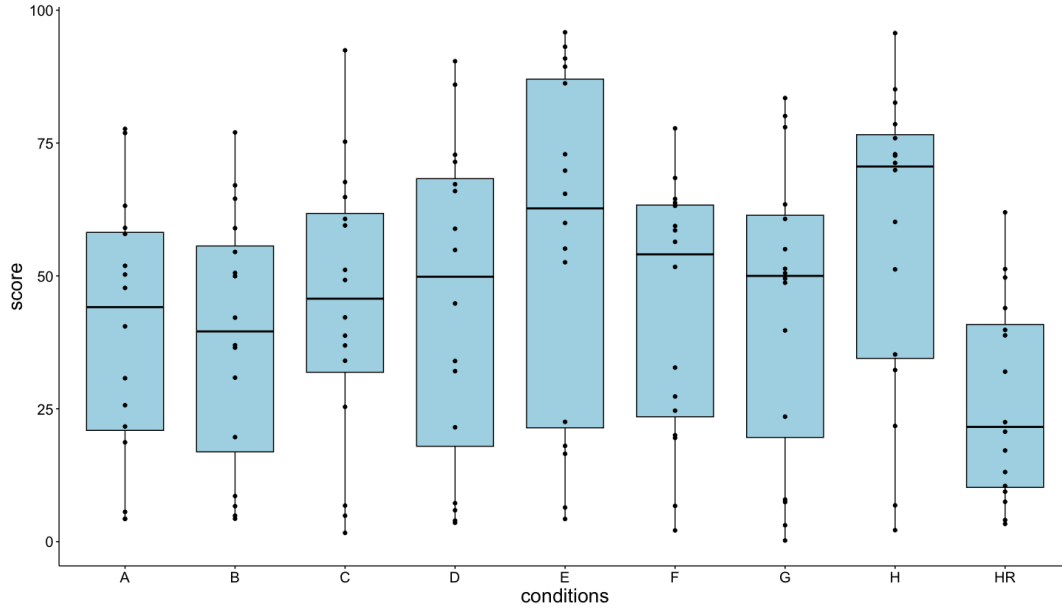


FIGURE 5.4: Boxplot of user ratings for each condition.

speech sound, played back through the real speaker, had been used. Even if it could be paused and restarted at any time, all participants carried out the experiment without interruption. Overall, the experiment lasted about 15 minutes, including 5 minutes of configuration. At the end of the experimental task, each participant determined his/her individual score in tracking conditions.

5.3.3 Results and Discussion

The pilot test aims mainly to evaluate the test procedure and the D&D MUSHRA interface. Collected rankings are shown in Fig. 5.4. Repeated Measures ANOVA shows a significant effect of conditions, obtaining $F(4.06, 60.96) = 2.603$ ($p < 0.05$) and differences HR-C ($p < 0.01$), HR-E ($p < 0.01$), and HR-F ($p < 0.05$). Since this pilot test based upon a small subject sample, no post-hoc analysis has been performed on data. These preliminary results exhibit a tendency for the H condition: HRTF and reverb personalization help render a more plausible AAR. Extensive tests are needed to confirm this trend. Subjects typically rated the condition with a higher Volume (+75 db(A), including D, E, and H) with higher scores. This result can be attributed to the environmental noise of the corridor, which was frequently reported by subjects following the experimental session, and which acoustics were not controlled in order to obtain an ecological AAR. Presumably, increased volume improved hearing clarity, influencing participants' perceptions. Moreover, SDN parameters used in the virtual room simulation could influence the perceived plausibility and can be improved by using other environmental measures such as RT60 decay. Another important bias can be given by the use of *Audio Transparency Mode* of AirPods, which can alter the environmental room audio perception.

Users found it easy to operate on the D&D MUSHRA GUI and to navigate between conditions and the reference, and no subject reported problems. On the other hand, according to the experimenter, participants' tendency to compare experimental conditions to the real reference significantly impacts the evaluation results, making the requested comparison difficult. Often users did not compare

stimuli with the reference, preferring a direct comparison between the conditions. For future tests, the D&D MUSHRA interface should be modified to implicitly require the participants to use the real reference more frequently.

User tracking was another important factor that diminished the significance of the test. In [52], the authors illustrate that subjects tend to adopt a movement that is consistent with the tracking method, resulting in the adaptation of the user to the system changing their behavior within the VE. Consequently, tracking methods could significantly affect the results. It is important to note that using consumer electronics devices has several advantages in design and implementation cost, experience reproducibility, and portability but often contains proprietary technologies that cannot be fully controlled and require a certain trade-off between accuracy and responsiveness. Considering AirPods, the manufacturer provides no specifications about IMU sensors available. Informal post-experimental user feedback reports significant tracking errors, probably due to no compensation or smoothing algorithms of orientation data that could lead to drift error over time and high tracking latency⁷. During this pilot test, no information about the lost tracking rate was recorded. About this topic, further accuracy measurement studies are advisable and our object of future studies.

5.3.4 Conclusion

In this study, we implemented the D&D MUSHRA interface and applied it to an AAR evaluation experiment. The adopted scenario was developed in a non-laboratory environment using mobile AAR, which can be adapted to various contexts. Personalization and calibration of the VE help render a more plausible AAR, even though the number of subjects should be increased in future tests to confirm such a tendency. On the other hand, the specific contribution of each part of the system should be carefully evaluated in order to simplify the design of the VE. Moreover, using a noisy environment led participants to prefer higher-volume conditions. According to the participants, the GUI was easy to use. However, the pilot test revealed some issues that need to be fixed for subsequent testing: (i) how the participants interact with the GUI should be adjusted to force them to use the real reference more frequently; (ii) additionally, user tracking should be carefully improved to obtain a more precise user position and orientation in the VE, while maintaining a trade-off between portability and accuracy. Even if these important factors significantly reduced the interpretability of the final outcome of the ranking test, this preliminary assessment allowed us to identify valuable insights for an ecological evaluation of mobile AAR scenarios.

The results obtained in Experiments 5.3 and 5.2 highlight that both tracking accuracy and rendering factors contribute to perceptual plausibility, although not always in a strictly linear manner.

Building upon these findings, the following experiment investigates whether such technical variations remain perceptually relevant in more complex and ecologically valid interaction scenarios, where multiple users and environmental cues are involved. In particular, it explores the extent to which simplified rendering approaches may still support co-immersive experiences.

⁷We informally conducted a one-way pilot test showing a latency of 200-300ms

5.4 A new method for SAQI evaluation: D&DMUSHRA+

Section adapted from publication [17].

In experimental research involving MR technologies, particularly when tasks involve comparing different stimuli such as audio or visual elements, the inclusion of multiple factors in the design can significantly complicate the experimental procedure. This complexity often results in an increased number of experimental conditions and a more delicate choice between a between-subjects and a within-subjects design, as each approach involves specific trade-offs: between-subjects designs require larger participant samples to ensure statistical power, while within-subjects designs demand longer sessions and careful counterbalancing to mitigate learning and fatigue effects—making the process more resource-intensive and time-consuming. Furthermore, managing study populations becomes challenging due to individual variability, which complicates the application of consistent procedures across diverse participants. This variability, combined with lengthy and rigid testing protocols, can make it difficult to achieve generalised results, especially in preliminary testing, when usually few participants are involved. Additionally, the tedious nature of prolonged procedures may affect participants' motivation and performance, leading to evaluation errors in screening processes [288].

To overcome these challenges, developing new research methodologies that can accommodate individual differences while ensuring reliable outcomes would be highly beneficial. In this section, we aimed to assess the impact of different rendering factors using a novel approach that would improve standard Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests⁸. In particular, we aimed to verify the effectiveness of a procedure incorporating a mobile app featuring the drag-and-drop MUSHRA test [282] with the introduction of D&DMUSHRA+, a novel integrated methodological framework combining:

1. The navigation of an arbitrarily large number of stimuli and experimental conditions.
2. A machine learning-based analysis based on hierarchical clustering for the identification of descriptive patterns among experimental conditions.

The interaction design of the user interface for the evaluation procedure and protocol must ensure a low experimental time, even when using a large number of stimuli/conditions arising from multiple, interdependent factors. Considering many conditions helps discover similarities and disambiguate preferences among different combinations of factors by strengthening the descriptive power of the evaluation outcomes.

As a case study, we expand the evaluation described in Sec.5.3 of the technological platform described in Chapter 4. The main goal is to enhance the evaluation of various aspects of the acoustic personalization and calibration on an audio AR scenario, specifically their impact on the *plausibility*.

To achieve this, it is important to consider various aspects ranging from technological platforms to other, more perceptual and subjective elements. In particular, experienced listeners demonstrate significantly higher spatial audio localisation accuracy compared to those who are less familiar or reliable in active listening ([289]). More broadly, distinguishing between listener categories remains

⁸ITU-R recommendation BS.1534-3 available at
<https://www.itu.int/rec/R-REC-BS.1534-3-201510-I>

a complex task, influenced by factors such as multisensory calibration and integration, familiarity with immersive or spatial audio technologies, and musical background ([290]). Building upon these elements, we consider perceived plausibility by comparing multiple rendering factors—such as room acoustics, user acoustic personalisation, and device types—and different types of users, i.e., audio experts and non-experts.

A critical endeavour in assessing AR technology quality lies in identifying a minimal set of factors that have the most significant impact on the user's experience. This can be difficult due to the complexity of the factors involved and individual user variability. To address this, D&DMUSHRA+ allows researchers to efficiently handle a large number of stimuli and conditions, facilitating the screening phase, helping to reduce the initial conditions by indicating one or more minimal sets to focus on. Moreover, D&DMUSHRA+ supports the piloting phase, not only simplifying the experimental procedure but also enhancing the ability to discern similarities and preferences among different combinations of factors, thereby strengthening the descriptive power of the evaluation outcomes.

5.4.1 Piloting and Screening in VR/AR

Dealing with experiments with a large number of conditions, such as rendering factors in ecological VR/AR scenarios, makes it challenging to achieve sufficient statistical power. In traditional evaluation, testing many conditions heightens the need for large sample sizes to detect significant effects, especially when inter-subject variability is high or expected effect sizes are small ([291]). Moreover, the interaction between factors often introduces nonlinear perceptual effects and subtle perceptual shifts that are difficult to detect without extensive participant data. In this context, *screening* refers to the early stage of experimentation involving a large number of conditions or factors to be explored and reduced, whereas *piloting* denotes a small-scale preliminary study conducted prior to a main or definitive experiment. These early stages present substantial challenges, particularly in immersive audio research, where few participants are typically involved and the subjective variability in listening is high ([290]).

Table 5.3 provides a structured summary of six pilot and screening studies involving VR/AR environments, covering a spectrum of different methodologies spanning two critical experimental design dimensions: the number of experimental conditions (presented in ascending order) and participant involvement. Accordingly, we also reported the tasks and collected measures employed by the authors.

The study conducted by ([292]) examined how age differences influenced key factors such as presence, workload, and cybersickness when participants engaged in immersive virtual environments. The study opted for an exploratory format using a diverse range of assessment tools, including the Igroup Presence Questionnaire (IPQ), NASA Task Load Index (NASA-TLX), System Usability Scale (SUS), and the Simulator Sickness Questionnaire (SSQ). Despite testing only four experimental conditions, the study collected a substantial volume of data over an extended period of two hours per participant, making the design time-intensive. By contrast, ([293]) presented a more streamlined experimental design in a pilot usability study that engaged 14 participants. The study involved six different conditions in which participants were asked to answer binary “yes-no” questions within a VR interface with two buttons that varied in position, shape, and colour. It aimed to investigate how these six conditions influence user performance and

Study	Type	Domain	Task	N Cond.	N Part.	Measures
[292]	Pilot	VR, presence and aging	Brief VR experiences in HMD; counterb. order	4	40	IPQ, NASA-TLX, SUS, SSQ, CPQ
[293]	Pilot - Screening	VR, user interface usability	Yes-no interaction modes in VR; auto vs manual evaluation	6	14	Ad-hoc usability questionnaire
[285]	Screening	Room acoustics in MR	Within-subject: conditions varying in rendering parameters	8	21	Seven attributes partially from RAQI
[294]	Screening	Time perception in VR	Within-subject: conditions varying in speed and number of elements on screen	12	137	Endtime, vection, boredom, and tiredness with ad-hoc questionnaire

TABLE 5.3: Summary of the selected pilot and screening studies on MR environments.

evaluate the accuracy of automated measurements compared to usability questionnaires.

[285] conducted an exploration study on binaural auralisation factors in audio VR compared to a real loudspeaker. Their study included eight different conditions, varying in room-acoustic simulation settings, HRTF datasets, and the presence or absence of a head-mounted display. The study relied on a smaller sample of expert participants ($n = 21$), with an overall limited number of conditions, but measured five different attributes selected from the Room Acoustical Quality Inventory (RAQI) [295]: reproduction quality, reverberance, tone colour, loudness, and source distance. Finally, [294] explored time perception under simulated motion in an immersive VR tunnel environment, combining multiple levels of motion speed, visual density, and exposure duration into a fully crossed 12-condition design. With 137 participants, the study successfully examined non-linear perceptual effects in VR, but again illustrates how multi-condition studies typically depend on relatively large samples to maintain statistical power.

Together, these studies reflect the breadth and diversity of pilot and screening research in immersive technology domains. Notably, we did not identify any study that featured both a high number of experimental conditions and a small sample size. This combination is probably rarely used since it tends to produce low statistical power and unreliable effect sizes even for pilot or preliminary testing ([291]).

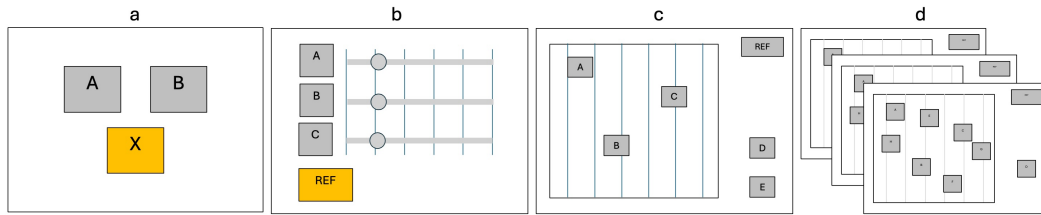


FIGURE 5.5: Example of different audio comparison test interfaces: a) ABX test, b) traditional MUSHRA test, c) D&D MUSHRA, and d) the proposed D&DMUSHRA+ with multiple pages.

While many studies adopt simplified experimental designs to cope with practical constraints, such as limited participant availability or the complexity of immersive systems, this often comes at the cost of ecological validity and a deeper understanding of user experience across varied contexts. For instance, studies such as those by [296], [297], and [298] highlight how controlled laboratory setups or reduced-factor experiments, although methodologically convenient, may fail to capture the perceptual dynamics and contextual variability typical of real-world VR/AR scenarios.

Our work aims to propose a methodology that balances design complexity with feasibility, exploring the possibility of including a richer set of experimental conditions in small-scale studies, ensuring interpretability and statistical reliability. Accordingly, the inclusion of a large number of experimental conditions with a limited number of participants and time is a challenging objective to pursue.

5.4.2 D&DMUSHRA+: D&D MUSHRA on multiple pages

For a comparison of audio technologies, many different solutions have been developed throughout the years. A first option is the traditional ABX test (Fig 5.5-1), in which two signals, A and B, and a third signal, X, which is randomly selected to be identical to either A or B, are provided. Participants are required to classify whether X is A or B, measuring their accuracy. MUSHRA⁹ (Fig 5.5-2) is a standard, widely used methodology for comparing audio stimuli based on perceived quality, introduced to simplify the use of different audio stimuli. It is commonly used to evaluate lossy audio compression algorithms. The standard test consists in the comparison of a certain number of audio stimuli with respect to a reference, a hidden version of it, and one or more anchors.

To enhance the usability of MUSHRA tests, especially to simplify the user interface for children and elderly people, several modifications of the standard MUSHRA have been proposed. Of particular interest, *Drag-and-drop* MUSHRA ([299, 282] (D&D MUSHRA, Fig 5.5-3, and Sec.5.5.3.2) is a modified MUSHRA test designed to simplify the rating interface, especially for touch screens. D&D MUSHRA modification introduces a unified interface for both playback and rating of stimuli. This modification has resulted in a reduction in the time required to complete the ranking task as compared to the traditional method ([283]). Stimulus playback is controlled by the user touching a button with a generic icon. Rating is achieved by moving the button on an XY-grid, where the horizontal axis adjusts the rating scale. The vertical axis can be used to visually organise

⁹ITU-R BS.1534-3: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-1!!PDF-E.pdf

ratings without affecting stimulus scoring. Furthermore, a button is available in the upper right corner of the GUI to play the reference stimulus on demand. The D&D MUSHRA interface can be easily modified to test audio quality taxonomies like SAQI.

In our evaluation process, we introduced the D&DMUSHRA+ (Fig 5.5-4), dividing the experimental conditions into multiple distinct *pages* for enhanced ease of navigation. Each page presented a maximum of nine randomly chosen conditions for evaluation, and upon completion, by pressing the *next* button, nine new conditions were presented, while the previous ones remained accessible on the grid. Throughout the experimental procedure, the participant can listen and compare stimuli to the reference and adjust the ratings of all available conditions as desired. The use of multiple pages encourages participants to revisit and reconsider previously evaluated stimuli, while this structured organisation of conditions enhances the overall ease of navigation through the various elements to evaluate. Regarding the presentation order of stimuli, various options can be considered:

1. Randomisation of all conditions across pages.
2. Following the traditional MUSHRA approach, including a hidden reference and one or more anchors, together with the representative test conditions.

The latter approach of stimulus presentation employs a representative set of stimuli, which appear in randomised order within the nine stimuli on the first page. This method aims to provide hidden references from the early stages of rating, relative to the spectrum of conditions involved in the evaluation. The remaining conditions are instead distributed randomly on the subsequent pages.

Handling Data

Beyond the user interface and the interaction methodology, the D&DMUSHRA+ suggests a hierarchical approach to data assessment, enhanced by machine learning analysis. This methodological approach is grounded in two key principles also applied in similar contexts ([300]): hierarchical ordering, which prioritizes the estimation of lower-order effects (main effects) before addressing higher-order interactions (e.g., two-way interactions), and the sparsity of effects assumption, which posits that only a limited number of factors in a factorial design will have meaningful or significant influence on the outcomes. These strategies enhance the interpretability of results and reduce the burden of exhaustive testing, which is particularly beneficial in pilot or exploratory studies with many experimental conditions. Unlike traditional MR audio evaluation methods, which typically test few rendering factors in isolation, our approach is designed as a screening methodology that allows multiple rendering factors to be explored simultaneously. The goal is not to provide a definitive perceptual characterisation, but to identify which conditions are perceptually distinct or similar, and therefore worth selecting for subsequent, more controlled experiments.

While the ITU-R BS.1534-3 recommendation suggests repeated-measures ANOVA or linear models for the analysis of MUSHRA results, our approach extends these methods, drawing inspiration from signal detection theory to better capture perceptual discriminability patterns rather than to replace standard inferential analyses. In MUSHRA tests, [301] raised important concerns about the statistical validity of such approaches. Their study demonstrated that MUSHRA data often violate key ANOVA assumptions—interval scaling, normality, homogeneity

of variances, and independence of observations—which can increase the risk of Type I errors (false positives).

To uncover general traits in the data collected with D&DMUSHRA+, we employed a combination of statistical modeling and exploratory analysis techniques: Mixed Linear Model Regression (MLMR,[302]) for fine-grained comparisons of statistical distribution beyond group-level differences of ANOVA, and machine learning analysis using Hierarchical Clustering Analysis (HCA) with the Ward linkage method to identify distances between factors ([303]). MLMR estimates fixed effects through maximum likelihood (REML) and expresses the significance of coefficients as Wald z-statistics ([304]). HCA is an unsupervised learning technique that groups similar data points into clusters based on their features, represented as vectors in a multi-dimensional space. In the Ward linkage method, given two clusters A and B, they are merged based on minimising the total within-cluster variance, represented mathematically as:

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2, \quad (5.1)$$

where $\Delta(A, B)$ denotes the increase in within-cluster variance resulting from merging clusters A and B; $\vec{m}_A$ and $\vec{m}_B$ are the centroid vectors of clusters A and B, respectively; and n_A and n_B are the numbers of vectors containing the scores assigned to a given experimental condition from all participants. The method iteratively merges the pair of clusters that result in the smallest increase in Δ , effectively minimising the total variance across all clusters. Ward linkage method tends to produce compact, homogeneous groups—an approach particularly suited to perceptual similarity data ([305]). Due to the relatively small sample size in our study, the data are insufficient to reliably estimate a dense distribution. Instead, clustering offers an alternative lens, enabling us to extract patterns of similarity between conditions.

The outcome is a hierarchical tree structure known as a dendrogram, which can be graphically represented to visualise the clustering process that helps identify natural groupings within the data. This approach allows us

1. to capture similarities in rating patterns, not just mean aggregated values.
2. to reveal latent perceptual groupings, including clusters dominated by specific conditions and factors, helping in revealing the most influential rendering features in the analysis.
3. to allow multi-level inspection, i.e., different cut heights in the dendrogram, enabling the identification of structural trends across conditions.

5.4.3 Materials and Methods

The exploratory test involves 16 participants (age 22 ± 5 , all male) with self-reported normal hearing capabilities and evaluates different calibrations of the acoustical space and degrees of personalisation. Participants were divided into two groups. The first included musicians, experts in audio and music technology, mainly recruited from the campus of music technology of the Music School of Vicenza (8 participants). The latter included non-musicians, recruited inside the campus of the University of Padua. This study adheres to the principles outlined in the Declaration of Helsinki ([306]). Technical setup and stimuli (Gaussian noise bursts) calibration are the same described in Sec. 5.3.2.

Stimuli and Conditions

The audio conditions were generated by using all possible arrangements, $3 \times 3 \times 2 \times 2 = 36$ in total, of the following factors:

- **Artificial reverberation** - 3 levels: Along with the *calibrated* corridor, a second *large* room (20m x 20m x 4m) was modelled with the same artificial reverberator described in Sec. 4.1. The no reverb was also considered.
- **HRTF** - 3 levels: Personalised HRTF (See Ch. 4.2 for details), a generic HRTF of a KEMAR dummy head ([307]), or no HRTF.
- **Headphone compensation** - 2 levels: audio impulses were not equalised or equalised to reduce the colouration caused by headphones. The compensation curve¹⁰, i.e., the inverse of the earpods transfer function, applied is shown in Fig. 4.5 (Ch. 4).
- **Reproduction volume** - 2 levels: The experiment involved two different SPLs: 70 dB(A), which corresponded to the real reference stimulus SPL, and 75 dB(A), which had a gain of +5 dB(A).

The experimental configurations are described in Table 5.4, in which condition PC70E is the fully customised setup, and HR is the hidden reference. The D&D MUSHRA used in this experiment was integrated inside the AR auralization framework and implemented in Unity. Overall, the stimuli were divided into four pages. In the first page of the D&D MUSHRA, nine conditions were chosen to have a representative set of the spectrum of all possibilities, presented in a randomized order, including the hidden reference (HR), the fully calibrated environment with matched (PC70E) and unmatched volume (PC75E), the environments with the isolated effect of each component (C70, P70E), no personalisation and calibration (KL75), or with the calibration of a specific component (PL70E, KC70E, PC75). All the other 27 conditions were presented in the following three pages in randomised order.

Procedure

After entering the room, the participant completed the Informed Consent, and the experimenter took front and side photos, ensuring the left ear was clearly visible, of their head to create custom HRTF files that were transferred to an iPad app at the end of the procedure. The processing of the pictures required only a few minutes (typically 2–3 minutes per participant). The participant was briefed about the experiment, wore AirPods and read the instructions on the iPad screen, which was put on a stand. To ensure that the participant's exploration did not rely on constant visual alignment with the virtual source, thereby avoiding front-back confusion ([286, 287]), the participant was instructed to move around the room without facing the virtual speaker. The instructions also explained that if the iPad lost user tracking, the interface would be temporarily turned black. As tracking was restored, the interface would resume full functionality. To help the subject perform the task, the iPad was put on a stand.

Before the experiment began, the participants were given a brief tutorial on using the GUI. During this tutorial phase, non-spatialized speech sound was used to prevent the subject from getting accustomed to the stimulus. The speech sound

¹⁰Impulse file available at <https://github.com/jaakkopasanen/AutoEq>

Name	HRTF	Reverb	db(A)	Hp Eq	Plausibility
HR	<i>none</i>	<i>none</i>	+70	<i>none</i>	26.60±18.87
PC70E	personal	calib.	+70	eq.	39.75±24.66
PC75E	personal	calib.	+75	eq.	45.03±30.15
PC75	personal	calib.	+75	<i>none</i>	57.14±28.89
PL70E	personal	large	+70	eq.	43.55±24.05
PL70	personal	large	+70	<i>none</i>	52.34±25.26
PL75E	personal	large	+75	eq.	40.40±23.99
PL75	personal	large	+75	<i>none</i>	61.06±23.16
P70E	personal	<i>none</i>	+70	eq.	44.46±26.01
P70	personal	<i>none</i>	+70	<i>none</i>	44.38±21.72
P75E	personal	<i>none</i>	+75	eq.	39.62±29.72
P75	personal	<i>none</i>	+75	<i>none</i>	60.53±27.52
KC70E	kemar	calib.	+70	eq.	43.92±27.82
KC70	kemar	calib.	+70	<i>none</i>	47.94±25.51
KC75E	kemar	calib.	+75	eq.	40.00±26.34
KC75	kemar	calib.	+75	<i>none</i>	51.99±27.89
KL70E	kemar	large	+70	eq.	39.09±22.95
KL70	kemar	large	+70	<i>none</i>	52.11±23.06
KL75E	kemar	large	+75	eq.	49.71±28.31
KL75	kemar	large	+75	<i>none</i>	56.19±32.72
K70E	kemar	<i>none</i>	+70	eq.	40.51±25.48
K70	kemar	<i>none</i>	+70	<i>none</i>	55.08±25.99
K75E	kemar	<i>none</i>	+75	eq.	45.08±33.03
K75	kemar	<i>none</i>	+75	<i>none</i>	50.67±33.55
C70E	<i>none</i>	calib.	+70	eq.	23.98±21.39
C70	<i>none</i>	calib.	+70	<i>none</i>	38.32±23.90
C75E	<i>none</i>	calib.	+75	eq.	28.18±23.47
C75	<i>none</i>	calib.	+75	<i>none</i>	35.38±31.70
L70E	<i>none</i>	large	+70	eq.	31.47±24.25
L70	<i>none</i>	large	+70	<i>none</i>	43.89±21.17
L75E	<i>none</i>	large	+75	eq.	37.34±21.61
L75	<i>none</i>	large	+75	<i>none</i>	51.17±31.38
70E	<i>none</i>	<i>none</i>	+70	eq.	31.08±22.37
75E	<i>none</i>	<i>none</i>	+75	eq.	47.52±22.67
75	<i>none</i>	<i>none</i>	+75	<i>none</i>	39.98±28.96

TABLE 5.4: Experimental conditions as combinations of the tested factors and the plausibility ratings provided by the participants (Mean $\pm$ SD). HR is a hidden reference condition, and PC70E is the fully calibrated condition.

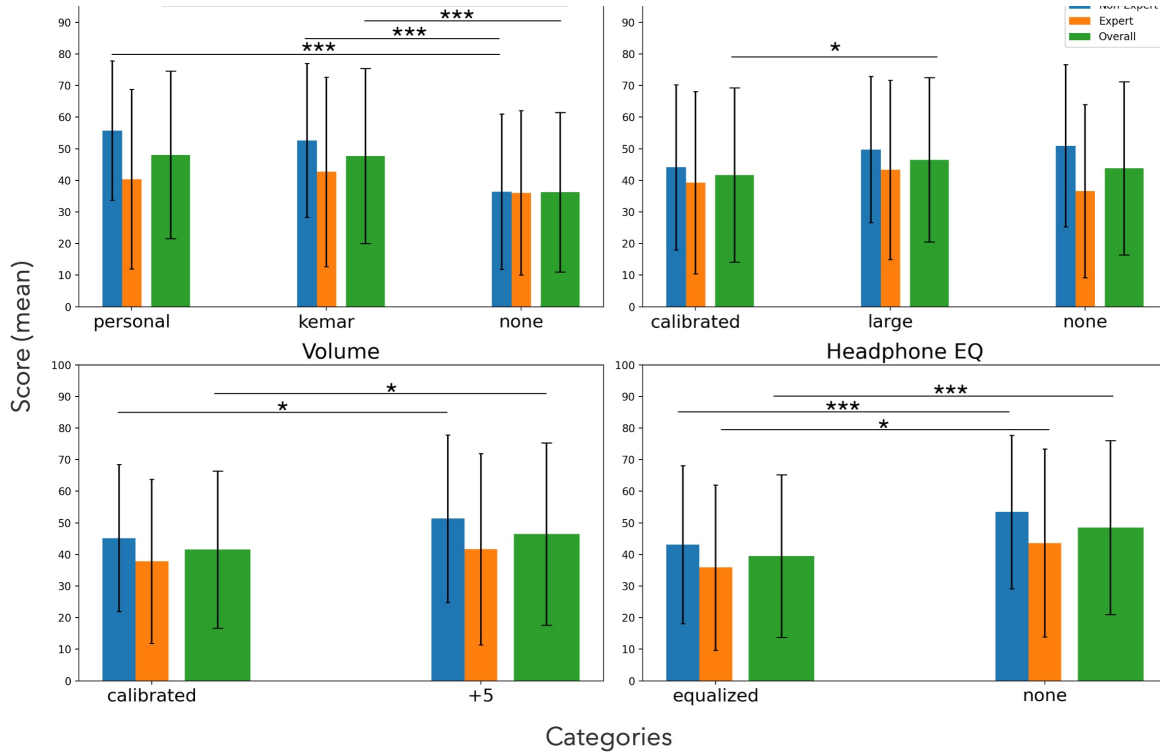


FIGURE 5.6: User ratings for each main factor. asterisks and bars indicate a significant difference (*: $p < .05$, **: $p < .01$, ***: $p < .001$ at posthoc test).

was played back through the real speaker, and the participants were able to pause and restart it at any time.

The experimental conditions, divided into 4 randomly chosen groups and the reference sound were available in the developed D&DMUSHRA+ GUI previously described and could be selected by touching the corresponding blue button. During the evaluation, the participant had to answer the following question by moving the stimulus button following the x-axis of the interface, representing a continuous scale within the range [0 100]:

"How similar is the stimulus to the reference?".

At any time, the participant could listen to any stimulus and the reference sound by touching the corresponding buttons. It is important to note that the participant did not remove AirPods while listening to the real reference, resulting in a mediated listening with distortion of the real source.

The entire experiment lasted on average 25 minutes plus 5 minutes of setup configuration. At the end of the experimental task, participants determined the individual scores for each condition. All participants conducted the experiment without interruption.

5.4.4 Results

Following subsection 5.4.2, we used *scikit-learn* and *statsmodels* libraries in Python for the Mixed Linear Model Regression and Hierarchical Cluster Analysis of the plausibility ranking and *matplotlib* for dendrogram and distribution visualisations.

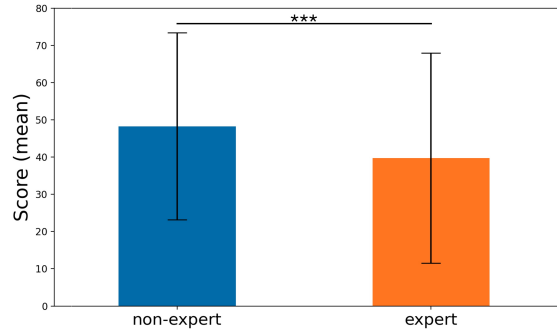


FIGURE 5.7: Total ratings for audio expert vs. non-expert users. asterisks and bars indicate a significant difference (***: $p < .001$ at posthoc test).

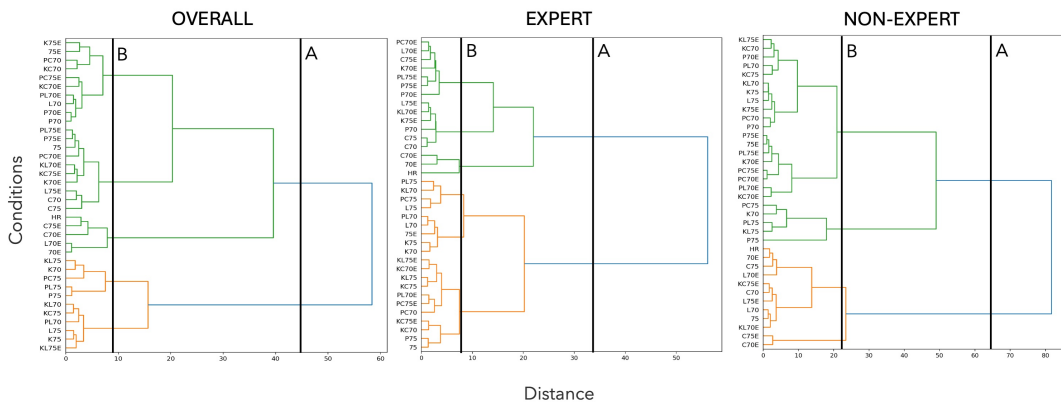


FIGURE 5.8: Dendrogram resulted from the hierarchical clustering analysis for expert and non-expert user groups and overall.

Users were able to easily operate the D&DMUSHRA+ GUI with stimuli divided into pages and navigate between experimental conditions and the reference. No user reported problems while operating the interface and after the evaluation procedure. User ratings for each condition are shown in Table 5.4. The condition with the highest score is PC75, although it does not present significant differentiation.

Considering all participants, the Mixed Linear Model Regression (MLMR) analysis (Fig. 5.6) indicated differences in HRTF between KEMAR HRTF and no HRTF ($z = -4.862$, $p < 0.000$) and between Personal HRTF and no HRTF ($z = -3.556$, $p < 0.0004$), in headphone compensation between equalized and not equalized groups ($z = 4.705$, $p < 0.000$), in volume between +70dB and +75dB ($z = -2.594$, $p < 0.009$), in reverb between calibrated reverb and large reverb ($z = 2.051$, $p < 0.04$). In the non-expert group, MLMR analysis underlines significant differences for HRTF between KEMAR HRTF and no HRTF ($z = -16.2266$, $p < 0.001$) and between Personal HRTF and no HRTF ($z = 19.3585$, $p < 0.0000$), for volume ($z = -6.1498$, $p < 0.04$), and headphone compensation ($z = 10.3536$, $p < 0.0004$). In the expert group, the MLMR test shows differences only in headphone compensation ($z = 7.7373$, $p < 0.02$).

Hierarchical clustering analysis (Ward linkage method - [303]), Fig. 5.8-overall) has been conducted in group conditions. By varying the number of clusters, different segmentation levels reveal interesting insights about the plausibility score. In the first considered level (2 clusters, level A), we observed a strong tendency

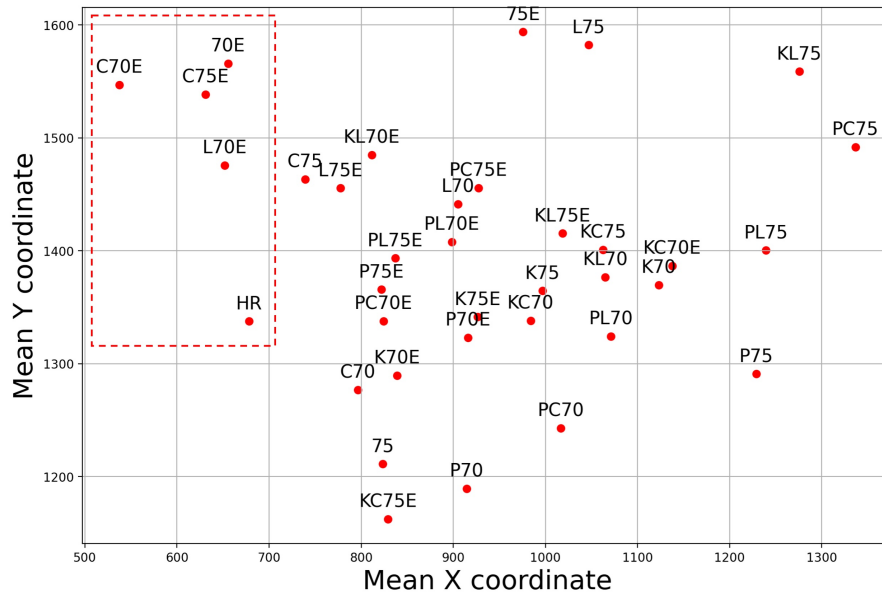


FIGURE 5.9: Mean coordinate in pixels at the end of the evaluation, for each condition, referred to iPad screen. The dotted box in the top-left corner shows the position of the conditions with the lowest scores identified by the hierarchical clustering analysis.

in headphone compensation. The first clade in the dendrogram contains a vast majority of compensated conditions (17 out of 25), while the other mainly contains conditions without compensation (10 out of 11). In the second considered level (5 groups, level B), an interesting observation was the emergence of a clade containing the 5 conditions with the lowest score (HR, C75E, C70E, L70E, 70E) characterized, apart from HR, by (i) the absence of HRTF, and (ii) the presence of headphone compensation. This group is also visually identifiable in Fig. 5.9, showing the average final position of various conditions inside the evaluation grid on the app GUI. Focusing on the top-left corner, one can identify such a group except for the HR condition that is in the center-left. Another group, containing the 5 highest rated conditions (KL75, K70, PC75, PL75, P75), is characterized by the presence of HRTF (2 Kemar, 3 personal), volume at +75dB (4 out of 5), and no headphone compensation.

The analysis of the separated dendrograms for experts and non-experts also shows interesting differences. For experts (Fig. 5.8-expert), in the first considered level (2 clusters, level A), again the first clade contains a vast majority of compensated conditions (12 out of 16) and many conditions with no compensation in the second one (14 out of 20). On the other side, non-experts, while still presenting a trend in preferring non-compensated conditions, show a smaller difference. In fact, in the first considered level (2 clusters, level A) the first clade contains only 11 out of 20 compensated conditions while the second 9 out of 16 non compensated conditions. At a deeper level, the analysis also underlines some interesting differences. The second considered level for experts (4 clusters, level B) contains a clade containing the conditions with lower expert ratings (C70E, 70E, HR) sharing no HRTF and volume calibration. Another one contains conditions with the highest expert ratings (PL75, KL70, PC75, L75) sharing the absence of headphone compensation, usually a higher volume (3 out of 4), and the presence of reverb (of which 3 out of 4 are large). For non-experts (Fig. 5.8-nonexpert), the second

level (4 clusters, level B) also contains a clade with lower scores (C75E, C70E) and a clade with the highest scores (PC75, K70, PL75, KL75, P75). The conditions contained in the first differ only in volume, indicating that users tended not to find plausible the absence of HRTF, headphone compensation, and calibrated reverb. On the other hand, conditions of the second one have no headphone compensation, higher volume (4 out of 5), and HRTF with no specific preference (3 personal, 2 Kemar).

Complementing the analysis described in Section 5.4.2, we performed one-way ANOVA analysis between experts and non-experts (Fig. 5.7). In particular, one-way ANOVA with a group factor shows that experts tend to evaluate conditions with lower scores ($z = 10.353$, $p < 0.0004$).

5.4.5 Discussion

The proposed evaluation method enables the reduction of experimental time, facilitating repeatability under varied conditions, such as different headphones, active noise cancellation settings, or environments, which could provide broader insights into how these variables impact the results. Given that this methodology applies to studies with a limited number of participants, it can be particularly effective in identifying dominant or disruptive elements. Furthermore, this approach can be particularly beneficial for pilot studies designed to identify potential limitations, specifically in the audio rendering process. As we scale the study to include more participants, the conditions that emerge as most promising in this phase will be prioritised for further investigation in subsequent experiments.

Piloting and Screening with D&DMUSHRA+

The use of the D&DMUSHRA+ interface required a very short time for each experimental session, with an average of about 42 seconds to evaluate a single condition. This can indicate the ease of use of the interface and the intuitiveness of the methodology. It is important to note that the primary aim of this work is to refine the evaluation methodology itself, rather than to conduct an in-depth study of audio AR. However, an additional outcome of this methodology is the identification of key conditions that can be further explored in future works. Notably, this identification has been achieved despite the small number of participants, underscoring the capacity of the methodology to effectively highlight the most relevant rendering factors even under limited-sample conditions and in a short amount of time. Similar approaches, which involve a rank-by-elimination procedure within a VR environment, are used in several studies. For instance, in a study by [296], 20 expert participants evaluated 12 conditions to assess hearing fatigue, with each evaluation taking about 10.8 ± 2.8 minutes. [308] used a similar method to evaluate speech intelligibility in different reverberation conditions, with a maximum session duration of 20 minutes to prevent hearing fatigue. Another example is the study by [15], where the same procedure was used to assess how different user tracking methods affected the perceived position of audio sources, with the entire process lasting approximately 20 minutes. In contrast, our methodology allows for the identification of key perceptual factors while keeping both the number of participants and the session duration low, and at the same time increasing the number of conditions under evaluation. This makes D&DMUSHRA+ particularly suitable for early-stage prototyping or scenarios where testing time and participant availability are limited.

We observed that the participants tended to compare the experimental conditions with each other rather than with the reference. While this behaviour is inherent to MUSHRA-like tests, it becomes more critical when a large number of conditions are presented. In such cases, listeners tend to stop revisiting the reference stimulus, and the perceptual memory of it fades over time, potentially leading to less accurate evaluations. This may lead to inaccurate comparisons and could have a significant impact on the evaluation results. However, this can be problematic when many conditions are available, as it can make the experiment lengthy and boring. To address this issue in future tests, the D&DMUSHRA+ interface could be modified to nudge participants to compare stimuli with the reference more frequently and/or introduce a gamified procedure. For instance, a mechanism can be introduced where the participants are required to listen to the reference after every specific amount of time.

Plausibility in Audio AR

In our case study, binauralized conditions are generally preferred to evaluate plausibility with higher scores, except for expert users, but there is no particular preference for personalised HRTF. However, it is crucial to note that the importance of HRTFs in a multimodal AR context remains an open question and seems to vary with user sensitivity and tracking methods ([65]), contrary to virtual reality, where its impact is generally more pronounced ([214]), although its relevance may vary depending on the specific VR use case.

Within our enhanced evaluation, the high-score clade exhibited the following characteristics: no headphone compensation, volume +75dB, and HRTF—whether personalised or generic—which provided the spatial directionality cues. In particular, the headphones compensation received a lower score than the conditions with no compensation. The possible reasons can be: (i) AirPods use adaptive equalization¹¹ that cannot be bypassed, and no specification from the vendor is available about the algorithm used, and (ii) calibration based on average curve measurements can lead to incorrect compensation and, thus, lower plausibility. Since the use of the *audio transparency mode* of AirPods is essential for perceiving the real environmental audio sources, it significantly alters the auditory perception of the environmental room, including compensation, soundwave phase, and latency ([260, 309]). This is further supported by [310], which underlines that using adaptive compensation while interaural time differences were largely preserved, transparency mode introduced systematic alterations in level differences, up to 8 dB, and eliminated spectral cues above 5 kHz. These modifications could result in a compromised spatial hearing ability, especially in noisy environments. However, the use of this device was intentional and aligned with the ecological objectives of the study. Rather than relying on laboratory-grade headphones with highly controlled acoustic behavior, the experiment aimed to reproduce realistic Audio AR usage conditions through portable, widely adopted consumer devices. This choice reflects the practical deployment scenarios of mobile Audio MR systems, where users typically experience augmented audio through commercially available wearable technologies. Consequently, the experimental setup prioritised ecological validity and real-world plausibility over strict electroacoustic control. From this perspective, the observed limitations should not be interpreted solely as methodological weaknesses, but also as indicators of the challenges involved in translating calibrated audio rendering into everyday listening

¹¹Technical specification: https://support.apple.com/kb/SP811?locale=en_US

devices. Accordingly, the trend in considering more plausible conditions with higher volume is evident and statistically significant. Although at first glance it may seem counter-intuitive, this result can be better understood considering that the real stimulus was mediated by the AirPods' transparency processing. Moreover, many subjects reported a high environmental noise during the experimental session, which was not controlled to obtain an ecological audio AR. This could have affected the perceived plausibility.

Furthermore, users tend to rate the plausibility of reverb with a similar score, preferring large reverberation overall, while assigning a similar score when virtual reverberation is absent. In the case of non-expert conditions with the lowest scores, all contain a calibrated reverb. In the post-experimental interviews, participants tend to describe the specific implementation of SDN reverb conditions as "metallic" and "unusual". The use of the SDN reverb technique, or its specific implementation, despite helping with audio source localisation when used with personalised HRTF ([214]), may sound audibly coloured to users in an AR environment compared to the real-room reverberation. Accordingly, the shape of the room, a corridor with a long prominent dimension, could have influenced the response of the SDN reverb, usually tested in "shoebox-shaped" environments ([240]), e.g., concert halls, offices, etc.

Another important factor to consider was user tracking, which can affect user behaviour and lead to inaccurate results ([311]). Although consumer electronics devices are cost-effective and portable, proprietary technologies can be challenging to control for accuracy and responsiveness. For example, AirPods lack information on IMU sensors, and informal feedback suggests the tracking may be prone to errors due to a lack of compensation.¹²

As a natural exploitation of these outcomes, this is a starting point for further investigations, which could focus on the factors and criticalities that emerged:

- Investigating conditions without equalisation.
- A more in-depth study on the impact of volume.
- Evaluating other types of HRTF personalisation.
- A deeper exploration of the differences between realistic reverberation and "aesthetic" reverberation.

Possible Limitations

As shown in our study, hierarchical clustering and mixed linear modelling foster the identification of consistent trends in plausibility judgments, despite a small number of participants, even with a small significance of common statistical tests. Accordingly, exploring alternative statistical methods could help uncover deeper insights or confirm the robustness of the observed effects, potentially involving techniques such as mixed-model analysis or non-parametric tests, depending on data distribution and the nature of the responses. These adjustments and considerations could help refine future experiments, leading to more reliable conclusions about the impact of different audio AR conditions.

It is also important to acknowledge that the sample size and composition could limit the generalizability of the results. Although this study was designed as a pilot, the representativeness of the participant pool remains an important

¹²Measurements extracted from an informal one-way pilot test exhibit a latency of 200 – 300 ms.

consideration. As this work forms part of a broader screening effort involving different populations, it would be interesting to verify whether demographic factors—such as gender—may influence spatial–auditory perception. This possibility should be kept in mind when interpreting the findings.

Furthermore, it is worthwhile to notice that the presence of noisy data can significantly affect the extraction of meaningful patterns, leading to decreased classification accuracy and hampered feature identification. In a systematic review, [312] highlighted the importance of noise management strategies, revealing that ensemble-based approaches are generally more accurate in identifying noisy instances, while single-model techniques tend to offer greater computational efficiency. To mitigate the impact of noise on pattern recognition, particularly in complex environments, Bayesian methods may offer a promising direction. Bayesian approaches consist of integrating prior knowledge into the analysis process, allowing a more robust analysis, and have already been successfully applied in MR contexts. For instance, [313] adopted Bayesian inference to validate the Virtual Reality Neuroscience Questionnaire (VRNQ), which connects perceptual quality to user discomfort during VR sessions. In their work, the Bayesian version of the t-test, Pearson correlation, and regression analysis have been used to evaluate experimental data in terms of Bayesian factor. In another example ([314]), Bayesian regression has been used to confirm practical equivalence in perceiving absolute distance in AR between realistic and stylized shadow rendering. Related to audio VR, Bayesian inference had been employed to evaluate the differences in mental effort on sound evaluations caused by different levels of detail in VR rendering, demonstrating that auditory judgment depends primarily on individual sound properties, regardless of the visual environment ([315]).

Reflections on Approximation and Perceptual Tolerance

The results presented throughout this chapter suggest that the relationship between technical fidelity and perceptual plausibility in Audio MR may be more complex than commonly assumed. Across the thesis, tracking precision and spatial audio accuracy are presented as critical components for achieving coherent and immersive Audio MR experiences. However, the experimental findings suggest a more nuanced interpretation.

While high-fidelity tracking and accurate spatial rendering contribute to perceptual realism, the reported results indicate that users may tolerate a certain degree of approximation without significantly degrading the overall experience. In some cases, even simplified interventions appear to produce perceptual benefits relative to the absence of augmentation, as shown in Fig.5.4.

This observation highlights that perceptual plausibility does not necessarily scale linearly with technological precision. Rather, users may rely on a broader integration of contextual cues, narrative consistency, and interaction coherence when evaluating immersive experiences. Within Audio MR, perceptual acceptance may therefore emerge once a minimum threshold of spatial consistency is achieved, beyond which improvements in technical accuracy yield diminishing perceptual returns.

These findings suggest that the relationship between technological fidelity and perceived immersion should not be interpreted as absolute, but rather as context-dependent and mediated by user expectations, task demands, and environmental complexity.

Consequently, approximate tracking and rendering methods may remain valid solutions in ecological or mobile Audio MR scenarios where computational efficiency and real-time responsiveness are necessary constraints.

5.5 Evaluation of Co-Immersion

Section adapted from publication [16]. The author contributed to the auralization pipeline and to the analysis of perceptual externalization.

In the context of audio AR, co-immersion is defined as the property of simulated sound sources to be perceived as belonging to a coherent auditory scene rather than being artificially superimposed [70, 316]. A common approach to evaluate co-immersion consists in assessing the ability of users to discriminate inconsistencies in acoustic rendering [266].

In this work, co-immersion is investigated in a fully virtual Audio Augmented Virtuality (AAV) scenario, enabling controlled and repeatable conditions. Multiple sound sources are rendered through headphones using dynamic binaural auralization [317, 19], where anechoic signals are convolved with measured room impulse responses (RIRs) [318].

Within this broader experimental framework, the contribution of this thesis focuses specifically on the role of *externalization* as a key perceptual factor supporting co-immersion.

5.5.1 Material and Methods

The experiment compares three reverberation conditions, referred to as H , B , and F_d (Fig. ??). All conditions share identical early reflections (ERs) represented in Higher-Order Ambisonics (HOA) and dynamically rendered according to head rotations. The late reverberation (LR) component differs across conditions.

The reference condition H employs the full measured HOA room impulse response (RIR). Condition B uses a simplified LR obtained through a binaural downmix of the HOA response using a dummy head HRTF [319], resulting in a static binaural rendering. Condition F_d replaces the LR with an artificial reverberator based on a Schroeder structure [320], whose parameters are automatically tuned to approximate the target response [321].

The rationale behind this design is that preserving dynamic spatial cues in the ERs is sufficient to maintain sound source localization and spatial impression [322, 323, 324], while the perceptual contribution of directional information in the LR is limited [325, 326]. Thus, the experiment isolates the perceptual impact of simplifying the LR component.

The auralization pipeline is based on dynamic binaural rendering of Ambisonics signals using individualized HRTFs. Anechoic speech signals are convolved with the corresponding HOA RIRs to simulate realistic acoustic environments [318]. Head tracking is implemented with 3 degrees of freedom (yaw, pitch, roll), enabling dynamic updates of the binaural rendering. Headphone compensation is applied to mitigate coloration effects using pre-measured transfer functions of the playback system.

The experiment was conducted using a virtual reality setup including a head-mounted display for visual rendering and head tracking, and high-quality open-back headphones for audio playback. The audio rendering pipeline was implemented using a digital audio workstation and Ambisonics processing tools [327].

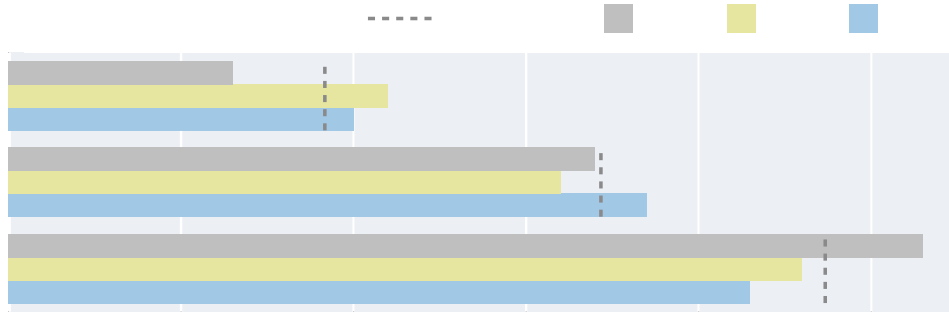


FIGURE 5.10: Externalization rates frequencies in the preliminary screening test grouped by reverberation condition.

5.5.2 Experiment design

Three rendering conditions were compared: a reference condition (H) and two simplified conditions (B and F_d).

Each trial involved a variable number of simultaneously active virtual speakers, placed in different positions within virtual acoustic environments (living room, classroom, concert hall). In the reference scenario (None), all sources were rendered using condition H . In the other scenarios (Bin and Fv), one source (the *impostor*) was rendered using a simplified late reverberation. In the Bin scenario, the impostor uses condition B , where the late reverberation is obtained through a static binaural downmix of the HOA response. In the Fv scenario, the impostor uses condition F_d , where the late reverberation is generated by an artificial reverberator tuned to approximate the reference response. All remaining sources are rendered using the reference condition H .

Participants were asked to perform two tasks:

1. *Detection*: determine whether all sources belonged to the same acoustic environment (i.e., presence or absence of an impostor).
2. *Identification*: when an impostor was detected, identify which source was inconsistent.

This paradigm allows co-immersion to be evaluated as the perceptual ability to detect acoustic inconsistencies, following a signal detection framework [328]. In this context, trials without impostors correspond to *noise*, while trials with impostors correspond to *signal* conditions.

Participants were instructed to focus on acoustic coherence, particularly reverberation characteristics, rather than spatial position or source identity. They were allowed to freely move their head and explore the auditory scene during each trial, enabling dynamic binaural perception.

Overall, the experimental design isolates the perceptual contribution of late reverberation to co-immersion, while controlling for spatial and contextual factors.

5.5.3 Results

Preliminary externalization test

The results of the preliminary screening test (Fig. 5.10) show that all rendering conditions are generally perceived as externalised. The reference condition (H) exhibited the highest rate of externalization.

The simplified conditions (B and F_d) also produced predominantly externalized perceptions, although with a higher proportion of responses classified as “on the edge” or “inside the head”. In particular, B showed a slightly higher tendency toward internalization, while F_d more frequently resulted in intermediate percepts.

Overall, the results confirm that all rendering conditions provided a sufficiently externalised auditory experience, supporting the validity of the experimental setup.

Detection performance

The detection results indicate a limited ability of participants to reliably identify acoustical inconsistencies between rendering conditions. Across conditions, hit rates remained moderate, and no statistically significant differences were observed between scenarios in the overall analysis.

While some trends emerged depending on environmental context and scene complexity, these effects were not consistent across conditions and did not substantially alter the overall outcome. In particular, simplified rendering conditions were not systematically more detectable than the reference condition.

Signal Detection Theory analysis

To further quantify perceptual sensitivity, we analyzed the results using Signal Detection Theory (SDT) [328].

The mean sensitivity values were $\bar{d}'_{Bin} = 0.35$ and $\bar{d}'_{F_v} = 0.63$. While d'_{F_v} was significantly above the minimum perceptual threshold, d'_{Bin} did not reach statistical significance.

Overall, these values indicate a relatively low sensitivity to the presence of impostor sources, suggesting that deviations introduced by simplified late reverberation were often perceptually subtle.

The response criterion c did not show significant bias, indicating that participants did not systematically favor either the presence or absence of an impostor.

Summary

Taken together, the results indicate that simplified rendering of late reverberation does not strongly affect perceived co-immersion. Despite measurable differences in rendering conditions, participants were often unable to reliably detect inconsistencies, especially in the case of binaural simplification.

Combined with the externalization results, this suggests that perceptual plausibility is largely preserved even under reduced rendering complexity.

5.5.4 General Discussion

In the literature, the Level Of Audio Detail (LOAD) [329] refers to the adaptation of the audio rendering complexity. LOAD techniques aim to reduce the rendered

details to save computational cost while minimizing the perceived quality degradation. Accordingly, our study can be interpreted as an evaluation of the LOAD required for LR to preserve the perceived co-immersion in AAV scenes. The obtained promising results on simplified LR have a potentially relevant impact on the saving of computational and storage resources compared to convolution with RIR or wave and geometric acoustics-based approaches.

We found that the detection tasks for Bin and Fv scenarios are very close to the chance level both accounting for detection rates and SDT metrics. Such results suggest the participants' inability to reliably detect the impostor in the proposed experiment.

For Bin scenario, we found that the mean sensitivity $\bar{d}'_{Bin}$ is not significantly greater than d'_{min} , thus it does not represent a meaningful effect. This result suggests that the LOAD can be reduced by substituting the dynamic rendering of an HOA LR with a static binaural version. Thus, we can answer to ?? that a static LR does not affect the co-immersion of a sound source in the evaluated AAV scenes.

For Fv scenario, the mean sensitivity $\bar{d}'_{Fv}$ is significantly greater than d'_{min} but is still very close to this value. To answer ??, this finding suggests that the static LR can be approximated with a simple artificial reverberator, such as Freeverb, with a perceivable but limited influence on co-immersion. It is worthwhile to notice that the LOAD is further reduced in Fv scenario because artificial reverberators are more efficient compared to convolution. However, it comes at the cost of a greater impact than Bin on the perceived degradation.

Finally, the LR matching method leads to satisfactory results if evaluated with the co-immersion criterion compared to the authenticity one reported by its authors [321]. Since in their study the matched reverberation was clearly distinguishable from the reference in direct comparison, co-immersion can be considered a narrow criterion that can be exploited in specific applications.

Although the LR simplified versions are difficult to detect, we found important differences with respect to the overall case by analyzing separately each VE and complexity. A significant finding is the easier detectability of Fv for the classroom compared to the other VEs. A possible explanation can be identified in the larger percentage errors of the matched classroom's RIRs for most of the acoustical parameters reported in ?? (ℓ_{end} , T_S and C_{80}). On the other hand, we found that participants were essentially unable to detect the impostor in the concert hall with Bin scenario. From an acoustic point of view, the directivity of LR is mitigated in a large environment, such as the concert hall. As a result, a static LR can be perceptually indistinguishable from a dynamic one in such an environment. For the living room, both Bin and Fv achieved low detection performances. This could be explained by the acoustic features of the living room whose short reverberation time leaves a small amount of information for the listener to focus on. This is an advantage for the application of our findings since indoor environments, like the living room, are common for end-users.

As in the overall case, Bin is less detectable than Fv in complexities 3 and 4, and in classroom and concert hall VEs. This was the expected behavior since the impostor condition *hoafv* is an approximated version of *hoabin*. However, we noticed that this pattern is reverted for complexity 2 and the living room. Though we found some statistically significant evidence of this pattern only for the logistic regression test (see ??) and not for d' values, this result represents an important finding for the use of an artificial reverberator to approximate LR. Nevertheless, for complexity 2, both Bin and Fv are quite detectable since is the

most simple complexity level. As the complexity increases the task is harder, and F_v is less co-immersed than Bin which is almost undetectable.

In this paper, we proposed an experimental design in an AAV framework originally intended for the case study of LR. However, it can be exploited to evaluate the perceptual impact of an arbitrary rendering approach of virtual sound sources in an AAV scene. The main advantage of designing such experiences in AAV scenes is that the challenges involved in the creation of AAR scenes with real-world sound sources are avoided (e.g., tracking, real-time and responsive interactions, high-quality simulations, to name but a few [70]). As a result, more controlled and repeatable experimental settings can be designed within AAV. In particular, one can decide which aspects of the physical reality to capture and which interactions to handle in VR. Furthermore, any real environment can be simulated with a pair of headphones by means of the acquisition of spatial RIRs in such an environment.

Although the focus of this work is on the timbral characteristics of the late reverb, which is mainly diffuse, we used personalized HRTFs to improve the stimuli's ecological validity. However, the usefulness of personalized HRTFs is debatable and strongly influenced by context and task. Visual “capture” effects, where visual stimuli associated with a sound source affect its localization in space, have been extensively studied in experimental psychology [330]. Similar effects were recently observed in VR using generic HRTFs [331]. Thus, in future works, we plan to investigate the effect of different approaches and degrees of HRTF personalization, as well as to compare our acoustic simulation approach to other auralisation engines and simulators, such as Project Triton [332], in relation to the level of HRTF personalization.

Chapter 6

Entanglement

The transition from the theoretical foundations of entanglement HCI to a concrete design practice requires tools capable of capturing the complexity of the "We-relationship" emerging between the user and the technological artifact. To this end, we introduce the Entanglement eXperience Map (EXMap) as a preliminary analysis instrument specifically tailored for egocentric audio Augmented Reality (AR). This mapping phase is not merely descriptive but serves as the essential precursor to the experimental evaluation presented in Section 6.2. This progression—from the preliminary "a-priori" design mapping to the empirical observation of "entanglement traces" ensures that design intentions are systematically aligned with the lived, emergent experiences discussed in the subsequent experimental results.

6.1 Entanglement eXperience Map

Section adapted from publication [10].

Frauenberger's entanglement HCI [5] suggests abandoning a user-centered design of the digital artifact in favor of participatory, speculative, and agonistic methods with the end goal of obtaining meaningful relationships and not merely optimized processes relating to the human pole or the machine pole, or their interaction. We may assume that entanglement follows the practice-based design of experiences and their emergent patterns, with methods, tactics, goals, and topics unfolding and changing as all actors adapt and learn in the course of their interaction. According to [223], the emergent nature of practice-based design research can lead to new insights, innovative solutions, and unexpected outcomes.

Critical viewpoints on design's interplay with non-human actors should serve as a vital proving ground and be systematically included in a design process. By defining, simulating, and visualizing the dynamic interactions, researchers can develop a comprehensive understanding of the sonic experience. In this section, we will introduce the Entanglement eXperience Map (EXMap), an analysis tool that is specifically designed for immersive experiences and egocentric audio technologies. EXMap incorporates by design the Actor network Theory [333] including the experience characterization introduced inSIVE, i.e., immersion, coherence, and entanglement, by reflecting upon a modified actor-network graph.

The proposed graphical representation is intended to be purely descriptive but can be supplemented with specific metrics and values pertinent to the specific domain within the evolutions of theSIVE research field. The constitutive elements of an EXMap are:

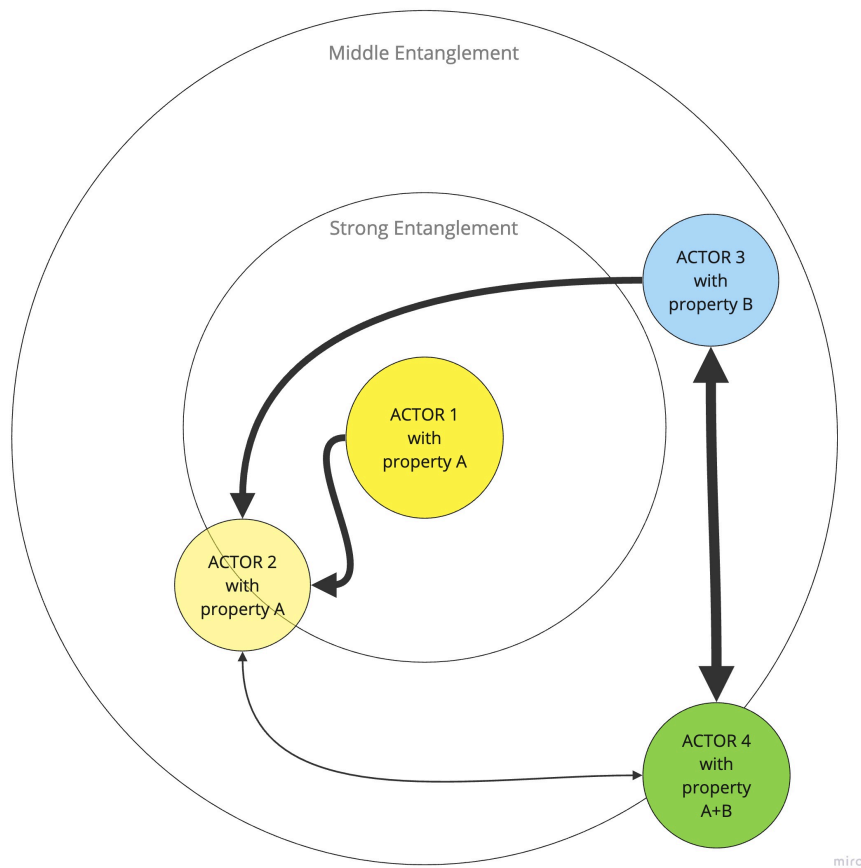


FIGURE 6.1: Graphical layout of an Entangled eXperience Map at time t for a hypothetical sonic experience.

- **Node:** As in actor-networks, each node represents an actor, which may be human or nonhuman, as well as real, virtual, or any models. Since we are working with entangled twins, virtual and real actors share the same action space. With respect to an actor-network map, these nodes can be connected or not based on their contribution to immersion and coherence in each phase of the experience, all in a dynamic and adaptable manner.
- **Arcs:** An oriented arc with weights representing a number of potential interactions between each pair of actors. The characterization of those interactions can dynamically define the network layer of immersion. Theoretically, the weight of arcs can be represented by the formula $\sum_{N=0}^i w_i$ where weight w_i of each kind of interaction can be a relative or absolute value assigned by designers or automatically computed. As the system can evolve throughout the experience, the weight of the arc can vary depending on the specific interaction that occurred. A possible example of interaction dynamics is proxemics. Distance between different actors inside a shared virtual/real space can influence the behavior of human actors, i.e., the users. In that case, the distance between actors can be used to determine, partially or totally, the weight of the corresponding arcs: arcs with a smaller distance will be represented by arcs with a higher weight.
- **Node Color Coding:** Color Codings are used to assign a semantic connotation

to the actor. Different colors indicate specific properties, e.g., human/non-human, real/virtual, etc.

- **Node Transparency Gradient:** The degree of transparency represents each actor's cognitive power over an action's expected outcome in the decision-making phase. Transparency can represent the level of coherence of a particular actor within the network. To aid in visualizing potential network issues, a node with high transparency represents an actor with a higher level of coherence. On the other hand, an opaque node represents a less coherent actor, which can lead to the so-called "break in presence" [49] within the experience.
- **Perspective:** A model based on the graphical representation in [334] has been created to visualize entanglement. The graph maintains an egocentric and monadic perspective that varies depending on the actor under consideration. As an actor is selected, the entire graph restructures, keeping the same relations, arc weight, and node properties but modifying the distance between the actor under consideration and the others: actors more entangled, i.e., in dynamic mutual re-configuration and intentional participation, are closer, whereas if they have less influence on each other, they will be further apart. Using different perspectives can help designers visualize how immersion and coherence contribute to maintaining entanglement during the experience. Moreover, emerging patterns of actions can be identified specifically for each actor.

Due to the dynamic nature of the entangled experience, the EXMap must be framed within a timeline as it evolves over time. Accordingly, digital media are the perfect display for supporting such a dynamic representation that varies depending on the monadic perspective of each actor under consideration.

6.1.1 Narrami: Actor-network Map

An actor-network map is a powerful graphical representation to identify system actors and the relations between them. Figure 6.2 shows an actor-network map for the previously described *Narrami* (Sec 3.7), in two different levels of detail. The color coding adopted in the Actor-Network Maps should be interpreted as a symbolic categorization rather than a perceptual or physical blending of colors. Although some nodes are represented using combined hues, these do not imply a continuous gradient or linear interpolation between categories, but rather indicate overlapping roles within the network. In Fig. 6.2-a, human and non-human actors within the entangled twins' relation can be defined as follows:

- **User:** The real user can directly interact with the real environment and indirectly influence the virtual environment using sensors and reproduction devices, with actions constrained by the technological setup.
- **Sensors:** This actor is responsible for user tracking in the real world. It includes hardware, e.g., Inertial Measurement Unit (IMU), accelerometer, gyroscope, eye-tracker, etc., and the algorithms that can be used for estimation and data fusion. Accurate tracking helps the system correctly represent space inside the VE, resulting in a coherent sonic experience [52]. The "Sensors" actor also directly relates to the AI engine, which uses data to generate a narrative and select the correct audio content. Moreover, the spatial audio engine uses user position and orientation to correctly position the virtual audio elements in the VE.

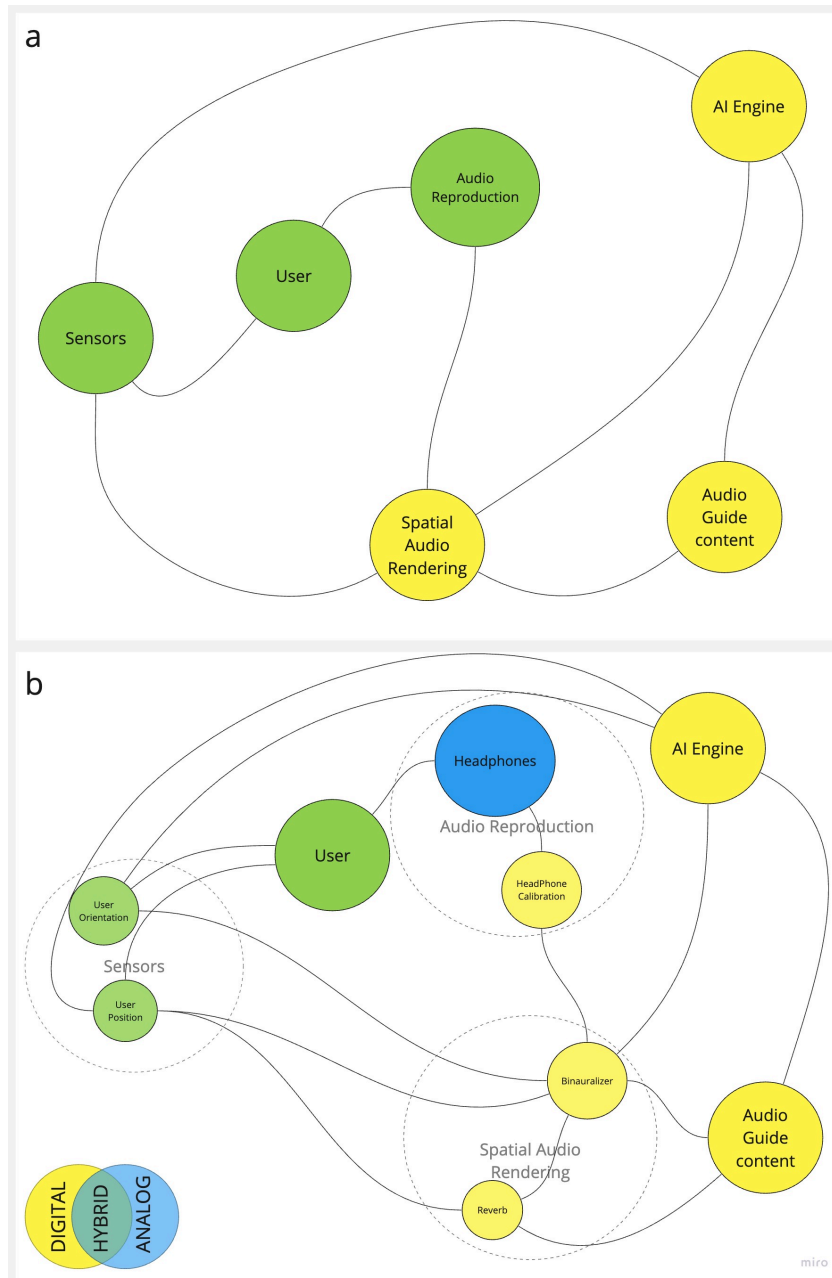


FIGURE 6.2: Actor-network map of *Narrami* with different levels of detail: (a) low-level details, and (b) higher-level details. At the bottom of (b), we provide a color descriptor supporting this example's categorical distinction.

- **Audio guide content:** This actor contains all the audio elements selected by museum curators to be used by an AI engine for generating narratives.
- **AI engine:** This actor uses data collected from sensors, museum curators selected contents, and spatial audio rendering to generate a personalized narrative for the user. Narrative generation could be implemented by integrating generative AI algorithms and reinforcement learning techniques [335].
- **Spatial audio rendering:** correct audio spatialization is fundamental for immersion in the VE (the reader can refer to the “Interactive and Immersive Audio” part of the recentSIVE book by Geronazzo and Serafin [19]). Audio is spatialized, by using the user’s position and orientation to correctly position audio sources within the VE, to be correctly perceived by the user via the audio subsystem.
- **Audio reproduction:** It contains the equipment for correctly reproducing sound generated during the audio rendering and is delivered to the user by using headphones provided by the museum.

Node colors represent the type of actor. Yellow nodes represent digital nodes, i.e., software components or data. Blue nodes represent analog actors that include physical elements or hardware devices. Some actors can contain both digital and analog components and are represented by green nodes. In *Narrami* actor-network maps, the user is represented by a green node, as it contains the embodied physical user that moves within the natural space (blue color), and all the objectification of the user’s configurations collected and learned by the system, including preferences, behaviors, emotional and psychological state, etc (yellow color).

Fig. 6.2-b provides more details for some specific actors of the *Narrami* cyber-physical system. In particular, the visualization shows that sensors, audio reproduction, and spatial audio rendering actors are now constituted by different parts. Sensors contain information about user orientation and user position. Audio reproduction is divided into the physical headphones and the algorithms that could help a more realistic audio rendering in terms of perceived audio quality, which is strictly related to the reproduction device. The spatial audio rendering actor is composed of two main parts, which includes the algorithms for enabling the users to correctly perceive virtual sound direction in space (binaural rendering) and the acoustics of a museum room (reverberation). As the details are increased, the map can contain more elements and show more specific relationships between the actors.

Finally, Fig. 6.2 provides a graphical representation of the connections between the actors defined in *Narrami*. This visual representation is crucial because it highlights that the user is not directly connected to the VE but the DT mediates based on sensors’ movement data collected within the physical environment. On the other hand, AI reinforcement learning algorithms rely on imperfect sensor data to generate content. This leads to unintentional interactions [52] and a partial adaptation of user behavior to the system. It is worthwhile to note how actor-network allows the identification of these relationships.

6.1.2 Narrami: EXMap

Although actor-network is a fundamental tool, identifying the specific characteristics of a sonic experience, i.e., immersion, coherence, and entanglement, remains challenging.

The proposed EXMap can be positioned in relation to existing approaches in HCI and network-based representations. While ANT provides a conceptual framework to describe relationships between human and non-human actors, it does not offer operational tools to capture the dynamic and experiential qualities of such relations in immersive environments.

In Fig. 6.3, three distinct EXMaps for different perspectives and phases of *Narrami* are depicted. Fig. 6.3-a and Fig. 6.3-b depict EXMap centered on the user during the system's initial phase and museum room navigation, respectively. Figure 6.3-c depicts the network centered on the sensors at the start of navigation. In this representation, the same actors were used as in Fig. 6.2's actor-network, and the corresponding nodes have the same color. On the other hand, the new representation includes the peculiar characteristics of egocentric audio technologies and SIVE.

Node Transparency represents the actor's coherence within the network. Particularly in Fig. 6.3-a/c, AI engine and Audio Guide have a low level of transparency, indicating that the user perceives both components as artificial within the egocentric system. As the experience progresses (Fig. 6.3-c), AI Engine adapts its response to the user's preferences, and the component becomes more consistent with the user's expectations. In addition, the user adapts its behavior to the MR, and the spatial audio rendering is perceived as more coherent. In this case, the transparency of the two nodes is increased. In the graphical representation provided by the EXMap, a node with low transparency (opaque) indicates an issue with the overall experience and is easily identifiable. The use of oriented arcs and arc weights is another feature of EXMap that makes it easier to analyze the mutual interaction between different actors. For instance, Fig. 6.3 shows that users strongly interact directly with sensors that have no direct effect on them in both phases. Nonetheless, Fig. 6.3-b demonstrates that the AI engine significantly impacts the audio content and a high level of interdependence with sensors despite not directly affecting it. Consequently, this illustration demonstrates that accurate tracking may be essential for a coherent and accurate audio experience. Moreover, AI engine influence over spatial audio rendering changes during the experience. In the initial phase, virtual sound starts with a random or standard position. As the experience progresses, the AI Engine indicates the correct location of all sound sources in the virtual space based on user movements and behaviors. Using the weight of the arc (thickness) between nodes representing the AI engine and spatial audio rendering, this change can be identified in the EXMap.

Entanglement is an important component of sonic experiences. Using the distance between nodes, EXMap depicts the degree of mutual influence between actors. In the initial phase, the level of entanglement between the user and the AI engine is moderate, as the algorithm has limited information about user behaviour and cannot provide narratives and content with a high level of engagement. During navigation, the AI agent indirectly receives information about the user from the sensors and modifies the generated content. The user receives engaging content and provides positive feedback, such as an increase in eye fixation,

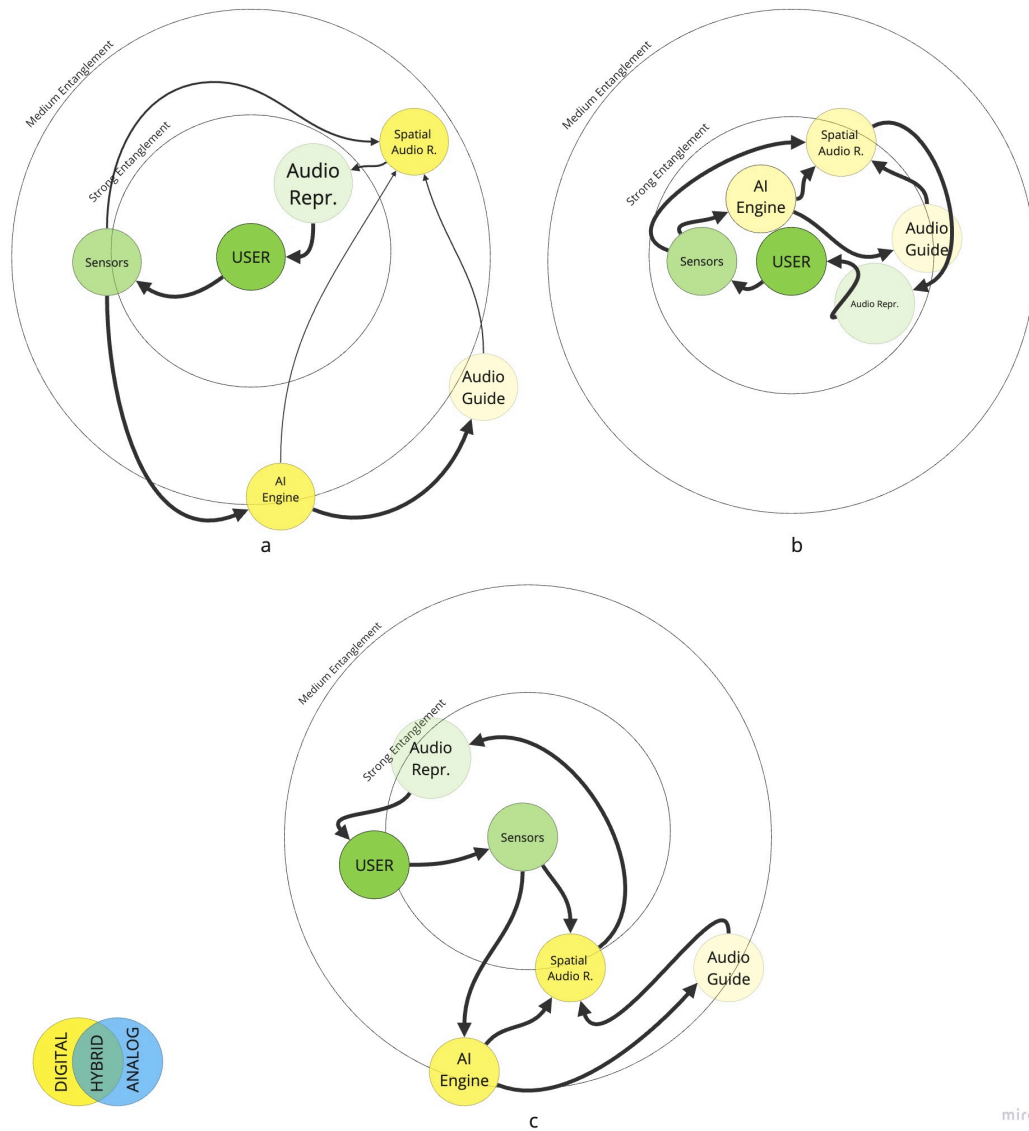


FIGURE 6.3: Entanglement eXperience Map of Narrami: a) Centered on the user, "Start of navigation phase". b) Centered on the user, "Personalized recommendations phase". c) Centered on the sensors, "Start of navigation phase".

patterns of head movement, etc. In comparison to the initial phase, mutual interaction and entanglement increase. It is possible to observe in the EXMap that the two respective nodes are closer. Similarly, as the experience progresses, all other actors tend to have a higher mutual interaction and adaptation with the user.

Similar to actor-network maps, EXMaps can have different levels of abstraction, depending on the information needed during the design phase. For example, as in Fig. 6.2, the Sensors node can be divided into User Orientation and User Position, and the degree of immersion, coherence, and entanglement of the two parts can be analyzed in more detail. For the sake of brevity and to avoid overburdening the theoretical speculation, we maintain a low level of detail in order to prioritize the explanation of the fundamental mechanics.

6.2 A Foundational Experiment

The work presented in this chapter builds upon an ongoing research effort currently being prepared for journal submission.

In this concluding Section, we addressed a complex experience in which a virtual agent interacts with a human user, realizing a genuine SIVE. To rigorously evaluate this experience, we employ the M-I-C-E framework described in Chapter 2 and Technological platform described in Chapter 4. For the purposes of this study, we chose to build upon an existing experimental framework [239], which provides a solid foundation for examining the shared intentionality characterizing ETs.

The study of Kastanis and Slater, involving 30 participants across three conditions based on proxemics where participants is particularly suitable for introducing the M-I-C-E methodology since it combines several key features: the use of a transparent and easily explainable RL algorithm, an adaptive and immersive simulation environment, the application of proxemics as a social relational construct, and the demonstration of persuasive technology in a general context rather than a domain-specific application.

6.2.1 Immersive and Adaptive Simulation

Immersive and adaptive simulations allow participants to experience dynamic environments where virtual agents can respond in real time to user actions [1]. Kastanis and Slater implemented an immersive simulation featuring a digital agent capable of adaptive behaviour, providing an interactive VE in which participants' movements could be influenced by the agent's responses.

6.2.2 Proxemics

In human communication, an important factor that influences behaviour is proxemics, which refers to how people use and perceive space and how this affects their interactions with others. Studying proxemics models, i.e., working with the implicit social rules of interpersonal distance between people [336], is relevant for the understanding of different social meanings. In his work, Hall identified four different proxemic zones: *Intimate distance* ($<0.46\text{m}$) for close friends and family members, *Personal distance* ($\geq 0.46\text{m}$, <1.2) for conversations with friends), *Social distance* ($\geq 1.2\text{m}$, <3.7) for interactions with strangers, *Public distance* ($\geq 3.7\text{m}$, <7.6) for public interactions. The amount of space that people maintain between themselves and others can communicate a variety of intentions and meanings,

such as their level of comfort, their relationship with the other person, and their cultural background.

Recent work extended proxemic theory to non-human entities, showing that navigation methods significantly influence proxemic behaviour in virtual environments with conversational agents [337] and virtual agents [338]. Proxemic frameworks, even with specific adaptations [339], can also be applied to human-robot interactions [340]. In this context, Kastanis and Slater [239] tested whether proxemic principles apply to virtual agents by using a $Q(\lambda)$ RL algorithm to control an avatar's spatial positioning in an immersive virtual corridor in three conditions: Intimate with 0.38m approach distance, Social with 41.2m minimum distance, and Random with no algorithmic guidance. Participants were unknowingly guided toward a target location through the avatar's proxemic behaviour. Results showed that intimate proxemic distances significantly accelerated participant movement to the target compared to social and random conditions, demonstrating that proxemic theory remains valid in virtual environments.

6.2.3 Explainability

Explainable algorithms provide interpretable insights into the decision-making processes of adaptive systems [341]. The algorithm used by Kastanis and Slater can be considered explainable in a tabular $Q(\lambda)$ reinforcement learning context, as the spatial behaviour of the agent is composed of few states (3 or 4 depending on the condition) and can be traced and understood in terms of learned state-action values, facilitating the analysis of how proxemic adjustments influence participant behaviour.

6.2.4 Persuasive Technologies

Persuasive technologies are designed to influence users' behaviours without coercion [342]. These technologies leverage principles from different disciplines to encourage desired behaviours, often by providing adaptive interventions and are proved to be effective in many domains [343, 344, 345]. By employing an adaptive digital agent, the Kastanis and Slater experiment exemplifies persuasive technology operating without explicit instructions shaping user behaviour while maintaining the sense of autonomy.

6.3 A-Priori Interaction Design

A priori interaction design is critical when setting up experimental conditions that must remain stable and replicable across multiple users or iterations [54]. This serves to anticipate possible user actions and system responses, thereby shaping the interactive landscape before any empirical data is gathered [346]. For the purposes of this study, the experimental framework [239] previously described in Sec. 6.2, provides a solid foundation. In this context, design choices include the types of interactions permitted, the modalities involved, and the expected user responses [347]. Re-adapting the original research hypotheses to our audio-based AR study, we posit two researchers' hypotheses (RHs) which belong to the design team:

- RH1 Participants guided by the RL agent will reach the target area more frequently and in less time than participants in the random guidance condition.

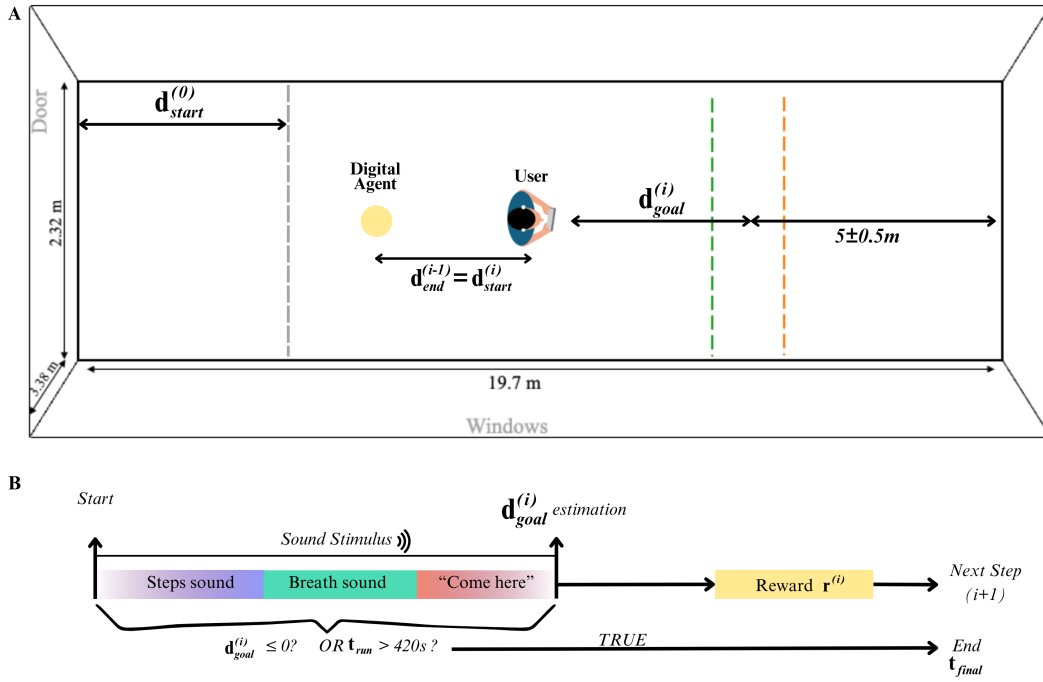


FIGURE 6.4: Schematic representations of the experiment in AR. a) Space arrangements for the user and the digital agent. The experiment starts with the digital agent placed close to the door and the user at the gray-dashed position yielding to a $d_{start}^{(0)}$ relative distance between the two. The area between the green and orange dotted lines identifies the goal region. b) The digital agent's RL cycle. For step i , ego-centric distance estimation ($d_{start}^{(i)}$, $d_{end}^{(i)}$, $d_{goal}^{(i)}$), temporal controls (t_{run} , t_{final}), and rewards ($r^{(i)}$) are computed.

RH2 Proxemic configuration will modulate guidance consistently with proxemics theory.

In designing the AR experience, several real-world constraints must be addressed to ensure that the experiment remained ecological, reproducible, and effective, and, for this study, we relied on the technical setup described in Chapter 4.

6.3.1 AI Agent

The $Q(\lambda)$ RL algorithm was employed to select the agent actions, following the specifications outlined in the original experiment [239]. $Q(\lambda)$ integrates parameters that bridge Monte Carlo methods and temporal-difference learning. This accelerates learning by efficiently propagating rewards, helping the agent to learn from sparse and delayed rewards. For this experimental design, the $Q(\lambda)$ parameters were set as follows: actions with the highest reward for the specific state were chosen 30% of the time, while random actions were chosen 70% of the time ($\lambda = 0.7$). The algorithm's learning rate (α) was set to 0.2, the discount factor (γ) was set to 0.15, and an ϵ -greedy policy improvement was used with ϵ set to 0.15.

The algorithm calculated the training space coefficients online during the experiment to adapt the RL agent's behaviour to each specific user. Fig. 6.4-b represents the timeline of each interaction step of the algorithm. The action space comprised 5 agent actions: moving forward or backward 1m or 2m (respectively

F1, F2, B1, B2), idle (ST). Each action had a different time computed using a fixed part based on the chosen action (0.8s for F1, 1.5s for F2, 1.8s for B1, 3.2s for B2, 0s for ST). During walking actions, a step sound was spatially generated from the voice agent's feet. After each action, except ST, the agent voiced "Come Here", with slight variations (*Hey, come, Come on, join me, etc.*), followed by a breath sound. The second part of the action took a random time between 2.4s and 2.8s. At the end of each step i , the RL algorithm estimated the relative distance between the user and the agent, $d_{end}^{(i)}$, to determine the reward and the next action. Respect to the original experiment, the breathing sound served a dual purpose: (1) to provided a continuous auditory cue that reinforced the perception of the virtual agent's proximity, thereby heightening the sense of presence, (2) to elicit feelings of discomfort or unease [348], to subtly influence the participant's behaviour, encouraging movement towards the target area. The state space was defined by the user-agent. Based on the user-target distance, $d_{goal}^{(i)}$, the RL agent's actions were rewarded as follows:

$$r^{(i)} = \begin{cases} 1 + \left(d_{goal}^{(i-1)} - d_{goal}^{(i)} \right) & \text{if } d_{goal}^{(i)} < d_{goal}^{(i-1)} \\ -1 & \text{otherwise} \end{cases} \quad (6.1)$$

Where $d_{goal}^{(i-1)}$ is the user-target distance at the beginning of the algorithmic step i and $d_{goal}^{(i)}$ is the distance at the end of the step. The algorithm ended when the participant reached the target area ($d_{goal}^{(i)} = 0$) or when the total experimental time $t_{run} \geq 420s$. To ensure a high level of plausibility, volume levels of auditory stimuli (voice, footsteps, breathing) were calibrated to match real-world SPL values (voice set to 60 dB SPL at 1.0m [349], footsteps at 37.6/44.7dBA [350], breathing set to 10dB SPL [349]). The calibration procedure involved the spatialised sounds being played through the AirPods placed on a silicone ear located at the measurement input of a sound level meter to ensurey.

Finally, if the participant moved backwards and reached the minimum distance allowed by the experimental condition, the virtual agent would adjust by moving backwards to maintain the prescribed interpersonal distance. On the other hand, if the virtual agent reached the end of the corridor, if a backwards movement was selected by the algorithm, the agent would instead remain stationary. Contextually, the agent immediately ended its movement but continued any corresponding voice and breath action, ensuring the total time remained constant.

6.3.2 Procedure

At the beginning of the experimental session, the participant was introduced to the study and provided with preliminary information. She/he was then asked to read the general information sheet and fill out an informed consent form and a questionnaire that included details such as their age and previous experiences with audio AR. Next, the experimenter took two photos of the participant (front and side of the head) to create the personalised HRTF to be loaded into the iPad app. In the meantime, the participant was provided with instructions for the experiment and how to use the simple User interface (UI) available on the iPad. The

instructions also advised that if the iPad lost user tracking, the screen would become black and the interface unusable. The experimental procedure was divided into two parts: the training and the main part.

Training

The participant was positioned on a virtual dot in the centre of the room and asked to move toward a voice agent saying “Come here” without turning towards it. The training started by pressing the “Start” button on the iPad UI. The voice agent’s position was randomly chosen between the front and the back of the room. When the participant reached the agent, it moved to the other part of the room. The training ended when the participant reached the voice in the second position.

Main experiment

Participants were asked to move to the starting position $d_{start}^{(0)}$ (see Fig. 6.4-a). The experimenter informed the participant that the aim of the experiment was *to feel comfortable*, while another person was entering the environment. To avoid any anxiety or uneasiness, she/he was free to walk backwards or forwards. During the experiment, the virtual agent moved according to the RL algorithm described in subsection 6.3.1. The experiment ended when the participant reached the goal area or a total time of $t_{run} \geq 420s$ (7 minutes) had passed from the start of the main experiment session, and the goal had not been reached. At the end, a short interview with open-ended questions was proposed to the participants (see Supplementary Material).

Experimental Conditions

The experiment comprised a between-subjects design with three distinct conditions: social, intimate, and random.

- **Social (S):** The agent maintained a minimum distance of 1.2 m, ensuring it did not enter the personal space. This condition included three states for the $Q(\lambda)$ algorithm: social, public, and not engaged.
- **Intimate (I):** The agent avoided entering the intimate distance of the user, maintaining a minimum distance of 0.46 m. This condition included four states for the $Q(\lambda)$ algorithm: personal, social, public, and not engaged.
- **Random (R):** The agent operated without the RL algorithm and included four states, as in I .

Depending on the experimental condition, the action space consists of 3×5 states for S and 4×5 for I .

6.4 From Experiential to Data Space

The experience involved 30 participants (24 male, age 25 ± 6 , 6 female, age 35 ± 19) with self-reported normal hearing and was held at the Department of – *anonimised for review* – and was approved by the local ethical committee. Participants

were randomly assigned to the three conditions, with 10 individuals per condition.

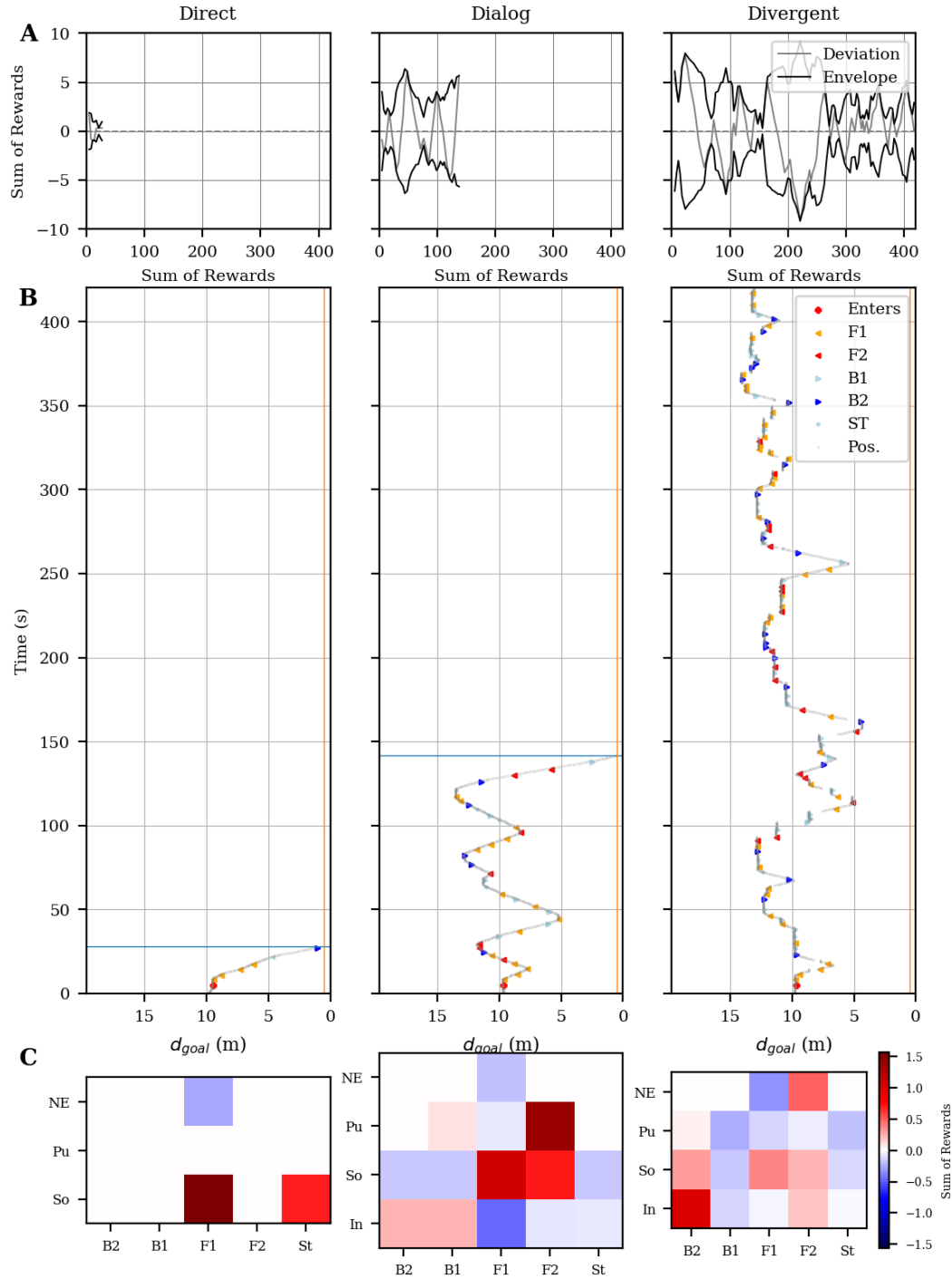


FIGURE 6.5: Three examples of reward curve, action taken and state space for each cluster. a) Hilbert transform of the sum of rewards deviations. b) User movements in terms of egocentric distance from the goal (d_{goal}) and agent's decisions, i.e., performer actions. The orange line identifies the final spatial goal. c) Final state matrix of the digital agents.

From the *researcher's perspective*, as in the original experiment [239], we evaluated the performance of the RL algorithm by analyzing the total percentage of successful reaches for each experimental condition with ANOVA and the cumulative distribution of t_{final} .

For the *computer perspective*, the collected data encapsulates various dimensions of algorithm behaviour. Specifically, for each computational step i , we logged the reward scores, $r^{(i)}$, the action-state space evolution, the type of executed actions, user-target distance at each step, d_{goal} , and the overall session duration t_{final}). Figure 6.5 shows three examples of collected data from the main experimental session. We analysed the $r^{(i)}$ as quantitative indicators of how effectively the agent's actions influenced the participant's movements toward the target area. The analysis comprised several steps. Figure 6.6 shows an example of the extracted data for a sample participant. The following processing steps lead to the indicators of the variability in the agent's influence over time:

1. The cumulative sum of rewards $R_{cum}(k) = \sum_{i=1}^k r^{(i)}$ accumulated throughout the session (Fig. 6.6-a). This time series represented the agent's overall performance in influencing the participant over time.
2. The linear regression on the cumulative reward time series $R_{reg}(k) = \alpha + \beta t \approx R_{cum}(k)$ for each participant (Fig. 6.6-b, blue line). The regression line represented the expected progression of rewards if the influence was consistent and linear over time.
3. Deviation between the actual cumulative rewards and the regression line $Res(k) = R_{cum}(k) - R_{reg}(k)$ (Fig. 6.6-b - black line), highlighting fluctuations in the agent's influence and indicating periods of stronger or weaker effectiveness.
4. The Hilbert Transform $\mathcal{H}(Res)(k)$ to obtain the absolute value of the analytic signal (Fig. 6.6-c), effectively capturing the envelope of the deviations of the agent's influence.

From the envelope, we derived the number of peak points $x_{peaks} = peaks \left[\mathbb{E}[\mathcal{H}(Res)(k)] \right]$, quantifying significant fluctuations in the envelope. K-Means was then applied to cluster data into groups.

For the *participants' perspective*, post-experiment interviews served as the primary means of collecting feedback, in line with established methodologies for assessing presence in VEs[351]. As a design choice, we did not use standard questionnaires, and this approach aligned with the exploratory nature of the study. Several established questionnaires exist to measure aspects of user experience in VR/AR, such as the Presence Questionnaire [352], the ITC-Sense of Presence Inventory [353], and the Game Experience Questionnaire [354], but none of them fully capture the intertwined dimensions of I-C-E. Moreover, the use of more than one questionnaire could have extended the duration of the experiment and potentially affected the participants' feedback.

To ensure data accuracy and reliability, all interviews were audio-recorded with participants' informed consent. The recordings were transcribed using automatic speech recognition software (Notta¹). Human supervision followed this step, where each transcript was reviewed and corrected to address any errors introduced by the automated transcription.

¹Available at <https://notta.ai>

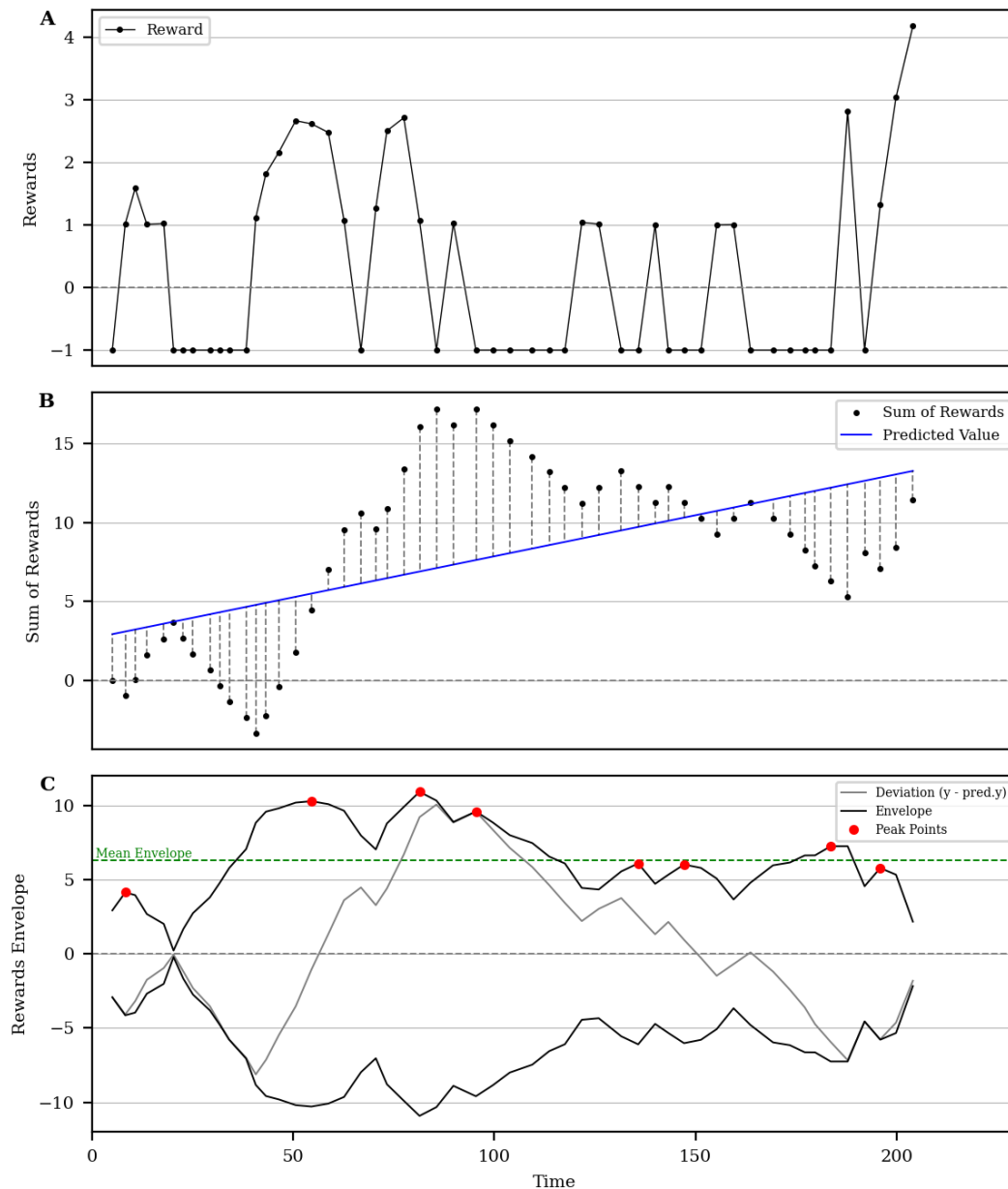


FIGURE 6.6: An example of analysis on the Data Space of participant U2336: a) Rewards over time. b) Sum of rewards and linear regression (blue solid line). c) The envelope (black solid line) of reward's deviation (grey solid line) computed via the Hilbert transform. Peak points (red circles) and average envelope (green dashed line) are also reported.

To gain deeper insights into the participants' subjective experiences, we conducted a thematic analysis of the post-experiment interviews using *Lumivero NVivo 15*², a qualitative data analysis software. NVivo facilitates the organisation, coding, and analysis of large volumes of text, allowing for easy grouping of similar concepts to identify underlying patterns and themes within participants' responses [355].

Finally, the exploration of patterns within the participants' identified themes,

²Available at <https://lumivero.com/products/nvivo/>

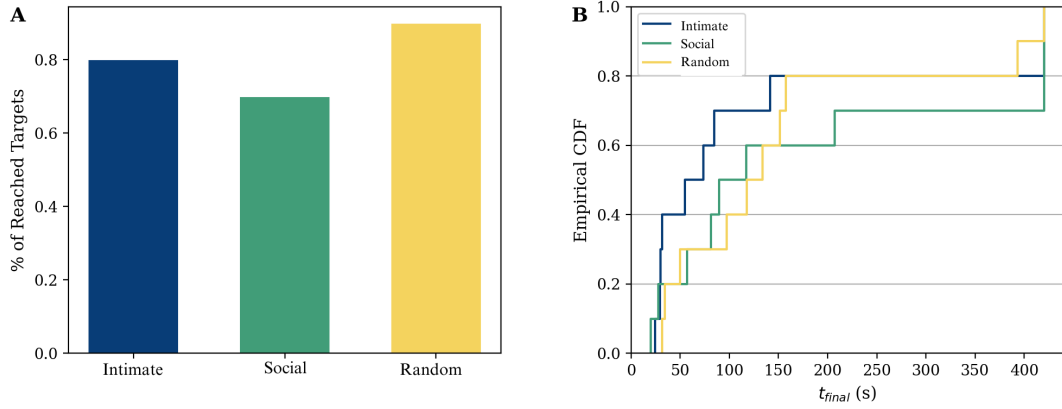


FIGURE 6.7: Researcher's viewpoints analysis. a) The percentage of participants reaching the target under the three conditions. b) Empirical cumulative distribution functions for the three conditions.

we performed clustering using the Gower distance metric [356], which is suitable for categorical variables.

6.5 Intra-Action Analysis: Single Perspectives

In alignment with the M-I-C-E methodology, our analysis examines the experiment from multiple perspectives, including the experiment designer and researcher, the user of the system (participant in our experiment), and the system itself. We began our investigation from the researcher's perspectives and subsequently entered the system by adopting the computer's viewpoint, to further characterise the system's internal logic, by examining the types of choices made by the algorithm. Finally, we compared these findings with the human participants' perspectives by employing the interviews in order to better understand the dynamics of the ETs within the interactive experience.

6.5.1 Researcher's Viewpoint

From the researcher's viewpoint, the primary objective was to analyse the effectiveness of the virtual agent's influence on participant behaviour across the different experimental conditions-I, S, R. Drawing inspiration from the approach used in Kastanis and Slater's experiment [239], we focused on evaluating algorithm's rewards $r^{(i)}$ to quantify the agent's impact on participants.

For the analysis, the primary response variables included the success rate (participants who reached the target in $t_{final} < 420$, Fig. 6.7-A), and the cumulative distribution functions (CDFs) of each condition's survival times (Fig. 6.7-B).

Our analysis observed that participants in the Random condition demonstrated the highest success rate (90%) Conversely, participants in the Intimate condition tended to reach the target more quickly (mean t_{final} was $I = 131 \pm 156s$ for I, $186 \pm 169s$ for P, $158 \pm 138s$ for R). Although post hoc analyses did not indicate significant differences among the conditions when applying the methodologies outlined by Kastanis and Slater, our findings identified a similar for I and S conditions compared to the original VR experience. Whereas Kastanis and Slater's study concluded at the comparison of experimental conditions, the present work

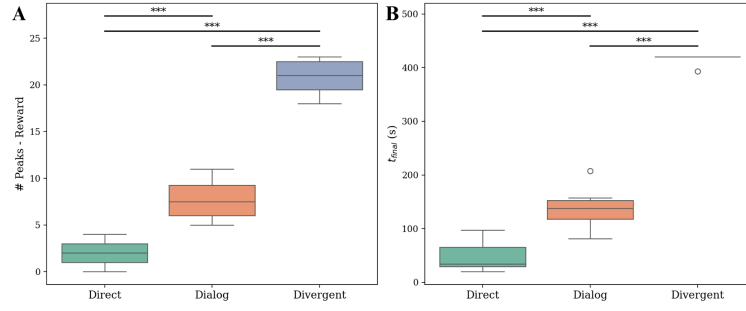


FIGURE 6.8: Aggregate statistics (average and standard deviation). a) x_{peaks} grouped by reward clusters. b) t_{final} grouped by reward cluster. Participants in the Divergent cluster did not reach the goal, i.e., $t_{run} > 420s$. Asterisks and bars indicate a significant difference (*: $p < .05$, **: $p < .01$, ***: $p < .001$ at post-hoc test).

extends their perspective by broadening the analytical scope beyond the original framework.

6.5.2 Computer's Viewpoint

From the computer's Viewpoint, particularly focusing on the virtual agent's decision-making process, we adopted a *global* perspective aimed at characterising different reward curves $r^{(i)}$, and a *local* perspective, consisting of the analysis of the agent's actions.

Rewards

Using the feature extracted from the agent's movement log, we conducted a cluster analysis on x_{peaks} . Before applying K-Means, the analysis using the Elbow Method [357] and Silhouette Analysis [358] indicated that the optimal number of clusters was three. K-Means was then applied to organise the data into groups (Fig. 6.8).

In the first cluster, named *Direct*, the cumulative reward curve was nearly linear, generally having few x_{peaks} , indicating consistent progress toward the target with minimal fluctuations. Participants in this group probably tended to respond directly to the agent's cues, moving toward the target area quickly (Fig 6.5, first column). Moreover, t_{final} is significantly shorter in this group with respect to the others. Notably, many participants in this group were from the I condition, suggesting that closer proximity enhanced the agent's influence.

In the second cluster, *Dialogue*, cumulative rewards reflect moderate deviations from the linear progression. Participants in this group showed periods of compliance interspersed with hesitation or exploration.

The last cluster, *Divergent*, presented longer t_{final} , high x_{peaks} , indicating substantial deviations and prolonged interaction times. Participants often did not follow the agent's cues and displayed inconsistent movement toward the target. A Tukey Honest Significant Difference (HSD) post-hoc analysis with Bonferroni correction among the three groups revealed statistically significant differences on t_{final} ($F(2,27) = 447.549$, $p \ll .001$ on all comparisons) and x_{peaks} ($F(2,27) = 283.884$, $p \ll .001$ on all comparisons). The peak-related feature thus provides an effective explanation of goal-reaching performance and can be further examined in its local patterns, potentially revealing explainable structures.

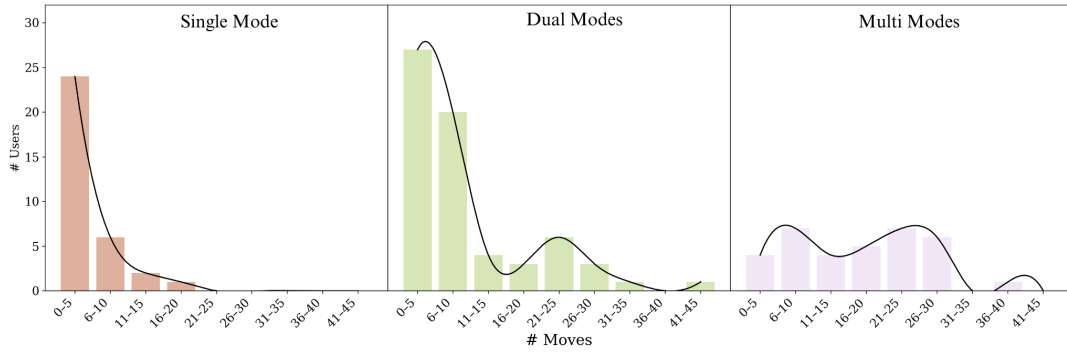


FIGURE 6.9: Computer viewpoint. Distribution of the digital agent's moves across peak clusters with the number of moves grouped by 5. Data are approximated using a third-degree polynomial curve.

Agent's Moves

We expanded the analysis, taking into account the virtual agent's local patterns, i.e., available actions. For each participant, we computed (i) the number of times each action type was selected by the agent throughout the session, (ii) the mean number of actions per type, and (iii) identified the actions that exceeded this mean in (ii). Fig. 6.9 illustrates how these groups align with the statistical modes.:

- **Single action type (*Single Mode*):** This group consists of participants for whom only one action type exceeded the mean (10 participants). All participants in this category reached the target and came from I(6) and S(4) conditions, with no representation from the R condition.
- **Two action types (*Dual Modes*):** This group includes participants for whom the agent used two different action types above the mean (14 participants). The majority (8 out of 14) were from the R condition, while the remaining participants were from the I(4) and S(2) conditions. Three participants did not reach the target.
- **More than two action types (*Multi Modes*):** This category comprises participants for whom more than three action types exceeded the mean (6 participants). This group was mainly composed of S(4) and R(2) conditions, with no participants from I. 50% did not reach the target.

6.5.3 Human Participant Viewpoint

For the human participant viewpoint, we identified the main themes for each interview by focusing on participants' subjective experiences, ensuring a broad spectrum of perspectives and attitudes was captured.

Goal Understanding

Regarding the perceived avatar goal, some participants believed the avatar clearly had a particular objective, regardless of whether this perception was entirely accurate, while others felt that the goal was unclear or that no goal existed at all.

To the question "What intentions did the other person seem to have?", accounts ranged from fear and distrust to more neutral or functional interpretations. One participant described it as almost cinematic: "During the first trial

it almost felt like a horror scene, because this person kept saying ‘come here, come here!’” (U1852). Another participant associated the insistence with a potential threat: “Clearly it felt like someone who wanted something... I don’t know, something unsafe [...] or something that would not benefit me, maybe even harm me” (U1852). Spatial orientation also played a key role in how intentions were interpreted. “They kept asking me to reach them, ‘come here,’ and I imagined that once I got to the exact point, the voice would stop speaking.” (U3236). Finally, for some participants, the uncertainty shaped their response: “Not knowing the entity of the voice, the statement ‘come here’ is one that makes you want to move away.” (U4214). From these responses, two broad categories were derived for classification purposes: *the goal is clear* to the participant and *the goal is not clear*.

Environmental Presence

Under the dimension of Perceived Coherence, participants noted whether they found the system fully coherent and immersive (I-C), not coherent and/or immersive at all (No I-C), or partially coherent and immersive (Quasi I-C). For instance, those who considered the system *Quasi I-C* often pointed to certain features that felt well-integrated, in contrast to other elements that seemed inconsistent or disconnected.

When the interviewer asked participants “Do you have any particular thoughts about the other person in the room?”, they once again offered diverse impressions, oscillating between perceptions of threat, artificiality, and monotony. For some, the dominant impression was one of aggression: “The impression was an aggressive manner” (U1852). For others, the non-human quality of the voice was the most salient aspect, noting that “It was a very robotic voice, a bit like Google, or Alexa, so it was clear it wasn’t an actual human being” (U2625). The monotony of repetition also stood out. As one participant explained: “Maybe the fact that they kept saying the same things made it a bit monotonous. [...] Other considerations: surely the sound of the footsteps, that was interesting [...] because it gave you more of a feeling like ‘they’re getting tired because you’re not going where they want you to go’” (U3584). This account illustrates how subtle auditory details beyond speech, such as footsteps and sighs, can enrich the sense of presence and intention.

Participants were also asked “How real did the experience seem to you on a scale from one to ten?” rating the experience, on average, 6.77 ± 1.59 . While breathing was often judged as more convincing than speech or footsteps, inconsistencies in rhythm or directionality undermined immersion. At the same time, the possibility of spatially locating the source in a precise and embodied way (28 out of 30 participants correctly identified the virtual agent as always behind the position) enhanced the perception of reality, showing how subtle acoustic cues can shift the experience from abstract to strikingly tangible.

User-Simulation Connectivity

The perceived directionality of the interaction between the user and the agent ranged from a fully bidirectional relationship (Bi), where both the user and the agent influenced each other, to unidirectional influence (Uni), in which either the avatar or the user held the primary influence (No). In certain cases, participants saw no meaningful interaction between the two. Participants were asked “Did you notice anything particular about your movement and that of the avatar?”.

One explained: “They tended to always stay behind me—they never went in front—they made small movements” (U2757). Another participant described the voice as repeatedly drawing them back to a fixed position: “When I moved away from the reference point, the voice kept asking me to return to the exact spot. During the test, I even moved away on purpose to understand better the exact spot, how the request varied” (U3236). A further account emphasized the impression of mobility of the avatar: “Basically, I tried to adjust my own movement according to the distance, and it seemed to me that they were moving back and forth along the corridor” (U3584). This suggests that participants actively calibrated their movements in response to auditory feedback, producing a dynamic loop between their embodied actions and the perceived positioning of the agent. Overall, these responses highlight the co-constitutive nature of movement and sound: participants adjusted their trajectories to make sense of auditory cues, while interpreting the virtual agent’s presence as variably fixed, guiding, or mobile. Such accounts stress the embodied strategies by which users negotiate presence and orientation in audio-based interaction.

Emerging Patterns

Using gopher distance-based clustering [356] on the three identified themes and the corresponding labelled interviews, three primary clusters emerged. Fig. 6.10 displays their distinguishing features, grouped in *Goal-Oriented*, *Self-Centered*, and *Shared-Discovery* clusters.

Cluster *Goal-Oriented* is composed by 16 subjects. All reported a clear goal of the agent, most interactions were reported to be *Uni* (9 out of 15), while 5 were reported *Bi*, and 2 *No*. The overall coherence was usually considered good: the vast majority was I-C (6) and Quasi I-C (6), while 4 was No I-C. Interestingly, half of the participants reported elements of hostility (8 out of 16), while 3 revealed a lack of understanding. These findings suggest that, even amidst uncertainty, users remain sufficiently involved to perceive some level of coherence—though friction or misalignment can still trigger adverse reactions.

Cluster *Self-Centered* is composed of 9 subjects. In the interviews, all the subjects described the goal as not clear but usually was I-C (7). The remaining 2 participants was No I-C. As regards interaction, only 2 subjects reported *Bi* interaction, 3 subjects considered the system *Uni*, and 4 *No*. Additionally, random elements accounted for half of the observations, pointing to a divergent or unpredictable user response. Marked by generally poor experiences, this cluster lacked entanglement, immersion and coherence, indicating minimal reinforcement to sustain meaningful interaction.

The last cluster, *Shared-Discovery*, was composed of 5 subjects. 4 participants reported *Bi* interaction, one reported *Uni*. All was Quasi I-C and the goal was not goal. This fluidity demonstrates how greater user-agent interactivity can foster more pronounced responses, both positive and negative, and underscores the dynamic nature of entanglement in user experiences.

6.6 Intra-Action Analysis: Systemic View

Both the human and the simulation adapt and respond to one another, creating a dynamic, relational space that could not have been predicted solely based on the

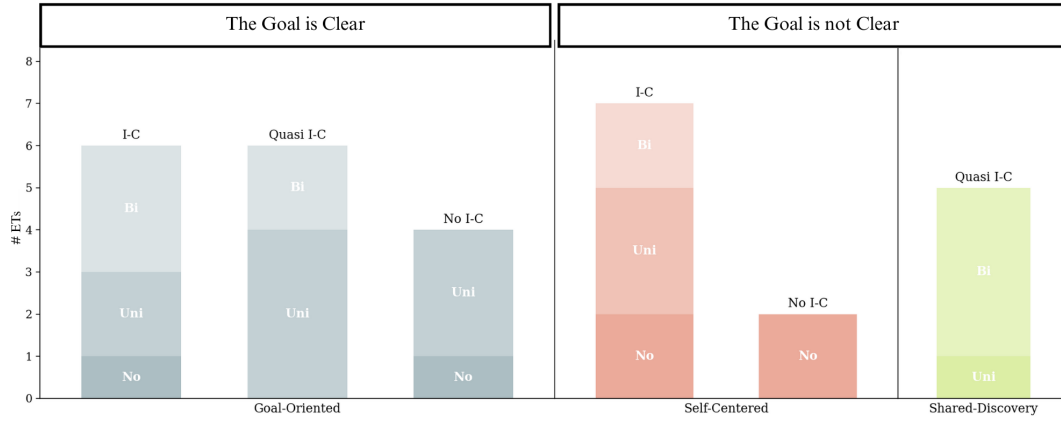


FIGURE 6.10: Human participants' viewpoint: clustering of questionnaire's themes. On top of the stacked bar chart, a visual distinction between the perceived goals is provided. Within each cluster, the bars on environmental presence are stacked based on user-simulation connectivity.

initial hypotheses. Thus, the systemic view, refined by detailed intra-action analysis, reveals the first tangible evidence of entanglement within the experimental environment.

6.6.1 Comparison between Researcher and Computer Viewpoints

The digital agent possesses a different perspective compared to the researchers' team based on its interactions with human participants. As depicted by the alluvial plot in Fig. 6.11-a, a direct correspondence between the experimental conditions and resulting reward clusters does not emerge. Although experimental conditions offer a useful framework set of possible interactions, they do not fully account for the nuances of behaviors observable during the experiential sessions. Moreover, statistical analysis reveals that the experimental conditions only yield trends, not statistically significant outcomes. For instance, while users in the Intimate condition tend to cluster within the Direct behaviour while the Public and Random conditions yield a more diffuse result. This demonstrates that t_{final} and agent-user interaction dynamics are influenced by more than just the designed conditions, and in complex systems, user-specific traits and situational responses play a substantial role.

6.6.2 The Relational Mirror Offered by the Computer Viewpoint

The computer agent's perspective also encompasses multiple perspectives or features, which acts as an experiential mirror of each participant. In particular, the analysis of the most frequent actions permits a more granular analysis, thereby furnishing insights into the complexity of relationality between the human participant and the agent.

In the comparison between the two perspectives, the Direct cluster largely corresponds to the *Single Mode* action pattern, with half of the cases falling into the *Dual Modes* pattern, where participants exhibit one dominant action type above the mean. When a second action is present, it often belongs to the same category, such as two types of forward movements. Importantly, no Direct subject is associated with the *Multi Modes* pattern, indicating that this group is characterized by

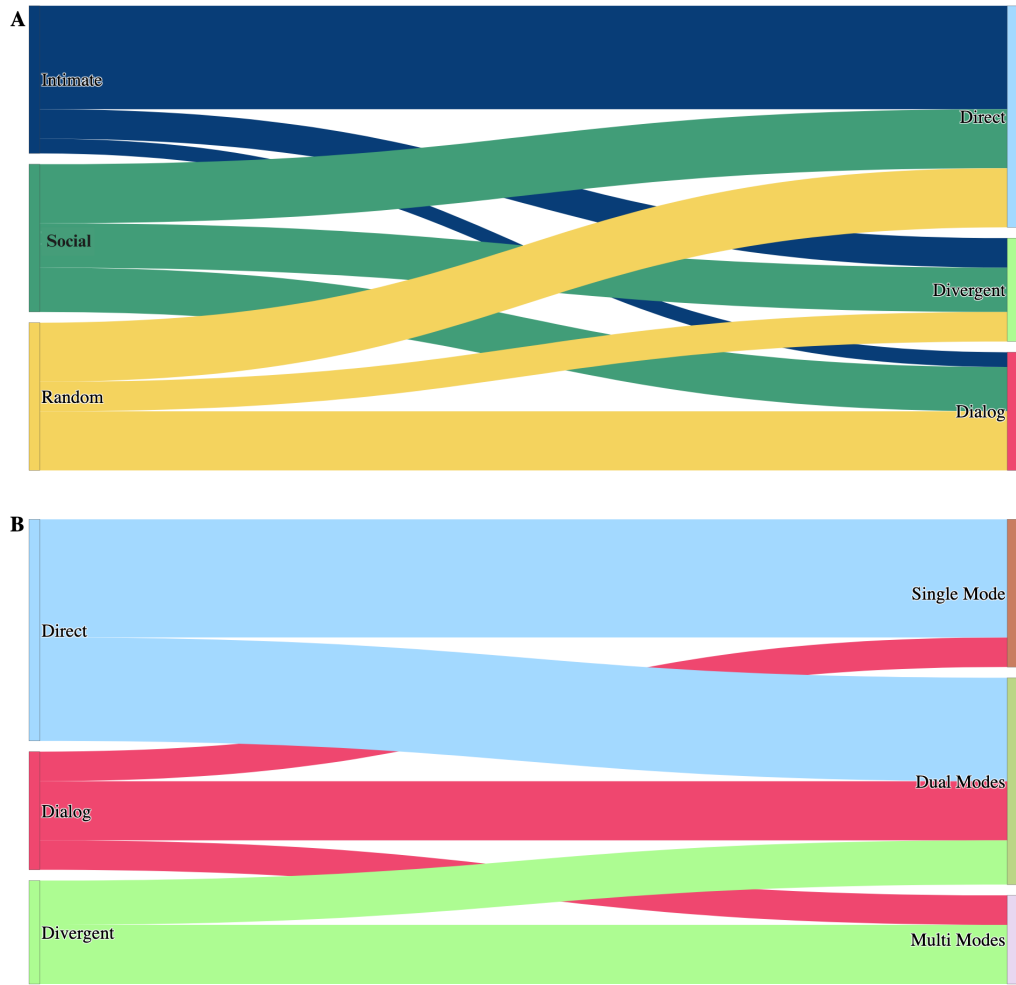


FIGURE 6.11: Systemic view. Alluvial plots a) from experimental conditions to reward clusters. b) from reward cluster to agent's local patterns. Color coding across categories is consistent with earlier figures.

brief, focused interactions that do not require the agent to explore a wider action repertoire to achieve task completion.

In contrast, the Dialog cluster aligns predominantly with *Dual Modes* behavior, and notably, these action pairs typically span different categories (e.g., forward and backward, forward and stop), highlighting a more diverse interaction strategy. While most Dialog participants are confined to this *Dual Modes* structure, some cases fall into *Multi Modes* and *Direct*, suggesting that a broader interactional range may be required for dialogic engagement under certain conditions.

The Divergent type consistently falls into the *Multi Modes* or, less frequently, dual-mode action patterns. All participants in this group have at least three action types above the mean, confirming their complex and variable behavior. This diversity aligns with longer interaction times and the need for the agent to explore multiple action paths in an attempt to engage meaningfully. Notably, no Direct-type participant is found in the *Multi Modes* group, reinforcing the interpretation that more exploratory interactions are required to produce Divergent behavioral profiles.

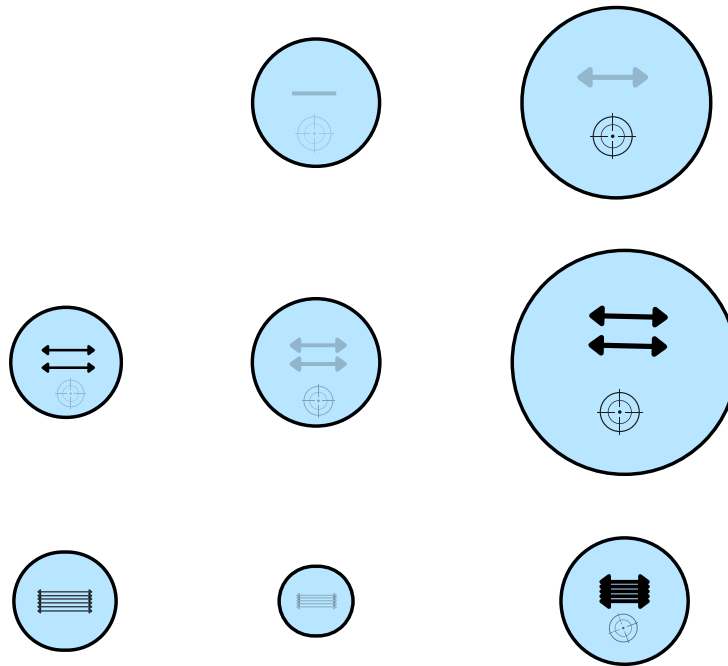


FIGURE 6.12: Typology of entanglement traces between entities **C**omputer and **H**uman participant, represented through circular diagrams. Rows indicate relationality characterization, while columns indicate perceived intentionality. The diameter of each circle is proportional to the numerosity of ETs in the corresponding configuration. Arrows identifies relationality modes within a greyscale of emergence. The target icon indicates the clarity of the perceived goal in a greyscale.

Overall, the partial alignment between global and local features in the reward clusters highlights a meaningful relationship between the agent's two perspectives. The diversity of user experience clearly transcends such categorical boundaries, underscoring the necessity of a broader systemic approach.

6.6.3 Perceived Intentionality

Another crucial comparison emerges when aligning the local-mirroring computer viewpoint, i.e., action characterization, with the local-human viewpoint, derived from the clustering of interview-based thematic analyses on perceived intentionality (as described in Sec. 6.5.3). These qualitative clusters—*Self-Centered*, *Goal-Oriented*, and *Shared-Discovery*—are grounded in participants' subjective perceptions of the interaction and built around emergent patterns of perceived presence, relational intentionality, and entanglement characterization.

When mapping perceived intentionality of the ETs onto the relational action patterns (see Fig. 6.12), an informative and consistent alignment becomes evident. One notable finding is that ET experiences classified as *Single Mode* never

fall into the *Shared-Discovery* cluster. This absence aligns with the core nature of *Shared-Discovery*, which presupposes an exploratory orientation and a diversified interaction repertoire. Participants in this cluster often reported engaging in detailed, exploratory narratives, demonstrating attentive focus on immersive elements such as the spatial dynamics of the AI agent or the supporting sound effects (footsteps and breathing), indicative of a high degree of cognitive engagement and active meaning-making that exceeds what a single dominant interaction mode can support.

In contrast, participants within the *Self-Centered* cluster seldom display *Multi Modes* behavior. Their interactions tend to be fragmented or shallow, and their lack of perceived coherence or clear agent intentionality appears to limit the development of mutual alignment. This correlates with a reduced need for the AI agent to explore alternative interaction strategies, confirming the only behavioral pattern observed, which appeared effective.

The *Goal-Oriented* cluster occupies a hybrid position. Participants in this group exhibit both *Single*, *Dual*, and *Multi Modes* patterns, a reflection of the variability in their interpretation of the agent's intentionality. While most participants do perceive a goal, their understanding of what that goal is tends to diverge significantly. Multiple, sometimes conflicting goals are reported, indicating that although these participants are engaged, the agent's intentions may be ambiguously communicated. This ambiguity is plausibly the result of an interplay between algorithmic constraints in modeling evolving goal structures and user-specific interpretative responses, which jointly contribute to the observed behavioral heterogeneity.

6.7 General Discussion

According to OBJ-1, the M-I-C-E methodology allows us to deepen the analysis of the ETs experience in the experimental settings inspired by Kastanis and Slater [239], who focused on the observable effects of proxemics-driven movement persuasion in controlled conditions. From a traditional hypothetico-deductive perspective, the hypotheses RH1 and RH2 did not yield statistically significant results with respect to the researcher's viewpoint and Kastanis' original research questions on the link between proxemics and t_{final} . However, our analysis shifts the focus toward the traces of entanglement emerging from the We-subject systemic vision (OBJ-2). It is worthwhile to notice that none of the participants explicitly identified the coded goal of the digital agent. This highlights a fundamental phase of the intra-action: rather than selecting a predefined objective, participants entered a process of predicting the agent's intentionality, i.e., an alignment phase to co-constitute a shared intentionality of the We-subject. It is precisely within this exploratory, uncertain space that the first signs of entanglement emerge (OBJ-3). This phase can be interpreted as the relational *handshake*, a point where digital and human agents begin to understand each other, unfolding early-stage engagement patterns.

Entanglement traces, far from being a singular construct, manifested in multiple configurations throughout the proposed experiment. Figure 6.12 displays the eight entanglement trace types identified through the M-I-C-E methodology as a subset of the 3×3 combinations of the relationality patterns, i.e., computer's experiential mirror, and perceived intentionality clusters from interview thematic

analysis. Following the goal-understanding distinction as in Fig. 6.10, two main families of traces emerged when participants believed a goal was present and coherent, and those emerging in conditions of ambiguity or confusion. The majority of the ETs belong to the first family when human participants can report a perceived intentionality:

- Trace 1 (6 ETs) - *Goal-Oriented, Single Mode*. The experience followed a strict path with minimal variation, likely driven by internal belief rather than adaptive response. This strategy may suggest a more selfish form of engagement.
- Trace 2 (7 ETs) - *Goal-Oriented, Dual Modes*. These interactions were focused and ETs show moderate adaptation. This reflects a clear and effective approach to the task, whereby ETs rapidly converged on a perceived intentionality without the need to revisit or revise previous steps.
- Trace 3 (3 ETs) - *Goal-Oriented, Multi Modes*. The experience showed high engagement but seemed to misread or misinterpret the agent's intentionality. The result was intense interaction that lacked coherence, as participants' assumptions shifted over time, prompting them to explore multiple paths before gaining confidence in the goal.

On the other hand, when the goal has been perceived as unclear, diverse exploratory intra-actions unfold:

- Trace 4 (4 ETs) - *Self-Centered, Single Mode*. These happened in short experiential sessions. Participants completed the task quickly with minimal adaptation or sense of connection. The interaction was possibly driven by low perceived relevance of the experience.
- Trace 5 (4 ETs) - *Self-Centered, Dual Modes*. ETs alternated between exploring the digital agent's behavior and relying on human agent's preconceptions, producing disorientation, exposing the difficulty of sustaining coherence against shifting assumptions.
- Trace 6 (3 ETs) - *Shared-Discovery, Dual Modes*. This group also showed oscillation between behaviors, indicating an active search for meaning. ETs were likely engaged but lacked sufficient clarity to stabilize their interactional pattern.
- Trace 7 (2 ETs) - *Shared-Discovery, Multi Modes*. These ETs interacted deeply and across multiple modes, yet failed to converge on a shared goal. Their exploration was rich but unfocused, possibly highlighting a disconnection between the agent intentions.
- Trace 8 (1 ETs) - *Self-Centered, Multi Modes*. this single ET group was open to negotiation but did not exhibit sufficient consistency. With only one example, interpretation remains tentative but points to a space of unresolved mutual engagement.

An open methodological question is how specific experimental a-priori procedure shapes each I-C-E component. The number of possible configurations explodes combinatorially: proxemic boundaries (tighter vs. looser thresholds), agent's sound design (voice timbre, footstep salience, call frequency), goal salience (implicit vs. lightly primed), adaptation rate of the RL algorithm (fast vs. gradual), avatar embodiment (audio-only vs. audio-visual), and experiment duration.

Considering the latter, the short duration of the sessions privileged early-stage mutual negotiations and handshake-like alignment while impairing the development of a more stable We-relation. This temporal constraint likely amplified oscillations (Dual/Multi-Modes) and under-sampled later stabilisation phases, i.e., almost all participants achieved the goal without differentiation based on proxemics and a high number of random choices for the RL algorithm (invalidated RH1 and RH2). Over longer, more complex encounters, traces are expected to evolve through phases (e.g., handshake, mutual calibration, stabilisation, and breakdown/recalibration). Purposeful piloting can be used to open or narrow the spectrum of intra-actions, thereby sharpening trace differentiation. Moreover, capturing this dynamics requires a dynamic network representation, e.g., evolving graph-style representations where tracking immersion, coherence, and entanglement over time [359]. However, developing such tools is out of the scope of the present study and a primary avenue for future work.

It is worthwhile to notice that if the proxemics-driven persuasion hypotheses, HP1, had been confirmed, we would have observed a convergence toward the researcher-coded intentionality in the simulation. Methodologically, this would strengthen OBJ-1 from the researcher-persuasive perspective but may compress the variety of traces and mask early relational negotiation, potentially flattening the emergent We-subject dynamics that are central to the multi-perspective richness required by OBJ-2 and OBJ-3.

6.7.1 Challenges

Since the interviews were based on open questions and were not explicitly designed to probe relationality and activity types, our thematic analysis allowed us to sketch the M-I-C-E framework from participants' lived experiences. A promising and challenging next step would be to complement this with targeted, relation-centred qualitative evaluation tools. For example, the integration of standardized questionnaires linked to the three ICE components could combine multiple presence and agency scales, such as the Igroup Presence Questionnaire (IPQ) [360], the assessment tool for perceived agency [361]. Ultimately, designing and validating a dedicated M-I-C-E questionnaire with an ad hoc set of items would provide a standardized tool to formalize the data space, enabling consistent assessment across studies and fostering cross-modal comparisons.

From the proposed study, it is evident that the efficacy of the M-I-C-E methodology presupposes a precise mirroring of the system dynamics, aiming at understanding human perception. Without AI explainability and transparency, the entanglement traces risk becomes speculative. The proposed digital agent based on the interpretability of $Q(\lambda)$ prioritised policy transparency in logging state – action – reward sequences, offering post-hoc policy visualisations for explainability due to its limited size in state space. However, the challenge of Explainable AI (XAI) is particularly critical in RL [362], where increasing complexity can easily turn models into opaque black boxes. Beyond the current $Q(\lambda)$ setup and the relational nature of I-C-E, it is worth exploring alternative approaches, such as actor-critic methods [363], relational RL frameworks such as Relational Markov Decision Processes (RMDP), and relational Bayes networks [364]. Finally, policy distillation into interpretable models, such as decision-rule surrogates, represents another avenue could enhance transparency [365]. Finally, adopting human-in-the-loop strategies [366], e.g., policy or reward shaping, ensures that the mirroring of the system dynamics remains inspectable, keeping the I-C-E framework

analytically robust.

6.8 Conclusion

This study has shown how the M-I-C-E methodology provides a new framework for investigating human–AI relationality with immersive technologies. By triangulating researcher, simulation, and participant perspectives, the framework captures not only measurable behaviors but also the lived and often divergent experiences that emerge from negotiation with digital agents. The persuasive proxemics-based experience of an audio-voice agent demonstrated that entanglement traces cannot be reduced to top–down experimental conditions induced by the research team: instead, they reveal themselves as *We*-subject relational configurations that exceed design intentions, underscoring the need for multi-perspective approaches in HCI.

From a theoretical standpoint, the observed dissonances across viewpoints illuminate the limits of predefined models of user response and support a shift toward relational ontologies. In our ETs framework, Barad’s agential realism resonates strongly with our findings, highlighting how agency is enacted between human and non-human agents rather than possessed by either. In this sense, the proposed methodology does not simply evaluate interaction quality but offers a way to account for the co-constitution of subjectivity, intentionality, and space in VR/AR environments.

Finally, the M-I-C-E methodology can be extended beyond proxemics and persuasive technologies to domains such as collaborative learning, industrial training, or therapeutic applications, to name but a few, where emergent traces of entanglement are decisive for human well-being. More broadly, it invites design research practices to embrace systemic, cross-disciplinary analyses capable of integrating behavioral data, phenomenological concepts, and computational models. M-I-C-E contributes to shaping a responsible and transparent vision of future human–AI ecologies where technology participates in the co-creation of shared experiential worlds.

Chapter 7

Conclusions

7.1 Main Thesis Achievements

This research has contributed to the evolving fields of Sonic Interaction in Virtual Environments (SIVE) and HCI by redefining the role of audio as a primary agential force within intelligent environments. By shifting the perspective from traditional sound reproduction to a framework of agential realism [38], this work has established how egocentric audio technologies within a MR context can serve as mediators in the mutual interaction between human (users) and non-human (technology) entities.

Chapter 2 addressed a research gap in understanding the bidirectional, agential relationship between human users and AI-driven interfaces. This led to the definition of the M-I-C-E framework (Multiperspective, Immersion, Coherence, Entanglement), providing a foundational framework for analysing human-technology interaction. Building upon this, the M-I-C-E methodology was introduced as a significant innovation to integrate the perspectives of designers, users, and virtual agents. This conceptual development evolved the theory of the Auditory Digital Twin into the "Entangled Twin", shifting the discourse from viewing interfaces as passive tools to recognising them as active participants in a collaborative "we-relationship".

Chapter 3 introduced a possible application domain for Audio AR: Cultural Heritage Storytelling. A systematic review discussed the role of audio in the field, highlighting a crucial shift toward personalization. This was practically demonstrated through "Narrami," a hypothetical case study on adaptive storytelling. On the technical front, **Chapter 4** established a robust auralization framework incorporating environmental room acoustics and user-specific HRTF personalization, as well as headphones equalisation and user tracking, reconstructing a real room (a corridor) in the virtual realm. Rigorous evaluations of consumer hardware, such as the iPad, ensured spatial audio fidelity and tracking accuracy for the ecological deployment.

Methodological advancements continued in **Chapter 5** with a deep discussion about Immersion and Conherence.

Section 5.2 is dedicated to the role of tracking introducing an evaluation of tracking using consumer-grade devices. The following sections discussed D&DMUSHRA and its expanded version, D&DMUSHRA+, a screening protocols that allow for the simultaneous testing of multiple rendering factors like HRTF and equalisation. Experimental validation confirmed that personalisation is essential for maintaining *plausibility* and *co-immersion*. Moreover, to visualize these complex dynamics, the EXMap was introduced to track interactions between human users and technological systems over time. The chapter also explored co-immersion. By applying Signal Detection Theory (SDT), we analyse how effectively users can

discriminate between sounds coming from real and virtual speakers using different rendering techniques.

Chapter 6, introduced the Entanglement eXperience Map (EXMap), a design tool aimed at representing and tracking SIVE dimensions in an actor-network map. The second part of the chapter detailed an ecological study using a Reinforcement Learning virtual agent, which led to the identification of eight distinct *entanglement traces*. These traces provided a novel taxonomy for understanding how intentionality and meaning are co-constituted between human users and intelligent audio agents.

7.2 Future Directions

The outcomes of this thesis open promising avenues for future research in SIVE. One primary direction involves the improvement and expansion of the Entanglement Traces taxonomy. While the current research primarily focused on the initial *handshake* phase of interaction, future studies should consider implementing longitudinal experiments in ecological environments. This would enable the monitoring of entanglement evolving over extended periods, identifying traces that emerge only through sustained interaction and defining formal measurement metrics to quantify the depth and quality of the human-agent interaction.

Building on this, future research should focus on applying the M-I-C-E framework to other multimodal contexts. Although this thesis prioritised the auditory dimension, the methodology is designed to be extensible to MR environments where the visual component is integrated. Investigating the coherence between spatial audio and visual stimuli through the lens of entanglement will be crucial for developing truly immersive experiences. Furthermore, the framework's application can be extended to the study of interaction with general UIs and social networking platforms. In these contexts, entanglement can manifest through algorithmic feedback dynamics and personalised information flows, where the system is an agent that actively shapes social content and user perception. This necessitates an evaluation of coherence and immersion even in the absence of a three-dimensional spatial component.

Furthermore, a significant opportunity can come from applying M-I-C-E framework to different fields. While Cultural Heritage represented the first elected context, the framework and technologies developed here can be applied to industrial, educational, and medical fields, as well as to socio-collaborative scenarios involving human-human-technology interaction. Exploring how egocentric audio facilitates collaboration in shared environments where multiple humans and technologies interact simultaneously could be an interesting area of development.

The technical integration of advanced algorithms and Large Language Models with the Auditory Digital Twin presents another vital opportunity. Future work should explore how generative AI can synthesize spatialized dialogue and environmental soundscapes that respond dynamically to a user's verbal input and emotional state. Implementing more sophisticated RL models will allow for a more natural and fluid interaction, deepening the sense of co-presence.

In addition to the broader research perspectives outlined above, this work opens several concrete and short-term research directions directly grounded in the findings of this thesis.

A first direction concerns the systematic investigation of perceptual thresholds in Audio MR. The results of Chapter 5 suggest that users tolerate a certain

degree of approximation in tracking and audio rendering. Future studies should aim to quantify these thresholds more precisely, identifying the minimum technical requirements needed to achieve perceptual plausibility under different contextual conditions.

A second direction relates to the role of sound design in shaping interaction and entanglement. As highlighted in Chapter 6, auditory cues such as spatialization, and temporal patterns may significantly influence users' interpretation of agent behavior. Controlled experiments focusing on these parameters would enable a more fine-grained understanding of audio-driven interaction.

A third direction involves the development of adaptive rendering strategies. Building on the findings of Chapter 5, future work could explore dynamic allocation of computational resources based on perceptual relevance, user context, or interaction complexity, enabling scalable Audio MR systems.

Finally, the EXMap framework introduced in Chapter 6 could be further refined and validated across different application domains, including collaborative XR scenarios and cultural heritage experiences, to assess its generalizability and analytical robustness.

To conclude, an ambitious future direction should be a better understanding of the impact that human-technology interaction has on individuals. Future research in the field should aim to culminate in the creation of comprehensive guidelines for the safe and ethical design of entangled systems. Ensuring that these pervasive and intelligent environments respect user agency and well-being is essential as they become increasingly integrated into daily life.

Bibliography

- [1] Mel Slater and Maria V. Sanchez-Vives. “Enhancing Our Lives with Immersive Virtual Reality”. In: *Frontiers in Robotics and AI* 3 (Dec. 2016). ISSN: 2296–9144. DOI: [10.3389/frobt.2016.00074](https://doi.org/10.3389/frobt.2016.00074). URL: <https://doi.org/10.3389/frobt.2016.00074>.
- [2] Frank Biocca. “The Cyborg’s Dilemma: Progressive Embodiment in Virtual Environments [1]”. In: *Journal of Computer-Mediated Communication* 3.2 (June 2006), 0–0. ISSN: 1083-6101. DOI: [10.1111/j.1083-6101.1997.tb00070.x](https://doi.org/10.1111/j.1083-6101.1997.tb00070.x). URL: <http://dx.doi.org/10.1111/j.1083-6101.1997.tb00070.x>.
- [3] Jeremy Bailenson. *Experience on demand*. en. New York, NY: WW Norton, Mar. 2019.
- [4] Richard Skarbez et al. “Immersive Analytics: Theory and Research Agenda”. In: *Frontiers in Robotics and AI* 6 (Sept. 2019). ISSN: 2296-9144. DOI: [10.3389/frobt.2019.00082](https://doi.org/10.3389/frobt.2019.00082). URL: <http://dx.doi.org/10.3389/frobt.2019.00082>.
- [5] Christopher Frauenberger. “Entanglement HCI The Next Wave?” In: *ACM Transactions on Computer-Human Interaction* 27.1 (Nov. 2019), pp. 1–27. DOI: [10.1145/3364998](https://doi.org/10.1145/3364998). URL: <https://doi.org/10.1145/3364998>.
- [6] Michele Geronazzo and Stefania Serafin. “Sonic Interactions in Virtual Environments: The Egocentric Audio Perspective of the Digital Twin”. In: *Sonic Interactions in Virtual Environments*. Springer International Publishing, Oct. 2022, pp. 3–45. DOI: [10.1007/978-3-031-04021-4_1](https://doi.org/10.1007/978-3-031-04021-4_1). URL: https://doi.org/10.1007/978-3-031-04021-4_1.
- [7] Nick Yee and Jeremy Bailenson. “The Proteus Effect: The Effect of Transformed Self-Representation on Behavior”. In: *Human Communication Research* 33.3 (July 2007), 271–290. ISSN: 1468-2958. DOI: [10.1111/j.1468-2958.2007.00299.x](https://doi.org/10.1111/j.1468-2958.2007.00299.x). URL: <http://dx.doi.org/10.1111/j.1468-2958.2007.00299.x>.
- [8] James Broderick, Jim Duggan, and Sam Redfern. “The importance of spatial audio in modern games and virtual environments”. In: *2018 IEEE Games, Entertainment, Media Conference (GEM)*. Galway: IEEE, Aug. 2018.
- [9] Stefano Delle Monache, Nicolas Misdariis, and Elif Özcan. “Semantic models of sound-driven design: Designing with listening in mind”. en. In: *Des. Stud.* 83.101134 (Nov. 2022), p. 101134.
- [10] Alessandro Giuseppe Privitera and Michele Geronazzo. “Designing sonic interactions in intelligent reality with egocentric audio technologies”. In: *The Routledge Handbook of Sound Design*. Focal Press, 2024, pp. 219–251.
- [11] Alessandro Giuseppe Privitera and Michele Geronazzo. “Traces of Entanglement with an AI Agent in Augmented Reality: The M-I-C-E Methodology”. Manuscript in preparation. 2025.

- [12] Floriana Ferro, Alessandro Giuseppe Privitera, and Michele Geronazzo. "From Digital Twins to Entangled Twins: A Phenomenological Model for Industry 5.0". In: *Submitted for publication at Zagadnienia Filozoficzne w Nauce (Philosophical Problems in Science)* (2025).
- [13] Alessandro Giuseppe Privitera, Federico Fontana, and Michele Geronazzo. "The Role of Audio in Immersive Storytelling: a Systematic Review in Cultural Heritage". In: *Multimedia Tools and Applications* 84.16 (June 2024), 16105–16143. ISSN: 1573-7721. DOI: [10.1007/s11042-024-19288-4](https://doi.org/10.1007/s11042-024-19288-4). URL: <http://dx.doi.org/10.1007/s11042-024-19288-4>.
- [14] Alessandro Giuseppe Privitera, Marco Noro, Michele Geronazzo, et al. "Preliminary evaluation of the auralization of a real indoor environment for augmented reality research". In: *Conv. Eur. Acoust. Assoc* (2023).
- [15] Alessandro Giuseppe Privitera, Federico Fontana, and Michele Geronazzo. "On the effect of user tracking on perceived source positions in mobile audio augmented reality". In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*. 2023, pp. 1–9.
- [16] Davide Fantini et al. "Co-immersion in audio augmented virtuality: the case study of a static and approximated late reverberation algorithm". In: *IEEE transactions on visualization and computer graphics* 29.11 (2023), pp. 4472–4482.
- [17] Alessandro Giuseppe Privitera, Marco Noro, and Michele Geronazzo. "Piloting and Screening Multiple Factors in Audio Augmented Reality with D&DMUSHRA+: a Case Study on Plausibility". Submitted for publication at *Frontiers in Virtual Reality - Augmented Reality*. 2025.
- [18] Alessandro Giuseppe Privitera and M. Geronazzo. *Designing Sonic Interactions in Intelligent Reality with Egocentric Audio Technologies*. Routledge Handbook on Sound Design (to appear). Routledge, Taylor & Francis Group, 2024.
- [19] Michele Geronazzo and Stefania Serafin. "Sonic Interactions in Virtual Environments: the Egocentric Audio Perspective of the Digital Twin". In: *Sonic Interactions in Virtual Environments*. Human–Computer Interaction Series. Springer London, 2023, pp. 3–48. ISBN: 978-3-031-04020-7.
- [20] Paul Milgram and Fumio Kishino. "A taxonomy of mixed reality visual displays". In: *IEICE TRANSACTIONS on Information and Systems* 77.12 (1994), pp. 1321–1329.
- [21] Vesa Valimaki et al. "Assisted Listening Using a Headset: Enhancing audio perception in real, augmented, and virtual environments". In: *IEEE Signal Processing Magazine* 32.2 (Mar. 2015), pp. 92–99. DOI: [10.1109/msp.2014.2369191](https://doi.org/10.1109/msp.2014.2369191). URL: <https://doi.org/10.1109/msp.2014.2369191>.
- [22] Roman Schlieper, Stephan Preihs, and Jürgen Peissig. "An open dataset of measured HRTFs perturbed by headphones". In: *Audio Engineering Society Convention* 152. May 2022. URL: <http://www.aes.org/e-lib/browse.cfm?elib=21724>.
- [23] Fabian Brinkmann, Alexander Lindau, and Stefan Weinzierl. "On the authenticity of individual dynamic binaural synthesis". In: *The Journal of the Acoustical Society of America* 142.4 (Oct. 2017), pp. 1784–1795. DOI: [10.1121/1.5005606](https://doi.org/10.1121/1.5005606). URL: <https://doi.org/10.1121/1.5005606>.

- [24] Xiao li Zhong and Bo sun Xie. "Head-Related Transfer Functions and Virtual Auditory Display". In: *Soundscape Semiotics - Localisation and Categorisation*. InTech, Mar. 2014. DOI: [10.5772/56907](https://doi.org/10.5772/56907). URL: <https://doi.org/10.5772/56907>.
- [25] Stefania Serafin et al. "Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions". In: *IEEE Computer Graphics and Applications* 38.2 (Mar. 2018), pp. 31–43. DOI: [10.1109/mcg.2018.193142628](https://doi.org/10.1109/mcg.2018.193142628). URL: <https://doi.org/10.1109/mcg.2018.193142628>.
- [26] Stefania Serafin et al. "Reflections from five years of Sonic Interactions in Virtual Environments workshops". In: *Journal of New Music Research* 49.1 (Jan. 2020), pp. 24–34. ISSN: 0929-8215. DOI: [10.1080/09298215.2019.1708413](https://doi.org/10.1080/09298215.2019.1708413). URL: <https://doi.org/10.1080/09298215.2019.1708413> (visited on 01/31/2020).
- [27] Louis Nisiotis and Lyuba Alboul. "Evaluating Presence and Technology Acceptance of an Intelligent Reality Virtual Museum Prototype". In: *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*. IEEE, Dec. 2022. DOI: [10.1109/icir55739.2022.00016](https://doi.org/10.1109/icir55739.2022.00016). URL: <https://doi.org/10.1109/icir55739.2022.00016>.
- [28] Andriani Piki, Louis Nisiotis, and Lyuba Alboul. "A Preliminary Exploration of the Learning and Engagement Potential of an Intelligent Virtual Environment". In: *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*. IEEE, Dec. 2022. DOI: [10.1109/icir55739.2022.00021](https://doi.org/10.1109/icir55739.2022.00021). URL: <https://doi.org/10.1109/icir55739.2022.00021>.
- [29] Takeo Kanade and Martial Hebert. "First-Person Vision". In: *Proceedings of the IEEE* 100.8 (Aug. 2012). Conference Name: Proceedings of the IEEE, pp. 2442–2453. ISSN: 1558-2256. DOI: [10.1109/JPROC.2012.2200554](https://doi.org/10.1109/JPROC.2012.2200554).
- [30] Mel Slater. "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1535 (Dec. 2009), pp. 3549–3557. ISSN: 0962-8436. DOI: [10.1098/rstb.2009.0138](https://doi.org/10.1098/rstb.2009.0138). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2781884/> (visited on 11/16/2021).
- [31] Virginia Best et al. "Sound Externalization: A Review of Recent Research". In: *Trends in Hearing* 24 (Jan. 2020). Publisher: SAGE Publications Inc, p. 233121652094839. ISSN: 2331-2165. DOI: [10.1177/2331216520948390](https://doi.org/10.1177/2331216520948390). URL: <https://doi.org/10.1177/2331216520948390> (visited on 09/17/2021).
- [32] Ulrike Schultze. "The Avatar as Sociomaterial Entanglement: A Performative Perspective on Identity, Agency and World-Making in Virtual Worlds". In: *Proceedings of the International Conference on Information Systems, ICIS 2011, Shanghai, China, December 4-7, 2011*. Ed. by Dennis F. Galletta and Ting-Peng Liang. Association for Information Systems, 2011. URL: <http://aisel.aisnet.org/icis2011/proceedings/hci/11>.
- [33] Stein Tore Johansen et al. "Hybrid and cognitive digital twins for the process industry". In: *Open Engineering* 13.1 (Jan. 2023). DOI: [10.1515/eng-2022-0418](https://doi.org/10.1515/eng-2022-0418). URL: <https://doi.org/10.1515/eng-2022-0418>.
- [34] Behrang Ashtari Talkhestani et al. "An architecture of an Intelligent Digital Twin in a Cyber-Physical Production System". In: *at - Automatisierungstechnik* 67.9 (Sept. 2019), pp. 762–782. DOI: [10.1515/auto-2019-0039](https://doi.org/10.1515/auto-2019-0039). URL: <https://doi.org/10.1515/auto-2019-0039>.

- [35] Joakim Vindenes and Barbara Wasson. "A Postphenomenological Framework for Studying User Experience of Immersive Virtual Reality". In: *Frontiers in Virtual Reality* 2 (Apr. 2021). DOI: [10.3389/frvir.2021.656423](https://doi.org/10.3389/frvir.2021.656423). URL: <https://doi.org/10.3389/frvir.2021.656423>.
- [36] Floris den Hengst et al. "Reinforcement learning for personalization: A systematic literature review". In: *Data Science* 3.2 (Nov. 2020). Ed. by Izabela Moise, pp. 107–147. DOI: [10.3233/ds-200028](https://doi.org/10.3233/ds-200028). URL: <https://doi.org/10.3233/ds-200028>.
- [37] Elena Pretel et al. "Multi-Agent Systems in Support of Digital Twins: A Survey". In: *Bio-inspired Systems and Applications: from Robotics to Ambient Intelligence*. Springer International Publishing, 2022, pp. 524–533. DOI: [10.1007/978-3-031-06527-9_52](https://doi.org/10.1007/978-3-031-06527-9_52). URL: https://doi.org/10.1007/978-3-031-06527-9_52.
- [38] Karen Barad. *Meeting the Universe Halfway*. Duke University Press, Jan. 2006. DOI: [10.1515/9780822388128](https://doi.org/10.1515/9780822388128). URL: <https://doi.org/10.1515/9780822388128>.
- [39] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. New York: Psychology Press, Dec. 2014. ISBN: 978-1-315-74021-8. DOI: [10.4324/9781315740218](http://dx.doi.org/10.4324/9781315740218). URL: <http://dx.doi.org/10.4324/9781315740218>.
- [40] M. Geronazzo. "Egocentric Audio in the Digital Twin of Virtual Environments". In: *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*. IEEE, Dec. 2022, pp. 7–10. DOI: [10.1109/ICIR55739.2022.00017](https://doi.org/10.1109/ICIR55739.2022.00017). URL: <https://doi.org/10.1109/ICIR55739.2022.00017>.
- [41] Edward Glaessgen and David Stargel. "The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles". In: *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA*. Apr. 2012. DOI: [10.2514/6.2012-1818](http://dx.doi.org/10.2514/6.2012-1818). URL: <http://dx.doi.org/10.2514/6.2012-1818>.
- [42] Shaun Gallagher and Dan Zahavi. *The phenomenological mind*. en. 3rd ed. London, England: Routledge, Oct. 2020.
- [43] Edmund Husserl. *Cartesian Meditations: An Introduction to Phenomenology*. Trans. by D. Cairns. M. Nijhoff. en. The Hague, 1960.
- [44] Maurice Merleau-Ponty. *Phenomenology of Perception*. Trans. by C. Smith. London: Routledge, 1958.
- [45] Jean-Paul Sartre. *Being and Nothingness: An Essay in Phenomenological Ontology*. Trans. by H.E. Barns. Washington: Washington Square Press, 1992. ISBN: 9780429434013. DOI: [10.4324/9780429434013](http://dx.doi.org/10.4324/9780429434013). URL: <http://dx.doi.org/10.4324/9780429434013>.
- [46] Karen Barad. *Meeting the universe halfway*. Durham, NC: Duke University Press, Mar. 2007.
- [47] Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, July 2005. ISBN: 9781383039658. DOI: [10.1093/oso/9780199256044.001.0001](http://dx.doi.org/10.1093/oso/9780199256044.001.0001). URL: <http://dx.doi.org/10.1093/oso/9780199256044.001.0001>.

- [48] Luciano Floridi. "AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis". In: *Philosophy and Technology* 38.1 (Feb. 2025). ISSN: 2210-5441. DOI: [10.1007/s13347-025-00858-9](https://doi.org/10.1007/s13347-025-00858-9). URL: <http://dx.doi.org/10.1007/s13347-025-00858-9>.
- [49] Mel Slater, Andrea Brogni, and Anthony Steed. "Physiological Responses to Breaks in Presence: A Pilot Study". In: *Presence 2003: The 6th annual international workshop on presence*. Vol. 157. 2003. URL: <https://api.semanticscholar.org/CorpusID:17414751>.
- [50] Michele Geronazzo et al. "The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios". In: *2018 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE. 2018, pp. 90–97.
- [51] Richard Skarbez, Frederick P. Brooks, and Mary C. Whitton. "Immersion and Coherence: Research Agenda and Early Results". In: *IEEE Transactions on Visualization and Computer Graphics* 27.10 (Oct. 2021), 3839–3850. ISSN: 2160-9306. DOI: [10.1109/tvcg.2020.2983701](https://doi.org/10.1109/tvcg.2020.2983701). URL: <http://dx.doi.org/10.1109/TVCG.2020.2983701>.
- [52] Alessandro Giuseppe Privitera, Federico Fontana, and Michele Geronazzo. "On the Effect of User Tracking on Perceived Source Positions in Mobile Audio Augmented Reality". In: *CHIItaly '23*. Torino, Italy: Association for Computing Machinery, 2023.
- [53] Bruno Latour et al. "'The whole is always smaller than its parts' - a digital test of Gabriel Tarde's monads". In: *The British Journal of Sociology* 63.4 (Dec. 2012), pp. 590–615. DOI: [10.1111/j.1468-4446.2012.01428.x](https://doi.org/10.1111/j.1468-4446.2012.01428.x). URL: <https://doi.org/10.1111/j.1468-4446.2012.01428.x>.
- [54] Jenny Preece, Helen Sharp, and Yvonne Rogers. *Interaction Design*. en. 4th ed. Nashville, TN: John Wiley & Sons, Feb. 2015.
- [55] Kristina Höök et al. "Unpacking Non-Dualistic Design: The Soma Design Case". In: *ACM Trans. Comput.-Hum. Interact.* 28.6 (Nov. 2021), 40:1–40:36. ISSN: 1073-0516. DOI: [10.1145/3462448](https://doi.org/10.1145/3462448). URL: <https://dl.acm.org/doi/10.1145/3462448> (visited on 09/11/2025).
- [56] Elisa Giaccardi. "Histories and Futures of Research through Design:" en. In: 13.3 (2019).
- [57] John W. Creswell and Vicki L. Plano Clark. *Designing and conducting mixed methods research*. SAGE Publications, 2011.
- [58] Clara Bassano et al. "Storytelling about places: Tourism marketing in the digital age". In: *Cities* 87 (Apr. 2019), 10–20. ISSN: 0264–2751. DOI: [10.1016/j.cities.2018.12.025](https://doi.org/10.1016/j.cities.2018.12.025). URL: <http://dx.doi.org/10.1016/j.cities.2018.12.025>.
- [59] Hubert Cecotti. "Cultural Heritage in Fully Immersive Virtual Reality". In: *Virtual Worlds* 1.1 (Sept. 2022), 82–102. ISSN: 2813-2084. DOI: [10.3390/virtualworlds1010006](https://doi.org/10.3390/virtualworlds1010006).
- [60] Barbara Balbi and Alessandra Marasco. "Co-designing Cultural Heritage Experiences for All with Virtual Reality: a Scenario-Based Design approach". en. In: *Umanistica Digitale* No. 11 (2021 2022), rights and values in the Digital Age. DOI: [10.6092/ISSN.25328-8816/13686](https://doi.org/10.6092/ISSN.25328-8816/13686).

- [61] Maud Ceuterick and Chris Ingraham. "Immersive storytelling and affective ethnography in virtual reality". In: *Review of Communication* 21.1 (Jan. 2021), 9–22. ISSN: 1535–8593. DOI: [10.1080/15358593.2021.1881610](https://doi.org/10.1080/15358593.2021.1881610).
- [62] Seung Jae Lee. "A review of audio guides in the era of smart tourism". In: *Information Systems Frontiers* 19.4 (July 2016), pp. 705–715. DOI: [10.1007/s10796-016-9666-6](https://doi.org/10.1007/s10796-016-9666-6).
- [63] Mafkereseb Kassahun Bekele et al. "A Survey of Augmented, Virtual, and Mixed Reality for Cultural Heritage". In: *Journal on Computing and Cultural Heritage* 11.2 (Mar. 2018), pp. 1–36. DOI: [10.1145/3145534](https://doi.org/10.1145/3145534). URL: <https://doi.org/10.1145/3145534>.
- [64] Tony Regia-Corte et al. "Perceiving affordances in virtual reality: influence of person and environmental properties in perception of standing on virtual grounds". In: *Virtual Reality* 17.1 (Oct. 2012), 17–28. ISSN: 1434–9957. DOI: [10.1007/s10055-012-0216-3](https://doi.org/10.1007/s10055-012-0216-3).
- [65] Alessandro Giuseppe Privitera, Federico Fontana, and Michele Geronazzo. "On the Effect of User Tracking on Perceived Source Positions in Mobile Audio Augmented Reality". In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*. CHIItaly '23. , Torino, Italy: Association for Computing Machinery, 2023. ISBN: 9798400708060. DOI: [10.1145/3605390.3605422](https://doi.org/10.1145/3605390.3605422). URL: <https://doi-org.altais.uniud.it/10.1145/3605390.3605422>.
- [66] James Broderick, Jim Duggan, and Sam Redfern. "The Importance of Spatial Audio in Modern Games and Virtual Environments". In: *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, Aug. 2018. DOI: [10.1109/gem.2018.8516445](https://doi.org/10.1109/gem.2018.8516445).
- [67] Stefano Delle Monache, Nicolas Misdariis, and Elif Özcan. "Semantic models of sound-driven design: Designing with listening in mind". In: *Design Studies* 83 (Nov. 2022), p. 101134. ISSN: 0142-694X. DOI: [10.1016/j.destud.2022.101134](https://doi.org/10.1016/j.destud.2022.101134). URL: <https://doi.org/10.1016/j.destud.2022.101134>.
- [68] Charalampos A. Dimoulas. "Cultural Heritage Storytelling, Engagement and Management in the Era of Big Data and the Semantic Web". In: *Sustainability* 14.2 (Jan. 2022), p. 812. DOI: [10.3390/su14020812](https://doi.org/10.3390/su14020812). URL: <https://doi.org/10.3390/su14020812>.
- [69] Silviu Vert et al. "User Evaluation of a Multi-Platform Digital Storytelling Concept for Cultural Heritage". In: *Mathematics* 9.21 (Oct. 2021), p. 2678. DOI: [10.3390/math9212678](https://doi.org/10.3390/math9212678). URL: <https://doi.org/10.3390/math9212678>.
- [70] Jing Yang, Amit Barde, and Mark Billinghurst. "Audio Augmented Reality: A Systematic Review of Technologies, Applications, and Future Research Directions". In: *Journal of the Audio Engineering Society* 70.10 (Nov. 2022), pp. 788–809. DOI: [10.17743/jaes.2022.0048](https://doi.org/10.17743/jaes.2022.0048). URL: <https://doi.org/10.17743/jaes.2022.0048>.
- [71] Isak de Villiers Bosman et al. "The effect of audio on the experience in virtual reality: a scoping review". In: *Behaviour & Information Technology* (Jan. 2023), pp. 1–35. DOI: [10.1080/0144929x.2022.2158371](https://doi.org/10.1080/0144929x.2022.2158371).
- [72] Gene Youngblood. *Expanded cinema*. en. Meaning Systems. Fordham University Press, Mar. 2020.

- [73] Benjamin B. Bederson. "Audio augmented reality: a prototype automated tour guide". In: *Conference companion on Human factors in computing systems - CHI '95*. CHI '95. ACM Press, 1995. DOI: [10.1145/223355.223526](https://doi.org/10.1145/223355.223526). URL: <http://dx.doi.org/10.1145/223355.223526>.
- [74] P. Marti et al. "Adapting the museum: a non-intrusive user modeling approach". In: *UM99 User Modeling*. Springer Vienna, 1999, 311–313. ISBN: 9783709124901. DOI: [10.1007/9783-7091-2490-1_34](https://doi.org/10.1007/9783-7091-2490-1_34). URL: http://dx.doi.org/10.1007/9783-7091-2490-1_34.
- [75] D.-M. Popovici et al. "Interactive distributed guided tours of historical sites". In: *Proceedings. 2003 International Conference on Cyberworlds*. CW-03. IEEE Comput. Soc, 2003. DOI: [10.1109/cyber.2003.1253489](https://doi.org/10.1109/cyber.2003.1253489). URL: <http://dx.doi.org/10.1109/CYBER.2003.1253489>.
- [76] Anna Poletti. "Coaxing an intimate public: Life narrative in digital storytelling". In: *Continuum* 25.1 (Feb. 2011), pp. 73–83. DOI: [10.1080/10304312.2010.506672](https://doi.org/10.1080/10304312.2010.506672).
- [77] M. Meadows. *Pause & Effect: The Art of Interactive Narrative*. Pearson Education, 2002. ISBN: 9780132798259. URL: https://books.google.it/books?id=W1WP_mUJbYQC.
- [78] Marshall McLuhan. *Understanding media: the extensions of man*. en. The MIT Press. London, England: MIT Press, Oct. 1964.
- [79] Kathleen C. Sitter, Natalie Beausoleil, and Erin McGowan. "Digital Storytelling and Validity Criteria". In: *International Journal of Qualitative Methods* 19 (Jan. 2020), p. 160940692091065. ISSN: 1609–4069. DOI: [10.1177/1609406920910656](https://doi.org/10.1177/1609406920910656).
- [80] Nicola Grove, ed. *Using Storytelling to Support Children and Adults with Special Needs*. en. London, England: Routledge, Nov. 2012.
- [81] Davide Rocchesso et al. "Sonic interaction design: sound, information and experience". In: *CHI '08 Extended Abstracts on Human Factors in Computing Systems*. CHI '08. ACM, Apr. 2008. DOI: [10.1145/1358628.1358969](https://doi.org/10.1145/1358628.1358969). URL: <http://dx.doi.org/10.1145/1358628.1358969>.
- [82] John R. Pierce. "The Nature of Musical Sound". In: *The Psychology of Music (Second Edition)*. Ed. by Diana Deutsch. Second Edition. Cognition and Perception. San Diego: Academic Press, 1999, pp. 1–23. ISBN: 978-0-12-213564-4. DOI: https://doi.org/10.1016/B978-0122135644_4/500025-0. URL: <https://www.sciencedirect.com/science/article/pii/B9780122135644500020>.
- [83] Artur Lugmayr et al. "Serious storytelling – a first definition and review". In: *Multimedia Tools and Applications* 76.14 (Sept. 2016), pp. 15707–15733. DOI: [10.1007/s11042-016-3865-5](https://doi.org/10.1007/s11042-016-3865-5).
- [84] Veera Ojala. "Chernobyl dreams: investigating visitors' storytelling in the Chernobyl exclusion zone". In: *International Journal of Tourism Cities* (Dec. 2022). DOI: [10.1108/ijtc-04-2022-0094](https://doi.org/10.1108/ijtc-04-2022-0094).
- [85] Evgenios Vlachos, Jakob Povl Holck, and Mogens Kraghsig Jensen. "Narrating the Story of a Digitized Old Historical Map". In: *HCI International 2022 Posters*. Springer International Publishing, 2022, 296–303. ISBN: 9783031063916. DOI: [10.1007/9783-031-06391-6_39](https://doi.org/10.1007/9783-031-06391-6_39). URL: http://dx.doi.org/10.1007/9783-031-06391-6_39.

- [86] Anna Podara et al. "Digital Storytelling in Cultural Heritage: Audience Engagement in the Interactive Documentary New Life". In: *Sustainability* 13.3 (Jan. 2021), p. 1193. DOI: [10.3390/su13031193](https://doi.org/10.3390/su13031193).
- [87] Charles Forceville. "Interactive documentary and its limited opportunities to persuade". In: *Discourse, Context & Media* 20 (Dec. 2017), pp. 218–226. ISSN: 2211–6958. DOI: [10.1016/j.dcm.2017.06.004](https://doi.org/10.1016/j.dcm.2017.06.004). URL: <http://dx.doi.org/10.1016/j.dcm.2017.06.004>.
- [88] Joao M. F. Rodrigues et al. "Mobile Five Senses Augmented Reality System: Technology Acceptance Study". In: *IEEE Access* 7 (2019), pp. 163022–163033. DOI: [10.1109/access.2019.2953003](https://doi.org/10.1109/access.2019.2953003). URL: <https://doi.org/10.1109/access.2019.2953003>.
- [89] Marco Comunità et al. "Design and Evaluation of a Web- and Mobile-Based Binaural Audio Platform for Cultural Heritage". In: *Applied Sciences* 11.4 (Feb. 2021), p. 1540. DOI: [10.3390/app11041540](https://doi.org/10.3390/app11041540).
- [90] Fatima-Zahra Kaghat and Pierre Cubaud. "Fluid Interaction in Audio-Guided Museum Visit: Authoring Tool and Visitor Device". en. In: *VAST: International Symposium on Virtual Reality* (2010), Archaeology and Intelligent Cultural Heritage. DOI: [10.2312/VAST/VAST10/163\\$-\\$170](https://doi.org/10.2312/VAST/VAST10/163$-$170).
- [91] Daniela D'Auria et al. "Interactive Headphones for a Cloud 3D Audio Application". In: *2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. 2014, pp. 437–442. DOI: [10.1109/3PGCIC.2014.145](https://doi.org/10.1109/3PGCIC.2014.145).
- [92] E. Kabisch. "A periscope for mobile discovery and narrative". In: *CC2007 - Seeding Creativity: Tools, Media, and Environments* (2007), pp. 259–260.
- [93] K. Salo, D. Giova, and T. Mikkonen. "Backend infrastructure supporting audio augmented reality and storytelling". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Notes in Bioinformatics)* 9735 (2016), pp. 325–335.
- [94] Matthew McGinity et al. "AVIE: A versatile multi-user stereo 360° interactive VR theatre". In: *ACM International Conference Proceeding Series* (2007).
- [95] S. Beck. "The immersive computer-controlled audio sound theater: history and current trends in multi-modal sound diffusion". In: *SIGGRAPH 2009: Talks* (2009), p. 1.
- [96] Jason E. Summers. "Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms, and Acoustic Virtual Reality". In: *The Journal of the Acoustical Society of America* 123.6 (June 2008), pp. 4028–4029. DOI: [10.1121/1.2908264](https://doi.org/10.1121/1.2908264).
- [97] Michele Geronazzo, Jason Yves Tissieres, and Stefania Serafin. "A Minimal Personalization of Dynamic Binaural Synthesis with Mixed Structural Modeling and Scattering Delay Networks". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. IEEE, May 2020, pp. 411–415. DOI: [10.1109/icassp40776.2020.9053873](https://doi.org/10.1109/icassp40776.2020.9053873). URL: <http://dx.doi.org/10.1109/ICASSP40776.2020.9053873>.
- [98] Costas Boletsis and Dimitra Chasanidou. "Audio Augmented Reality in Public Transport for Exploring Tourist Sites". In: *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. NordiCHI '18. Oslo, Norway: Association for Computing Machinery, Sept. 2018, 721–725. ISBN: 9781450364379. DOI: [10.1145/3240167.3240243](https://doi.org/10.1145/3240167.3240243). URL: <https://doi-org.altais.uniud.it/10.1145/3240167.3240243>.

- [99] A. Kratky. "Walking in the Head: Methods of Sonic Augmented Reality Navigation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2019).
- [100] Y. Fu, D. Byrne, and L. Shea. "Evoking the Post-industrial Landscape Memories through Spectrality and Mixed Reality Soundscapes". In: *ACM International Conference Proceeding Series* (2021).
- [101] Jdp. Sardo et al. "A portable device for five sense augmented reality experiences in museums". In: *WSEAS Transactions on Environment and Development* 14 (2018), pp. 347–362.
- [102] Adv. Santiago, Pnm. Sampaio, and Lrs. Fernandes. "MOGRE-storytelling: Interactive creation of 3D stories". In: *Proceedings -2014 16th Symposium on Virtual and Augmented Reality, SVR 2014* (2014), pp. 190–199.
- [103] Adv. Santiago et al. "A digital approach to storytelling with MOGRE". In: *15th International Conference on Intelligent Games and Simulation* (2014), pp. 104–111.
- [104] Marek Hatala, Ron Wakkary, and Leila Kalantari. "Rules and ontologies in support of real-time ubiquitous application". In: *Journal of Web Semantics* 3.1 (July 2005), pp. 5–22. DOI: [10.1016/j.websem.2005.05.004](https://doi.org/10.1016/j.websem.2005.05.004).
- [105] Marek Hatala and Ron Wakkary. "Ontology-Based User Modeling in an Augmented Audio Reality System for Museums". In: *User Modeling and User-Adapted Interaction* 15.3–4 (Aug. 2005), 339–380. ISSN: 1573-1391. DOI: [10.1007/s11257-005-2304-5](https://doi.org/10.1007/s11257-005-2304-5). URL: <http://dx.doi.org/10.1007/s11257-005-2304-5>.
- [106] A. Damala and N. Stojanovic. "Tailoring the Adaptive Augmented Reality (A2R) museum visit: Identifying Cultural Heritage professionals' motivations and needs". In: *11th IEEE International Symposium on Mixed and Augmented Reality 2012 -Arts, Media, and Humanities Papers* (2012), pp. 71–80.
- [107] Brian F. G. Katz, Damian Murphy, and Angelo Farina. "The Past Has Ears (PHE): XR Explorations of Acoustic Spaces as Cultural Heritage". In: *Augmented Reality, Virtual Reality, and Computer Graphics*. Springer International Publishing, 2020, 91–98. ISBN: 9783030584689. DOI: [10.1007/978-3-300-03030-5_7](https://doi.org/10.1007/978-3-300-03030-5_7).
- [108] Sam Bilbow. "Evaluating polaris~ - An Audiovisual Augmented Reality Experience Built on Open-Source Hardware and Software". English. In: *NIME 2022*. PubPub, 2022. DOI: [10.21428/92fbeb44.8abb9ce6](https://doi.org/10.21428/92fbeb44.8abb9ce6). URL: <http://dx.doi.org/10.21428/92fbeb44.8abb9ce6>.
- [109] S. Meyer. "Right, left, high, low narrative strategies for non-linear storytelling". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2016).
- [110] E. Fujinawa et al. "Induction of human behavior by presentation of environmental acoustics". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9172 (2015), pp. 582–594.
- [111] Sarah Kenderdine, Ingrid Mason, and Lily Hibberd. "Computational Archives for Experimental Museology". In: *Communications in Computer and Information Science*. Springer International Publishing, 2021, 3–18. ISBN: 9783030836474. DOI: [10.1007/978-3-300-83647-4_1](https://doi.org/10.1007/978-3-300-83647-4_1).

- [112] Chris Rhodes. “Membrana Neopermeable”. In: *NIME 2022*. NIME. PubPub, 2022. DOI: [10.21428/92fbeb44.6e17eaf5](https://doi.org/10.21428/92fbeb44.6e17eaf5). URL: <http://dx.doi.org/10.21428/92fbeb44.6e17eaf5>.
- [113] A. Liberati et al. “The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration”. In: *BMJ* 339.jul21 1 (July 2009), b2700–b2700. DOI: [10.1136/bmj.b2700](https://doi.org/10.1136/bmj.b2700).
- [114] J. Vaquerizo-Serrano et al. “Catatonia in autism spectrum disorders: A systematic review and meta-analysis”. In: *European Psychiatry* 65.1 (Dec. 2021). DOI: [10.1192/j.eurpsy.2021.2259](https://doi.org/10.1192/j.eurpsy.2021.2259).
- [115] F. Kusunoki, M. Sugimoto, and H. Hashizume. “Toward an interactive museum guide system with sensing and wireless network technologies”. In: *Proceedings. IEEE International Workshop on Wireless and Mobile Technologies in Education. WMTE-02*. IEEE Comput. Soc, 2002. DOI: [10.1109/wmte.2002.1039228](https://doi.org/10.1109/wmte.2002.1039228). URL: <http://dx.doi.org/10.1109/WMTE.2002.1039228>.
- [116] Aaron F. Bobick et al. “The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment”. In: *Presence: Teleoperators and Virtual Environments* 8.4 (Aug. 1999), 369–393. ISSN: 1054–7460. DOI: [10.1162/105474699566297](https://doi.org/10.1162/105474699566297).
- [117] Anabela Marto et al. “A survey of multisensory VR and AR applications for cultural heritage”. In: *Computers & Graphics* 102 (Feb. 2022), pp. 426–440. DOI: [10.1016/j.cag.2021.10.001](https://doi.org/10.1016/j.cag.2021.10.001).
- [118] A. Tomi and Dra. Rambli. “An interactive mobile augmented reality magical playbook: Learning number with the thirsty crow”. In: *Procedia Computer Science* (2013), pp. 123–130.
- [119] P. Domínguez. “Audiovisual didactic storytelling through virtual reconstruction: The Iberian temple of la Alcudia in Elche”. In: *2016 International Symposium on Computers in Education, SIIE 2016: Learning Analytics Technologies* (2016).
- [120] S. Dow et al. “Exploring spatial narratives and mixed reality experiences in Oakland Cemetery”. In: *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology* (2005), pp. 51–60.
- [121] Mcr. Harrington. “Connecting User Experience to Learning in an Evaluation of an Immersive, Interactive, Multimodal Augmented Reality Virtual Diorama in a Natural History Museum and the Importance of Story”. In: *2020 6th International Conference of the Immersive Learning Research Network (iLRN)* (2020), pp. 70–78.
- [122] K. Rassmus-Gröhn, D. Szymczak, and C. Magnusson. “The time machine -An inclusive tourist guide application supporting exploration”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* 7989 (2013), pp. 127–136.
- [123] F. Fischnaller. “Virtual journey through the history of Fort Saint Jean, Marseille (VJ-FSJ Project) : Case Study: New Media exhibit at the Musée des civilisations de L’Europe et de la Méditerranée (MuCEM)”. In: *3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems Multimedia* (2018), pp. 1–8.
- [124] C. Zimmer et al. “War Children: Using AR in a Documentary Context”. In: *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2018), pp. 390–394.

- [125] Andrew E. Smith and Michael S. Humphreys. "Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping". In: *Behavior Research Methods* 38.2 (May 2006), pp. 262–279. DOI: [10.3758/bf03192778](https://doi.org/10.3758/bf03192778).
- [126] Matthew Chalmers and Paul Chitson. "Bead: explorations in information visualization". In: *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '92*. ACM Press, 1992. DOI: [10.1145/133160.133215](https://doi.org/10.1145/133160.133215).
- [127] Hwei Teeng Chong et al. "Comprehensive systematic review on virtual reality for cultural heritage practices: coherent taxonomy and motivations". In: *Multimedia Systems* 28.3 (Nov. 2021), 711–726. ISSN: 1432-1882. DOI: [10.1007/s00530-021-00869-4](https://doi.org/10.1007/s00530-021-00869-4). URL: <http://dx.doi.org/10.1007/s00530-021-00869-4>.
- [128] Jason Jerald. *The VR Book*. Association for Computing Machinery, Oct. 2015. ISBN: 9781970001129. DOI: [10.1145/2792790](https://doi.org/10.1145/2792790). URL: <http://dx.doi.org/10.1145/2792790>.
- [129] Nikunj Raghuvanshi and Hannes Gamper. "Interactive and Immersive Auratization". en. In: *Sonic Interactions in Virtual Environments*. Ed. by Michele Geronazzo and Stefania Serafin. Cham: Springer International Publishing, Oct. 2022, pp. 77–113. ISBN: 978-3-031-04021-4. DOI: [10.1007/978-3-031-04021-4_3](https://doi.org/10.1007/978-3-031-04021-4_3). URL: https://doi.org/10.1007/978-3-031-04021-4_3 (visited on 09/16/2024).
- [130] V. Välimäki, A. Franck, and J. Rämö. "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments". In: *IEEE Signal Processing Magazine* 32 (2015), pp. 92–99.
- [131] K. Kortbek and K. Grønbaek. "Interactive spatial multimedia for communication of art in the physical museum space". In: *Proceedings of the 16th ACM international conference on Multimedia* (2008), pp. 609–618.
- [132] K. Layng et al. "Cave: making collective virtual narrative". In: *ACM SIGGRAPH 2019 Art Gallery* (2019), pp. 1–8.
- [133] Daniela D'Auria et al. "Caruso: Interactive headphones for a dynamic 3D audio application in the cultural heritage context". In: *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*. 2014, pp. 525–528. DOI: [10.1109/IRI.2014.7051934](https://doi.org/10.1109/IRI.2014.7051934).
- [134] A. Tse et al. "Was I There?: Impact of Platform and Headphones on 360 Video Immersion". In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017), pp. 2967–2974.
- [135] M. Fakhour et al. "A Cultural Scavenger Hunt Serious Game Based on Audio Augmented Reality". In: *Advances in Intelligent Systems and Computing* 1102 AISC:1-8 (2020).
- [136] C. Popp and D. Murphy. "Establishment and Implementation of Guidelines for Narrative Audio-based Room-scale Virtual Reality using Practice-based Methods". In: *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference* (2022).
- [137] Michele Geronazzo et al. "Creating an Audio Story with Interactive Binaural Rendering in Virtual Reality". In: *Wireless Communications and Mobile Computing* 2019 (Nov. 2019), pp. 1–14. DOI: [10.1155/2019/1463204](https://doi.org/10.1155/2019/1463204). URL: <https://doi.org/10.1155/2019/1463204>.

- [138] F. Kaghat, A. Azough, and M. Fakhour. "SARIM: A gesture-based sound augmented reality interface for visiting museums". In: *2018 International Conference on Intelligent Systems and Computer Vision, ISCV 2018* (2018), pp. 1–9.
- [139] Fotis Giariskanis et al. "The Augmented Museum: A Multimodal, Game-Based, Augmented Reality Narrative for Cultural Heritage". In: *IMX 2022 - Proceedings of the 2022 ACM International Conference on Interactive Media Experiences*. Type: Conference paper. 2022, pp. 281–285. DOI: [10.1145/3505284.3532967](https://doi.org/10.1145/3505284.3532967).
- [140] Yannis Kritikos and Katerina Mania. "Interactive Historical Documentary in Virtual Reality". In: *2022 International Conference on Interactive Media, Smart Systems and Emerging Technologies (IMET)*. 2022, pp. 1–8. DOI: [10.1109/IMET54801.2022.9929500](https://doi.org/10.1109/IMET54801.2022.9929500).
- [141] Constantin Popp and Damian T. Murphy. "Creating Audio Object-Focused Acoustic Environments for Room-Scale Virtual Reality". In: *Applied Sciences (Switzerland)* 12.14 (2022). Type: Article. DOI: [10.3390/app12147306](https://doi.org/10.3390/app12147306).
- [142] Constantin Popp and Damian T. Murphy. "Establishment and Implementation of Guidelines for Narrative Audio-based Room-scale Virtual Reality using Practice-based Methods". In: *Proceedings of the AES International Conference*. Vol. 2022-August. Type: Conference paper. 2022, pp. 95–104.
- [143] Juan Carlos Vasquez Gomez. "Interactive Gamification for New Experimental Music: Initial Findings". In: *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '22. event-place: Bremen, Germany. Bremen, Germany: Association for Computing Machinery, 2022, 103–106. ISBN: 9781450392112. DOI: [10.1145/3505270.3558341](https://doi.org/10.1145/3505270.3558341).
- [144] Z. Erol et al. "SOUND OF(F): Contextual Storytelling Using Machine Learning Representations of Sound and Music". In: *Social-Informatics and Telecommunications Engineering* 422 (2022), pp. 332–345.
- [145] Richard Warp et al. "Validating the effects of immersion and spatial audio using novel continuous biometric sensor measures for Virtual Reality". en. In: *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. Singapore, Singapore: IEEE, Oct. 2022, pp. 262–265. ISBN: 978-1-66545-365-3. DOI: [10.1109/ISMAR\\$-Adjunct57072.2022.00058](https://doi.org/10.1109/ISMAR$-Adjunct57072.2022.00058).
- [146] Tycho Zaal, Alkim Almila Akdag Salah, and Wolfgang Hürst. "Toward inclusivity: Virtual reality museums for the visually impaired". In: *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. ISSN: 2771-7453. Dec. 2022, pp. 225–233. DOI: [10.1109/AIVR56993.2022.00047](https://doi.org/10.1109/AIVR56993.2022.00047).
- [147] Zizhen Fan et al. "The display and dissemination of Qiantang River poetry road culture based on Virtual Reality and Human-computer Interaction". In: *Proceedings - 2021 2nd International Conference on Intelligent Design, ICID 2021*. Type: Conference paper. 2021, pp. 120–128. DOI: [10.1109/ICID54526.2021.00032](https://doi.org/10.1109/ICID54526.2021.00032).
- [148] F. Heller et al. "Multi-user interaction in virtual audio spaces". In: *Conference on Human Factors in Computing Systems -Proceedings* (2009), pp. 4489–4494.
- [149] Michalis Tsepapadakis, Damianos Gavalas, and Panayiotis Koutsabasis. "3D Audio + Augmented Reality + AI Chatbots + IoT: An Immersive Conversational Cultural Guide". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13445 LNCS (2022), pp. 249–260.

- [150] Michalis Tsepapadakis and Damianos Gavalas. "Are you talking to me? An Audio Augmented Reality conversational guide for cultural heritage". In: *Pervasive and Mobile Computing* 92 (May 2023), p. 101797. ISSN: 1574-1192. DOI: [10.1016/j.pmcj.2023.101797](https://doi.org/10.1016/j.pmcj.2023.101797).
- [151] A. Steynberg. "Using Reverse Interactive Audio Systems (RIAS) to Direct Attention in Virtual Reality Narrative Practices: A Case Study". In: *International Conference on Interactive Digital Storytelling* (2020), pp. 353–356.
- [152] Susana Sanchez et al. "Embodied Reading: A Multisensory Experience." en. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. San Jose California USA: ACM, May 2016, pp. 1459–1466. ISBN: 978-1-4503-4082-3. DOI: [10.1145/2851581.2892353](https://doi.org/10.1145/2851581.2892353).
- [153] Larry Cutler et al. "From quest to quill". In: *ACM SIGGRAPH 2021 Production Sessions*. SIGGRAPH '21. Virtual Event, USA: ACM, Aug. 2021. ISBN: 9781450383226. DOI: [10.1145/3446368.3452126](https://doi.org/10.1145/3446368.3452126). URL: <https://doi.org/10.1145/3446368.3452126>.
- [154] Alicia Butterworth. "Beyond sonic realism: a cinematic sound approach in documentary 360° film". In: *Studies in Documentary Film* 16.2 (2022). Type: Article, pp. 156–173. DOI: [10.1080/17503280.2022.2048234](https://doi.org/10.1080/17503280.2022.2048234).
- [155] Hitesh Kumar Chaurasia and Manoj Majhi. "Sound Design for Cinematic Virtual Reality: A State-of-the-Art Review". In: *Lecture Notes in Networks and Systems* 391 (2022). Type: Conference paper, pp. 357–368.
- [156] Gerard T. Mulvany. "Because the Night - Immersive Theatre for Digital Audiences: Mapping the affordances of immersive theatre to digital interactions using game engines". In: *CHI PLAY 2022 - Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*. Type: Conference paper. 2022, pp. 291–296. DOI: [10.1145/3505270.3558374](https://doi.org/10.1145/3505270.3558374).
- [157] Justin Paterson and Oliver Kadel. "Audio for extended realities: A case study informed exposition". In: *Convergence* (June 2023). Publisher: SAGE Publications Ltd, p. 13548565231169723. ISSN: 1354-8565. DOI: [10.1177/13548565231169723](https://doi.org/10.1177/13548565231169723).
- [158] Jin Cui, Feifei Dong, and Wenli Zhang. "Immersive Virtual Experience Design Research of Jiaoshan Forest of Steles Calligraphy Art". In: *2020 International Conference on Innovation Design and Digital Technology (ICIDDT)*. Zhenjing, China: IEEE, Dec. 2020, pp. 165–169. ISBN: 978-1-66540-367-2. DOI: [10.1109/ICIDDT52279.2020.00037](https://doi.org/10.1109/ICIDDT52279.2020.00037).
- [159] Achim Hättich and Martina Schweizer. "I hear what you see: Effects of audio description used in a cinema on immersion and enjoyment in blind and visually impaired people". In: *British Journal of Visual Impairment* 38.3 (2020). Type: Article, pp. 284–298. DOI: [10.1177/0264619620911429](https://doi.org/10.1177/0264619620911429).
- [160] S. Andolina et al. "Designing for Mixed Reality Urban Exploration". In: *Interaction Design and Architecture(s)* 48 (2021), pp. 33–49.
- [161] Anabela Marto et al. "Development and Evaluation of an Outdoor Multisensory AR System for Cultural Heritage". In: *IEEE Access* 9 (2021), pp. 16419–16434. DOI: [10.1109/access.2021.3050974](https://doi.org/10.1109/access.2021.3050974).
- [162] P.B. Fellgett. "Ambisonic reproduction of sound". In: *Electronics and Power* 19.20 (1973), p. 492. DOI: [10.1049/ep.1973.0597](https://doi.org/10.1049/ep.1973.0597).

- [163] G. Struck, R. Böse, and U. Spierling. "Trying to get trapped in the past - Exploring the illusion of presence in virtual drama". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* 5334 (2008), pp. 114–125.
- [164] Timothy M Gmeiner and Eito Murakami. "Pigments of Imagination: An Interactive Virtual Reality Composition". en. In: *29th ACM Symposium on Virtual Reality Software and Technology*. Christchurch New Zealand: ACM, Oct. 2023. ISBN: 9798400703287. DOI: [10.1145/3611659.3617201](https://doi.org/10.1145/3611659.3617201).
- [165] M. Lopez and S. Pauletto. "The Sound Machine: A study in storytelling through sound design". In: *Proceedings of the 5th Audio Mostly -A Conference on Interaction With Sound* (2010).
- [166] Marta Gospodarek et al. "Sound design and reproduction techniques for co-located narrative VR experiences". In: *Audio Engineering Society Convention* 147 (2019).
- [167] Y. Li, P. Tennent, and S. Cobb. "Appropriate Control Methods for Mobile Virtual Exhibitions". In: *Duguleană M, Carrozzino M, Gams M, Tanea I (eds) VR Technologies in Cultural Heritage* (2019), pp. 165–183.
- [168] C. Magnusson, K. Rassmus-Gröhn, and D. Szymczak. "Exploring history: A mobile inclusive virtual tourist guide". In: *Proceedings of the NordiCHI 2014: The 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (2014), pp. 69–78.
- [169] D. Szymczak et al. "A realworld study of an audio-tactile tourist guide". In: *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services* (2012), pp. 335–344.
- [170] Olivia Guntarik, Hugh Davies, and Troy Innocent. "Indigenous Cartographies: Pervasive Games and Place-Based Storytelling". In: *Space and Culture* (Mar. 2023). Publisher: SAGE Publications Inc, p. 12063312231155348. ISSN: 1206-3312. DOI: [10.1177/12063312231155348](https://doi.org/10.1177/12063312231155348).
- [171] Lee Scott. "Creating opera for mobile media: Artistic opportunities and technical limitations". In: *Proceedings - 14th International Symposium on Pervasive Systems, Algorithms and Networks, I-SPAN 2017, 11th International Conference on Frontier of Computer Science and Technology, FCST 2017 and 3rd International Symposium of Creative Computing, ISCC 2017 2017-November* (2017). DOI: [10.1109/ISPAN\\$-FCST\\$-ISCC.2017.86](https://doi.org/10.1109/ISPAN$-FCST$-ISCC.2017.86).
- [172] Boriana Koleva et al. "Designing Hybrid Gifts". In: *ACM Transactions on Computer-Human Interaction* 27.5 (Aug. 2020), pp. 1–33. DOI: [10.1145/3398193](https://doi.org/10.1145/3398193). URL: <https://doi.org/10.1145/3398193>.
- [173] Fatima Zahra Kaghat et al. "A new audio augmented reality interaction and adaptation model for museum visits". In: *Computers & Electrical Engineering* 84 (June 2020), p. 106606. DOI: [10.1016/j.compeleceng.2020.106606](https://doi.org/10.1016/j.compeleceng.2020.106606).
- [174] Jesús Gimeno et al. "Combining traditional and indirect augmented reality for indoor crowded environments. A case study on the Casa Batlló museum". In: *Computers & Graphics* 69 (Dec. 2017), pp. 92–103. DOI: [10.1016/j.cag.2017.09.001](https://doi.org/10.1016/j.cag.2017.09.001).
- [175] D. Ahmetovic et al. "MusA: artwork accessibility through augmented reality for people with low vision". In: *Proceedings of the 18th International Web for All Conference* (2021), W4A–2021.

- [176] Ana Javornik et al. "An experimental study on the role of augmented reality content type in an outdoor site exploration". In: *Behaviour & Information Technology* 38.1 (Aug. 2018), pp. 9–27. DOI: [10.1080/0144929x.2018.1505950](https://doi.org/10.1080/0144929x.2018.1505950).
- [177] A. Dünser and E. Hornecker. "An observational study of children interacting with an augmented story book". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* 4469 (2007), pp. 305–315.
- [178] X. He and Y. Hong. "The effect of augmented reality on the memorization in history and humanities education". In: *Advances in Intelligent Systems and Computing* 1217 AISC (2020), pp. 769–776.
- [179] E. Corrigan-Kavanagh, C. Scarles, and G. Revill. "Augmenting travel guides for enriching travel experiences". In: *e-Review of Tourism Research* 17 (2019), pp. 334–348.
- [180] F. Sorrentino, L. Spano, and R. Scateni. "SuperAvatar: Children and mobile tourist guides become friends using superpowered avatars". In: *Proceedings of 2015 International Conference on Interactive Mobile Communication Technologies and Learning* (2015), pp. 222–226.
- [181] Frédérique Krupa. "InnovART2 : À l'écoute des chantiers: An industrial heritage sound-and-augmented-reality walking tour". en. In: *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*. Nantes France: ACM, June 2023, pp. 376–378. ISBN: 9798400700286. DOI: [10.1145/3573381.3597223](https://doi.org/10.1145/3573381.3597223).
- [182] Wei-Hsin Huang, Huei-Ming Chiao, and Wei-Hsin Huang. "Innovative Research on the Development of Game-based Tourism Information Services Using Component-based Software Engineering". In: *Advances in Science, Technology and Engineering Systems Journal* 3.1 (Feb. 2018), pp. 451–459. DOI: [10.25046/aj030155](https://doi.org/10.25046/aj030155). URL: <https://doi.org/10.25046/aj030155>.
- [183] F. Bertacchini et al. "Learning in the smart city: A virtual and augmented museum devoted to chaos theory". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* 7697 (2014), pp. 261–270.
- [184] J. Ogle et al. "First World War tunnel warfare through haptic VR". In: *ACM SIGGRAPH 2020 Immersive Pavilion* (2020), pp. 1–2.
- [185] J. Bargsten. "Narrative and Spatial Design through Immersive Music and Audio". In: *Proceedings -2020 IEEE Conference on Virtual Reality and 3D User Interfaces* 2020 (2020), pp. 396–400.
- [186] Yannis Kritikos et al. "Audio Augmented Reality Outdoors". en. In: *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*. Nantes France: ACM, June 2023, pp. 199–204. ISBN: 9798400700286. DOI: [10.1145/3573381.3597028](https://doi.org/10.1145/3573381.3597028).
- [187] Stephen Barrass. "Creative practice-based research in interaction design". In: *Computers in Entertainment* 6.3 (Oct. 2008), pp. 1–17. DOI: [10.1145/1394021.1394026](https://doi.org/10.1145/1394021.1394026). URL: <https://doi.org/10.1145/1394021.1394026>.
- [188] S. Kenderdine. "Immersive visualization architectures and situated embodiments of culture and heritage". In: *Proceedings of the International Conference on Information Visualisation* (2010), pp. 408–414.

- [189] C. Kelling et al. "Implications of audio and narration in the user experience design of virtual reality". In: *ACM International Conference Proceeding Series* (2018), pp. 258–261.
- [190] Tck. Kwok et al. "Gaze-Guided Narratives: Adapting Audio Guide Content to Gaze in Virtual and Real Environments". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12.
- [191] Caitlin Kelly. "A Study on Sound and Sound Design for Accessibility in Cultural Heritage Contexts in Ireland". en. In: *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. Bologna, Italy: IEEE, Sept. 2023, pp. 1–9. ISBN: 9798350311044. DOI: [10.1109/I3DA57090.2023.10289582](https://doi.org/10.1109/I3DA57090.2023.10289582).
- [192] T. Deacon and M. Barthet. "Spatial Design Considerations for Interactive Audio in Virtual Reality". In: *Sonic Interaction in Virtual Environment* (2022).
- [193] P. Matthias, M. Billinghamurst, and Z. See. "This land AR: An Australian music and sound XR installation a transmedia storytelling approach". In: *Proceedings -VRCAI 2019: 17th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry* (2019).
- [194] S. Huws, A. John, and J. Kidd. "Evaluating the affective dimensions of Traces-Olion: a subtle mob at St Fagans National Museum of History, Wales". In: *3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems Multimedia (VSMM 2018)* (2018), pp. 1–8.
- [195] Kuo-En Chang et al. "Development and behavioral pattern analysis of a mobile guide system with augmented reality for painting appreciation instruction in an art museum". In: *Computers & Education* 71 (Feb. 2014), pp. 185–197. DOI: [10.1016/j.compedu.2013.09.022](https://doi.org/10.1016/j.compedu.2013.09.022).
- [196] Chiara Teneggi et al. "Social Modulation of Peripersonal Space Boundaries". In: *Current Biology* 23.5 (Mar. 2013), pp. 406–411. DOI: [10.1016/j.cub.2013.01.043](https://doi.org/10.1016/j.cub.2013.01.043).
- [197] Rebecca A. Grier et al. "The System Usability Scale". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57.1 (Sept. 2013), pp. 187–191. DOI: [10.1177/1541931213571042](https://doi.org/10.1177/1541931213571042).
- [198] Volkan Sevinc and Mehmet Ilker Berkman. "Psychometric evaluation of Simulator Sickness Questionnaire and its variants as a measure of cybersickness in consumer virtual environments". In: *Applied Ergonomics* 82 (Jan. 2020), p. 102958. DOI: [10.1016/j.apergo.2019.102958](https://doi.org/10.1016/j.apergo.2019.102958).
- [199] Rick Busselle and Helena Bilandzic. "Measuring Narrative Engagement". In: *Media Psychology* 12.4 (Nov. 2009), 321–347. ISSN: 1532–785X. DOI: [10.1080/15213260903287259](https://doi.org/10.1080/15213260903287259).
- [200] Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. "The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions". In: *PLOS ONE* 11.8 (Aug. 2016). Ed. by André Aleman, e0159915. DOI: [10.1371/journal.pone.0159915](https://doi.org/10.1371/journal.pone.0159915).
- [201] D. Jurica et al. "Comparison of Two Methods of Soundscape Evaluation". In: *2nd International Colloquium on Smart Grid Metrology (SMAGRIMET)* (2019), pp. 1–5.

- [202] W. Mansilla. "Interactive dramaturgy by generating acousmètre in a virtual environment". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4161 (2006), pp. 90–95.
- [203] Samer Y. Al-Imamy. "Blending printed texts with digital resources through augmented reality interaction". In: *Education and Information Technologies* 25.4 (Dec. 2019), pp. 2561–2576. DOI: [10.1007/s10639-019-10070-w](https://doi.org/10.1007/s10639-019-10070-w).
- [204] Shih-Jou Yu, Jerry Chih-Yuan Sun, and Oscar Tzyh-Chiang Chen. "Effect of AR-based online wearable guides on university students' situational interest and learning performance". In: *Universal Access in the Information Society* 18.2 (Nov. 2017), pp. 287–299. DOI: [10.1007/s10209-017-0591-3](https://doi.org/10.1007/s10209-017-0591-3).
- [205] Michele Geronazzo and Stefania Serafin. "Sonic Interactions in Virtual Environments". en. In: *Human-Computer Interaction Series*. Human-Computer Interaction Series (Oct. 2023). Ed. by Michele Geronazzo and Stefania Serafin, 77–113. ISSN: 2524-4477. DOI: [10.1007/978-3-031-04021-4](https://doi.org/10.1007/978-3-031-04021-4). URL: <http://dx.doi.org/10.1007/978-3-031-04021-4> (visited on 10/17/2022).
- [206] Mel Slater, Andrea Brogni, and Anthony Steed. "Physiological responses to breaks in presence: A pilot study". In: *Presence 2003: The 6th annual international workshop on presence*. Vol. 157. 2003.
- [207] D. Williams and I. Daly. "Neuro-curation: A case study on the use of sonic enhancement of virtual museum exhibits". In: *Audio Mostly 2021* (2021), pp. 121–125.
- [208] T. Piumsomboon et al. "Empathic Mixed Reality: Sharing What You Feel and Interacting with What You See". In: *2017 International Symposium on Ubiquitous Virtual Reality (ISUVR)* (2017), pp. 38–41.
- [209] M. Kitajima, S. Shimizu, and K. Nakahira. "Creating memorable experiences in virtual reality: Theory of its processes and preliminary eye-tracking study using omnidirectional movies with audio-guide". In: *3rd IEEE International Conference on Cybernetics* (2017), pp. 1–8.
- [210] German I. Parisi et al. "Continual lifelong learning with neural networks: A review". In: *Neural Networks* 113 (May 2019), pp. 54–71. DOI: [10.1016/j.neunet.2019.01.012](https://doi.org/10.1016/j.neunet.2019.01.012).
- [211] Lénaïc B. Cadet and Hanna Chainay. "Memory of virtual experiences: Role of immersion, emotion and sense of presence". In: *International Journal of Human-Computer Studies* 144 (Dec. 2020), p. 102506. DOI: [10.1016/j.ijhcs.2020.102506](https://doi.org/10.1016/j.ijhcs.2020.102506).
- [212] Samer Y. Al-Imamy. "Blending printed texts with digital resources through augmented reality interaction". In: *Education and Information Technologies* 25.4 (Dec. 2019), pp. 2561–2576. DOI: [10.1007/s10639-019-10070-w](https://doi.org/10.1007/s10639-019-10070-w).
- [213] Alessandro Giuseppe Privitera, Federico Fontana, Michele Geronazzo, et al. "Personalization in Audio Storytelling within Virtual and Augmented Reality: State of the Art and Insights". In: *10th Convention of the European Acoustics Association*. 2023.

- [214] Michele Geronazzo, Jason Yves Tissieres, and Stefania Serafin. "A Minimal Personalization of Dynamic Binaural Synthesis with Mixed Structural Modeling and Scattering Delay Networks". In: *Proc. IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP 2020)*. Barcelona, Spain, May 2020, pp. 411–415. ISBN: 978-1-5090-6631-5. DOI: [ICASSP40776.2020.9053873](https://doi.org/10.5824/1309-1581.2020.9053873).
- [215] Barboros Bostan and Tim Marsh. "Fundamentals Of Interactive Storytelling". In: *Summer 3.8* (Aug. 2012), pp. 19–42. DOI: [10.5824/1309-1581.2012.3.002.x](https://doi.org/10.5824/1309-1581.2012.3.002.x). URL: <https://doi.org/10.5824/1309-1581.2012.3.002.x>.
- [216] M. Riedl, D. Thue, and M. Gomez-Martín. "Game AI as Storytelling". In: *Artificial Intelligence for Computer Games* (2011), pp. 125–150.
- [217] Poo. Hernandez et al. "Emotion-Based Interactive Storytelling with Artificial Intelligence". In: *AIIDE 10* (2021), pp. 146–152.
- [218] Galena Pisoni et al. "Human-Centered Artificial Intelligence for Designing Accessible Cultural Heritage". In: *Applied Sciences* 11.2 (Jan. 2021), p. 870. DOI: [10.3390/app11020870](https://doi.org/10.3390/app11020870). URL: <https://doi.org/10.3390/app11020870>.
- [219] C. Boletsis and D. Chasanidou. "Smart tourism in cities: Exploring urban destinations with audio augmented reality". In: *ACM International Conference Proceeding Series* (2018), pp. 515–521.
- [220] Pavel Zahorik and Rick L. Jenison. "Presence as Being-in-the-World". In: *Presence: Teleoperators and Virtual Environments* 7.1 (Feb. 1998), pp. 78–89. DOI: [10.1162/105474698565541](https://doi.org/10.1162/105474698565541).
- [221] Francisco J Varela, Evan T Thompson, and Eleanor Rosch. *The embodied mind*. The MIT Press. London, England: MIT Press, Nov. 1992.
- [222] S. Perry et al. "Moving beyond the virtual museum: Engaging visitors emotionally". In: *23rd International Conference on Virtual System & Multimedia (VSMM)* (2017), pp. 1–8.
- [223] William Gaver et al. "Emergence as a Feature of Practice-based Design Research". In: *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. DIS '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 517–526. ISBN: 978-1-4503-9358-4. DOI: [10.1145/3532106.3533524](https://doi.org/10.1145/3532106.3533524). URL: <https://dl.acm.org/doi/10.1145/3532106.3533524> (visited on 08/19/2023).
- [224] Heather Wiltse. *Relating to Things: Design, Technology and the Artificial*. en. Google-Books-ID: SyTXDwAAQBAJ. Bloomsbury Publishing, May 2020. ISBN: 978-1-350-12427-1.
- [225] R. Skarbez et al. "Immersive Analytics: Theory and Research Agenda". In: *Frontiers in Robotics and AI* 6 (2019).
- [226] T. Chandler et al. "Immersive analytics". In: *Big Data Visual Analytics (BDVA)* (Sept. 2015), p. 7314296.
- [227] Thomas Hermann. "Taxonomy and definitions for sonification and auditory display". In: *Proceedings of the 14th International Conference on Auditory Display*. International Community for Auditory Display, 2008.

- [228] K. Koebel, D. Agotai, and A. undefinedöltekin. "EXPLORING CULTURAL HERITAGE COLLECTIONS IN IMMERSIVE ANALYTICS: CHALLENGES, BENEFITS, AND A CASE STUDY USING VIRTUAL REALITY". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B4-2020 (Aug. 2020), 599–606. ISSN: 2194-9034. DOI: [10.5194/isprs-archives-xliii-b4-2020-599-2020](https://doi.org/10.5194/isprs-archives-xliii-b4-2020-599-2020). URL: <http://dx.doi.org/10.5194/isprs-archives-XLIII-B4-2020-599-2020>.
- [229] Sai Anirudh Athaluri et al. "Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References". In: *Cureus* (Apr. 2023). ISSN: 2168-8184. DOI: [10.7759/cureus.37432](https://doi.org/10.7759/cureus.37432). URL: <http://dx.doi.org/10.7759/cureus.37432>.
- [230] Giuseppe Ciaburro et al. "The acoustics of ancient catacombs in Southern Italy". In: *Building Acoustics* 28.4 (Oct. 2020), pp. 411–422. DOI: [10.1177/1351010x20967571](https://doi.org/10.1177/1351010x20967571).
- [231] Olivier Warusfel and Sibylle Emerit. "Assessing the Acoustics of an Ancient Egyptian temple". In: *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, Sept. 2021. DOI: [10.1109/i3da48870.2021.9610973](https://doi.org/10.1109/i3da48870.2021.9610973). URL: <http://dx.doi.org/10.1109/I3DA48870.2021.9610973>.
- [232] H. Tahvanainen, H. Matsuda, and R. Shinoda. "Numerical simulation of the acoustic guitar for virtual prototyping". In: *Proceedings of ISMA 2019* (2019), pp. 13–17.
- [233] Bradley Lehman. "Bach's extraordinary temperament: Our rosetta stone—1". In: *Early music* 33.1 (2005), pp. 3–24.
- [234] Lucile Cossou et al. "Tangible Sounds: An Audio System for Interactive Tabletops". In: *Proceedings of the 1st International Conference on Digital Tools & Uses Congress - DTUC '18*. DTUC '18. ACM Press, 2018. DOI: [10.1145/3240117.3240129](https://doi.org/10.1145/3240117.3240129). URL: <http://dx.doi.org/10.1145/3240117.3240129>.
- [235] Mario Montagud, Pilar Orero, and Anna Matamala. "Culture 4 all: accessibility-enabled cultural experiences through immersive VR360 content". In: *Personal and Ubiquitous Computing* 24.6 (Jan. 2020), pp. 887–905. DOI: [10.1007/s00779-020-01357-3](https://doi.org/10.1007/s00779-020-01357-3).
- [236] A. Montuwy, B. Cahour, and A. Dommès. "Visual, auditory and haptic navigation feedbacks among older pedestrians". In: *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2017), pp. 1–8.
- [237] T. Barbieri, A. Bianchi, and L. Sbattella. "Minus-two: Multimedia, sound spatialization and 3D representation for cognitively impaired children". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* 3118 (2004), pp. 1054–1061.
- [238] Marco di Furia, Piergiorgio Guarini, and Francesca Finestrone. "Digital Storytelling as teaching methodology for special needs teachers: a case study". In: *Proceedings of the Third Workshop on Technology Enhanced Learning Environments for Blended Education*. 2022, pp. 10–11.
- [239] Iason Kastanis and Mel Slater. "Reinforcement learning utilizes proxemics". In: *ACM Transactions on Applied Perception* 9.1 (Mar. 2012), pp. 1–15. DOI: [10.1145/2134203.2134206](https://doi.org/10.1145/2134203.2134206). URL: <https://doi.org/10.1145/2134203.2134206>.

- [240] Enzo De Sena, Huseyin Hacıhabiboglu, and Zoran Cvetkovic. "Scattering Delay Network: An Interactive Reverberator for Computer Games". English. In: Audio Engineering Society, Feb. 2011. (Visited on 05/14/2015).
- [241] Michael Vorlander. *Auralization*. en. 2008th ed. RWTHedition. Berlin, Germany: Springer, Oct. 2007.
- [242] Bosun Xie. *Head-related transfer function and virtual auditory display*. en. 2nd ed. Boca Raton, FL: J Ross Publishing, Aug. 2013.
- [243] John C. Middlebrooks. "Individual differences in external-ear transfer functions reduced by scaling in frequency". In: *The Journal of the Acoustical Society of America* 106.3 (Sept. 1999), 1480–1492. ISSN: 1520-8524. DOI: [10.1121/1.427176](https://doi.org/10.1121/1.427176). URL: <http://dx.doi.org/10.1121/1.427176>.
- [244] Durand Begault and Elizabeth Wenzel. "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source". In: Oct. 2001, pp. 904–916.
- [245] Sean E. Olive. "The Perception and Measurement of Headphone Sound Quality: What Do Listeners Prefer?" In: *Acoustics Today* 18.1 (2021), p. 58. ISSN: 1557-0215. DOI: [10.1121/at.2022.18.1.58](https://doi.org/10.1121/at.2022.18.1.58). URL: <http://dx.doi.org/10.1121/AT.2022.18.1.58>.
- [246] Klaus Genuit. "The sound quality of vehicle interior noise: a challenge for the NVH-engineers". In: *International Journal of Vehicle Noise and Vibration* 1.1/2 (2004), p. 158. ISSN: 1479-148X. DOI: [10.1504/ijvnv.2004.004079](https://doi.org/10.1504/ijvnv.2004.004079). URL: <http://dx.doi.org/10.1504/IJNV.2004.004079>.
- [247] Stefan A. Wirler, Nils Meyer-Kahlen, and Sebastian J. Schlecht. "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes". In: *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Aug. 2020. URL: <http://www.aes.org/e-lib/browse.cfm?elib=20881>.
- [248] Jont B. Allen and David A. Berkley. "Image method for efficiently simulating small-room acoustics". In: *The Journal of the Acoustical Society of America* 65.4 (Apr. 1979), pp. 943–950. DOI: [10.1121/1.382599](https://doi.org/10.1121/1.382599). URL: <https://doi.org/10.1121/1.382599>.
- [249] Cédric Foy, Antoine Deleforge, and Diego Di Carlo. "Mean absorption estimation from room impulse responses using virtually supervised learning". In: *The Journal of the Acoustical Society of America* 150.2 (Aug. 2021), pp. 1286–1299. DOI: [10.1121/10.0005888](https://doi.org/10.1121/10.0005888). URL: <https://doi.org/10.1121/10.0005888>.
- [250] Hélène Bahu and David Romblom. "Optimization and Prediction of the Spherical and Ellipsoidal ITD Model Parameters Using Offset Ears". In: *Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science*. Audio Engineering Society. 2018.
- [251] Roberto Barumerli et al. "A Bayesian Model for Human Directional Localization of Broadband Static Sound Sources". en. In: *Acta Acustica* 7 (2023). Publisher: EDP Sciences, p. 12. ISSN: 2681-4617. DOI: [10.1051/aacus/2023006](https://doi.org/10.1051/aacus/2023006). URL: <https://acta-acustica.edpsciences.org/articles/aacus/abs/2023/01/aacus210056/aacus210056.html> (visited on 05/02/2023).

- [252] V.R. Algazi et al. "The CIPIC HRTF Database". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*. IEEE. IEEE, 2001, pp. 99–102. DOI: [10 . 1109 / ASPAA . 2001 . 969552](https://doi.org/10.1109/ASPAA.2001.969552).
- [253] Michele Geronazzo et al. "Applying a Single-notch Metric to image-guided Head-Related Transfer Function Selection for Improved Vertical Localization". In: *Journal of the Audio Engineering Society* 67.6 (June 2019), pp. 414–428. DOI: <https://doi.org/10.17743/jaes.2019.0010>.
- [254] Valentin Holzwarth et al. "Comparing the Accuracy and Precision of SteamVR Tracking 2.0 and Oculus Quest 2 in a Room Scale Setup". In: *2021 the 5th International Conference on Virtual and Augmented Reality Simulations*. ACM, Mar. 2021. DOI: [10 . 1145 / 3463914 . 3463921](https://doi.org/10.1145/3463914.3463921). URL: <https://doi.org/10.1145/3463914.3463921>.
- [255] David Poirier-Quinot and Martin S. Lawless. "Impact of wearing a head-mounted display on localization accuracy of real sound sources". en. In: *Acta Acustica* 7 (2023). Publisher: EDP Sciences, p. 3. ISSN: 2681-4617. DOI: [10 . 1051 / aacus / 2022055](https://doi.org/10.1051/aacus/2022055). URL: <https://acta-acustica.edpsciences.org/articles/aacus/abs/2023/01/aacus220027/aacus220027.html> (visited on 04/14/2023).
- [256] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, Oct. 1996. ISBN: 9780262268684. DOI: [10 . 7551 / mitpress / 6391.001.0001](https://doi.org/10.7551/mitpress/6391.001.0001). URL: <https://doi.org/10.7551/mitpress/6391.001.0001>.
- [257] Alexander Lindau and Stefan Weinzierl. "Assessing the plausibility of virtual acoustic environments". In: *Acta Acustica united with Acustica* 98.5 (2012), pp. 804–810. ISSN: 1610-1928. DOI: [doi:10 . 3813 / AAA . 918562](https://doi.org/10.3813/AAA.918562). URL: <https://www.ingentaconnect.com/content/dav/aaua/2012/00000098/00000005/art00014>.
- [258] Chris Pike, Frank Melchior, and Tony Tew. "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room". In: *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society. Aug. 2014.
- [259] Annika Neidhardt, Alby Ignatious-Tommy, and Anson Davis Pereppadan. "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets". In: *Audio Engineering Society Convention 144*. Audio Engineering Society. May 2018.
- [260] Annika Neidhardt and Anna Maria Zerlik. "The availability of a hidden real reference affects the plausibility of position-dynamic auditory AR". In: *Frontiers in Virtual Reality* 2 (2021), p. 102. ISSN: 2673-4192. DOI: [10 . 3389 / frvir . 2021 . 678875](https://doi.org/10.3389/frvir.2021.678875). URL: <https://www.frontiersin.org/articles/10.3389/frvir.2021.678875> (visited on 09/07/2021).
- [261] Johannes M Arend et al. "Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response". In: *Journal of the Audio Engineering Society* 69.7/8 (2021), pp. 557–575. DOI: <https://doi.org/10.17743/jaes.2021.0009>.
- [262] Kajetan Enge, Matthias Frank, and Robert Höldrich. "Listening experiment on the plausibility of acoustic modeling in virtual reality". In: *Fortschritte der Akustik–DAGA*. Vol. 46. 2020, pp. 13–16.

- [263] Will Bailey and Bruno Fazenda. "The effect of visual cues and binaural rendering method on plausibility in virtual environments". In: *Audio Engineering Society Convention 144*. Audio Engineering Society. May 2018.
- [264] Ilias Bergström et al. "The plausibility of a string quartet performance in virtual reality". In: *IEEE transactions on visualization and computer graphics* 23.4 (Apr. 2017), pp. 1352–1359. ISSN: 2160-9306. DOI: [10.1109/TVCG.2017.2657138](https://doi.org/10.1109/TVCG.2017.2657138). URL: <http://dx.doi.org/10.1109/TVCG.2017.2657138>.
- [265] Josefa Oberem, Bruno Masiero, and Janina Fels. "Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods". In: *Applied Acoustics* 114 (2016), pp. 71–78. ISSN: 0003-682X. DOI: <https://doi.org/10.1016/j.apacoust.2016.07.009>. URL: <https://www.sciencedirect.com/science/article/pii/S0003682X16302031>.
- [266] G Christopher Stecker et al. "Toward objective measures of auditory co-immersion in virtual and augmented reality". In: *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society. Aug. 2018.
- [267] Stefan A Wirler, Nils Meyer-Kahlen, and Sebastian J Schlecht. "Towards transfer-plausibility for evaluating mixed reality audio in complex scenes". In: *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society. Aug. 2020.
- [268] Julie R Williamson et al. "Digital proxemics: Designing social and collaborative interaction in virtual environments". In: *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022.
- [269] Toqeer Ali Syed et al. "In-Depth Review of Augmented Reality: Tracking Technologies, Development Tools, AR Displays, Collaborative AR, and Security Concerns". In: *Sensors* 23.1 (Dec. 2022), p. 146. DOI: [10.3390/s23010146](https://doi.org/10.3390/s23010146). URL: <https://doi.org/10.3390/s23010146>.
- [270] Alison Keogh et al. "Comparing the Usability and Acceptability of Wearable Sensors Among Older Irish Adults in a Real-World Context: Observational Study". In: *JMIR mHealth and uHealth* 8.4 (Apr. 2020), e15704. DOI: [10.2196/15704](https://doi.org/10.2196/15704). URL: <https://doi.org/10.2196/15704>.
- [271] Tiffany C.K. Kwok et al. "Gaze-Guided Narratives". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019. DOI: [10.1145/3290605.3300721](https://doi.org/10.1145/3290605.3300721). URL: <https://doi.org/10.1145/3290605.3300721>.
- [272] Fabio Arena et al. "An Overview of Augmented Reality". In: *Computers* 11.2 (Feb. 2022), p. 28. DOI: [10.3390/computers11020028](https://doi.org/10.3390/computers11020028). URL: <https://doi.org/10.3390/computers11020028>.
- [273] Olli S Rummukainen et al. "Audio Quality Evaluation in Virtual Reality: Multiple Stimulus Ranking with Behavior Tracking". English. In: *AES International Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA: Audio Engineering Society, Aug. 2018, pp. 1–10. URL: <http://www.aes.org/e-lib/browse.cfm?elib=19678>.
- [274] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Advances in Psychology*. Elsevier, 1988, pp. 139–183. DOI: [10.1016/s0166-4115\(08\)62386-9](https://doi.org/10.1016/s0166-4115(08)62386-9). URL: [https://doi.org/10.1016/s0166-4115\(08\)62386-9](https://doi.org/10.1016/s0166-4115(08)62386-9).

- [275] Peter A Hancock and N Meshkati, eds. *Human Mental Workload*. Advances in Psychology. North-Holland, May 2014.
- [276] Michele Geronazzo et al. “Acoustic selfies for extraction of external ear features in mobile audio augmented reality”. In: *Proc. 22nd ACM Symposium on Virtual Reality Software and Technology (VRST 2016)*. Munich, Germany: ACM, Nov. 2016, pp. 23–26. ISBN: 978-1-4503-4491-3. DOI: [10.1145/2993369.2993376](https://doi.org/10.1145/2993369.2993376).
- [277] Vani G. Rajendran and Hannes Gamper. “Spectral manipulation improves elevation perception with non-individualized head-related transfer functions”. en. In: *The Journal of the Acoustical Society of America* 145.3 (Mar. 2019), EL222–EL228. ISSN: 0001-4966. DOI: [10.1121/1.5093641](https://doi.org/10.1121/1.5093641). URL: <http://asa.scitation.org/doi/10.1121/1.5093641> (visited on 03/13/2019).
- [278] Rapolas Daugintis et al. “Classifying Non-Individual Head-Related Transfer Functions with A Computational Auditory Model: Calibration And Metrics”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–5. DOI: [10.1109/ICASSP49357.2023.10095152](https://doi.org/10.1109/ICASSP49357.2023.10095152).
- [279] Adnan Aijaz et al. “Demystifying the performance of Bluetooth mesh: Experimental evaluation and optimization”. In: vol. 2021-June. Cited by: 2; All Open Access, Green Open Access. 2021. DOI: [10.1109/WD52248.2021.9508308](https://doi.org/10.1109/WD52248.2021.9508308). URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125403211&doi=10.1109%2fWD52248.2021.9508308&partnerID=40&md5=0262fa8921ba7b95fb6e0b72600d0791>.
- [280] F. Brinkmann et al. “Audibility and interpolation of head-above-torso orientation in binaural technology”. In: *IEEE Journal of Selected Topics in Signal Processing* PP.99 (2015), pp. 1–1. ISSN: 1932-4553. DOI: [10.1109/JSTSP.2015.2414905](https://doi.org/10.1109/JSTSP.2015.2414905).
- [281] Mark McGill et al. “Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, Apr. 2020, pp. 1–16. ISBN: 978-1-4503-6708-0. DOI: [10.1145/3313831.3376702](https://doi.org/10.1145/3313831.3376702). URL: <https://doi.org/10.1145/3313831.3376702> (visited on 05/06/2020).
- [282] Christoph Völker et al. “Modifications of the MULTi stimulus test with Hidden Reference and Anchor (MUSHRA) for use in audiology”. In: *International Journal of Audiology* 57.sup3 (Sept. 2016), S92–S104. DOI: [10.1080/14992027.2016.1220680](https://doi.org/10.1080/14992027.2016.1220680). URL: <https://doi.org/10.1080/14992027.2016.1220680>.
- [283] Fabian Brinkmann and Stefan Weinzierl. “Audio Quality Assessment for Virtual Reality”. en. In: *Sonic Interactions in Virtual Environments*. Ed. by Michele Geronazzo and Stefania Serafin. Cham: Springer International Publishing, Oct. 2022, pp. 145–178. ISBN: 978-3-031-04021-4. DOI: [10.1007/978-3-031-04021-4_5](https://doi.org/10.1007/978-3-031-04021-4_5). URL: https://doi.org/10.1007/978-3-031-04021-4_5 (visited on 09/19/2024).
- [284] Alexander Lindau. *Spatial Audio Quality Inventory (SAQI). Test Manual*. en. Tech. rep. 2014. DOI: [10.14279/DEPOSITONCE-1.2](https://doi.org/10.14279/DEPOSITONCE-1.2). URL: <https://depositonce.tu-berlin.de/handle/11303/157.2>.

- [285] Felix Stärz et al. "Comparison of binaural auralisations to a real loudspeaker in an audiovisual virtual classroom scenario: effect of room acoustic simulation, HRTF dataset, and head-mounted display on room acoustic perception". In: *Acta Acustica* 9 (2025), p. 31. ISSN: 2681-4617. DOI: [10.1051/aacus/2025012](https://doi.org/10.1051/aacus/2025012). URL: <http://dx.doi.org/10.1051/aacus/2025012>.
- [286] Elizabeth M. Wenzel et al. "Localization using Nonindividualized Head-Related Transfer Functions". In: *The Journal of the Acoustical Society of America* 94.1 (July 1993), pp. 111–123. DOI: [10.1121/1.407089](https://doi.org/10.1121/1.407089). URL: <https://doi.org/10.1121/1.407089>.
- [287] Henrik Møller et al. "Binaural Technique: Do we need Individual Recordings?" In: *Journal of the Audio Engineering Society* 44.6 (June 1996), pp. 451–469.
- [288] Maria Meier, Corinna S. Martarelli, and Wanja Wolff. "Is boredom a source of noise and/or a confound in behavioral science research?" In: *Humanities and Social Sciences Communications* 11.1 (Mar. 2024). ISSN: 2662-9992. DOI: [10.1057/s41599-024-02851-7](https://doi.org/10.1057/s41599-024-02851-7). URL: <http://dx.doi.org/10.1057/s41599-024-02851-7>.
- [289] Guillaume Andéol and Brian D. Simpson. "Editorial: How, and Why, Does Spatial-Hearing Ability Differ among Listeners? What is the Role of Learning and Multisensory Interactions?" In: *Frontiers in Neuroscience* 10 (Feb. 2016). ISSN: 1662-453X. DOI: [10.3389/fnins.2016.00036](https://doi.org/10.3389/fnins.2016.00036). URL: <http://journal.frontiersin.org/Article/10.3389/fnins.2016.00036/abstract> (visited on 12/08/2017).
- [290] Markus von Berg et al. "Assessing room acoustic listening expertise". In: *The Journal of the Acoustical Society of America* 150.4 (Oct. 2021). Publisher: Acoustical Society of America, pp. 2539–2548. ISSN: 0001-4966. DOI: [10.1121/10.0006574](https://doi.org/10.1121/10.0006574). URL: <https://asa.scitation.org/doi/10.1121/10.0006574> (visited on 10/13/2021).
- [291] Marc Brysbaert. "How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables". In: *Journal of Cognition* 2.1 (2019). ISSN: 2514-4820. DOI: [10.5334/joc.72](https://doi.org/10.5334/joc.72). URL: <http://dx.doi.org/10.5334/joc.72>.
- [292] Andrew T. Dilanchian, Ronald Andringa, and Walter R. Boot. "A Pilot Study Exploring Age Differences in Presence, Workload, and Cybersickness in the Experience of Immersive Virtual Reality Environments". In: *Frontiers in Virtual Reality* 2 (Oct. 2021). ISSN: 2673-4192. DOI: [10.3389/frvir.2021.736793](https://doi.org/10.3389/frvir.2021.736793). URL: <http://dx.doi.org/10.3389/frvir.2021.736793>.
- [293] Dorota Kamińska, Grzegorz Zwoliński, and Anna Laska-Leśniewicz. "Usability Testing of Virtual Reality Applications—The Pilot Study". In: *Sensors* 22.4 (Feb. 2022), p. 1342. ISSN: 1424-8220. DOI: [10.3390/s22041342](https://doi.org/10.3390/s22041342). URL: <http://dx.doi.org/10.3390/s22041342>.
- [294] Maximilian Landeck et al. "Journey through a virtual tunnel: Simulated motion and its effects on the experience of time". In: *Frontiers in Virtual Reality* 3 (Jan. 2023). ISSN: 2673-4192. DOI: [10.3389/frvir.2022.1059971](https://doi.org/10.3389/frvir.2022.1059971). URL: <http://dx.doi.org/10.3389/frvir.2022.1059971>.

- [295] Stefan Weinzierl, Steffen Lepa, and David Ackermann. "A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI)". en. In: *The Journal of the Acoustical Society of America* 144.3 (Sept. 2018), pp. 1245–1257. ISSN: 0001-4966. DOI: [10.1121/1.5051453](https://doi.org/10.1121/1.5051453). URL: <http://asa.scitation.org/doi/10.1121/1.5051453> (visited on 09/16/2018).
- [296] Olli Rummukainen et al. "Audio Quality Evaluation in Virtual Reality: Multiple Stimulus Ranking with Behavior Tracking". In: *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR)* (Redmond, WA). herre_papers_virtual_and_augmented_reality_rendering. Audio Engineering Society, 2018, pp. 50–59. ISBN: 9781510870390. URL: <https://www.scopus.com/record/display.uri?eid=2-s2.0-85056869420&login=inward>.
- [297] Minghui Liu et al. "A Literature Review of Indoor Wayfinding in Virtual Environment: Comparability to Real Environment and Ecological Validity". In: *ACM Symposium on Spatial User Interaction. SUI '24*. ACM, Oct. 2024, 1–2. DOI: [10.1145/3677386.3688892](https://doi.org/10.1145/3677386.3688892). URL: <http://dx.doi.org/10.1145/3677386.3688892>.
- [298] Alexis Paljic. "Ecological Validity of Virtual Reality: Three Use Cases". In: *New Trends in Image Analysis and Processing – ICIAP 2017*. Ed. by Sebastiano Battiato et al. Cham: Springer International Publishing, 2017, pp. 301–310. ISBN: 978-3-319-70742-6.
- [299] Christoph Völker et al. "Adaptions for the Multi Stimulus test with Hidden Reference and Anchor (MUSHRA) for elder and technical unexperienced participants". In: Mar. 2015.
- [300] Linda M Collins, John J Dziak, and Runze Li. "Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs." In: *Psychological methods* 14.3 (2009), p. 202.
- [301] Catarina Mendonça and Symeon Delikaris-Manias. "statistical tests with mushra data". In: *journal of the audio engineering society* 10006 (May 2018).
- [302] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. "Mixed Linear Regression with Multiple Components". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/8248a99e81e752cb9b41da3fc43fbe7f-Paper.pdf.
- [303] Joe H. Ward. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301 (Mar. 1963), 236–244. ISSN: 1537-274X. DOI: [10.1080/01621459.1963.10500845](https://doi.org/10.1080/01621459.1963.10500845). URL: <http://dx.doi.org/10.1080/01621459.1963.10500845>.
- [304] Ludwig Fahrmeir et al. *Regression*. en. 2013th ed. Berlin, Germany: Springer, May 2013.
- [305] Vijaya, Shweta Sharma, and Neha Batra. "Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering". In: *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. IEEE, Feb. 2019, 568–573. DOI: [10.1109/COMITCon.2019.8862232](https://doi.org/10.1109/COMITCon.2019.8862232). URL: <http://dx.doi.org/10.1109/COMITCon.2019.8862232>.

- [306] Michael D E Goodyear, Karmela Krleza-Jeric, and Trudo Lemmens. "The Declaration of Helsinki". In: *BMJ* 335.7621 (Sept. 2007), 624–625. ISSN: 1468-5833. DOI: [10.1136/bmj.39339.610000.BE](https://doi.org/10.1136/bmj.39339.610000.BE). URL: <http://dx.doi.org/10.1136/bmj.39339.610000.BE>.
- [307] W. G. Gardner and K. D. Martin. "HRTF Measurements of a KEMAR". In: *J. of the Acoustical Society of America* 97.6 (June 1995), pp. 3907–3908.
- [308] Constantin Popp and Damian T. Murphy. "Speech Intelligibility versus Congruency: User Preferences of the Acoustics of Virtual Reality Game Spaces". In: *Virtual Worlds* 3.1 (Jan. 2024), 40–61. ISSN: 2813-2084. DOI: [10.3390/virtualworlds3010003](https://doi.org/10.3390/virtualworlds3010003). URL: <http://dx.doi.org/10.3390/virtualworlds3010003>.
- [309] Pedro Lladó, Petteri Hyvärinen, and Ville Pulkki. "The impact of head-worn devices in an auditory-aided visual search task". In: *The Journal of the Acoustical Society of America* 155.4 (Apr. 2024), 2460–2469. ISSN: 0001-4966. DOI: [10.1121/10.0025542](https://doi.org/10.1121/10.0025542). URL: <http://dx.doi.org/10.1121/10.0025542>.
- [310] Sebastian A. Ausili et al. "Sound Localization with Hearables in Transparency Mode". In: *Audiology Research* 15.3 (Apr. 2025), p. 48. ISSN: 2039-4349. DOI: [10.3390/audiolres15030048](https://doi.org/10.3390/audiolres15030048). URL: <http://dx.doi.org/10.3390/audiolres15030048>.
- [311] Yoonas A. Sekhavat and Jeffrey Parsons. "The effect of tracking technique on the quality of user experience for augmented reality mobile navigation". In: *Multimedia Tools and Applications* 77.10 (May 2017), 11635–11668. ISSN: 1573-7721. DOI: [10.1007/s11042-017-4810-y](https://doi.org/10.1007/s11042-017-4810-y). URL: <http://dx.doi.org/10.1007/s11042-017-4810-y>.
- [312] Shivani Gupta and Atul Gupta. "Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review". In: *Procedia Computer Science* 161 (2019), 466–474. ISSN: 1877-0509. DOI: [10.1016/j.procs.2019.11.146](https://doi.org/10.1016/j.procs.2019.11.146). URL: <http://dx.doi.org/10.1016/j.procs.2019.11.146>.
- [313] Panagiotis Kourtesis et al. "Validation of the Virtual Reality Neuroscience Questionnaire: Maximum Duration of Immersive Virtual Reality Sessions Without the Presence of Pertinent Adverse Symptomatology". In: *Frontiers in Human Neuroscience* 13 (Nov. 2019). ISSN: 1662-5161. DOI: [10.3389/fnhum.2019.00417](https://doi.org/10.3389/fnhum.2019.00417). URL: <http://dx.doi.org/10.3389/fnhum.2019.00417>.
- [314] Bobby Bodenheimer et al. "Perceiving Absolute Distance in Augmented Reality Displays with Realistic and Non-realistic Shadows". In: *ACM Symposium on Applied Perception 2023. SAP '23*. ACM, Aug. 2023, 1–9. DOI: [10.1145/3605495.3605800](https://doi.org/10.1145/3605495.3605800). URL: <http://dx.doi.org/10.1145/3605495.3605800>.
- [315] Hendrik Himmelein et al. "Influence of mental effort on sound evaluations in virtual and real experimental environments". In: *Frontiers in Virtual Reality* 6 (Oct. 2025). ISSN: 2673-4192. DOI: [10.3389/frvir.2025.1672595](https://doi.org/10.3389/frvir.2025.1672595). URL: <http://dx.doi.org/10.3389/frvir.2025.1672595>.
- [316] Francis Rumsey. "Evaluating AVAR: Goodbye Quality, Hello Plausibility?". In: *Journal of the Audio Engineering Society* 66.12 (12 Dec. 2018), pp. 1126–1130.
- [317] Durand R Begault and Leonard J Trejo. *3-D sound for virtual reality and multimedia*. Sept. 2000.
- [318] Mendel Kleiner, Bengt-Inge Dalenbäck, and Peter Svensson. "Auralization – An overview". In: *Journal of the Audio Engineering Society* 41.11 (1993), pp. 861–875.

- [319] Benjamin Bernschütz. “A spherical far field HRIR/HRTF compilation of the Neumann KU 100”. In: *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. German Acoustical Society (DEGA) Berlin. 2013, p. 29.
- [320] Julius O. Smith. *Physical Audio Signal Processing*. W3K Publishing, 2010. ISBN: 978-0-9745607-2-4. URL: <https://ccrma.stanford.edu/~jos/pasp/>.
- [321] Riccardo Bona et al. “Automatic Parameters Tuning of Late Reverberation Algorithms for Audio Augmented Reality”. In: *Proceedings of the 17th International Audio Mostly Conference*. AM '22. St. Pölten, Austria: Association for Computing Machinery, 2022, pp. 36–43. ISBN: 9781450397018. DOI: [10.1145/3561212.3561236](https://doi.org/10.1145/3561212.3561236). URL: <https://doi.org/10.1145/3561212.3561236>.
- [322] Andrew D Brown, G Christopher Stecker, and Daniel J Tollin. “The precedence effect in sound localization”. In: *Journal of the Association for Research in Otolaryngology* 16 (2015), pp. 1–28. DOI: [10.1007/s10162-014-0496-2](https://doi.org/10.1007/s10162-014-0496-2). URL: <https://doi.org/10.1007/s10162-014-0496-2>.
- [323] Tapio Lokki and Jukka Pätynen. “Auditory Spatial Impression in Concert Halls”. In: *The Technology of Binaural Understanding*. Ed. by Jens Blauert and Jonas Braasch. Cham: Springer International Publishing, 2020, pp. 173–202. ISBN: 978-3-030-00386-9. DOI: [10.1007/978-3-030-00386-9_7](https://doi.org/10.1007/978-3-030-00386-9_7). URL: https://doi.org/10.1007/978-3-030-00386-9_7.
- [324] Arthur Harold Marshall and Michael Barron. “Spatial responsiveness in concert halls and the origins of spatial impression”. In: *Applied Acoustics* 62.2 (2001), pp. 91–108. ISSN: 0003-682X. DOI: [10.1016/S0003-682X\(00\)00050-5](https://doi.org/10.1016/S0003-682X(00)00050-5). URL: <https://www.sciencedirect.com/science/article/pii/S0003682X00000505>.
- [325] Isaac Engel et al. “Perceptual implications of different Ambisonics-based methods for binaural reverberation”. In: *The Journal of the Acoustical Society of America* 149.2 (2021), pp. 895–910. DOI: [10.1121/10.0003437](https://doi.org/10.1121/10.0003437). URL: <https://doi.org/10.1121/10.0003437>.
- [326] Alexander Lindau, Linda Kosanke, and Stefan Weinzierl. “Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses”. In: *Journal of the Audio Engineering Society* 60.11 (Nov. 2012), pp. 887–898.
- [327] Leo McCormack and Archontis Politis. “SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods”. In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, Mar. 2019.
- [328] Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. *Detection theory: A user’s guide*. Routledge, 2021.
- [329] Federico Avanzini. “Procedural Modeling of Interactive Sound Sources in Virtual Reality”. In: *Sonic Interactions in Virtual Environments*. Ed. by Michele Geronazzo and Stefania Serafin. Cham: Springer International Publishing, Oct. 2022, pp. 49–76. ISBN: 978-3-031-04021-4. DOI: [10.1007/978-3-031-04021-4_2](https://doi.org/10.1007/978-3-031-04021-4_2). URL: https://doi.org/10.1007/978-3-031-04021-4_2.

- [330] Charles E. Jack and Willard R. Thurlow. "Effects of Degree of Visual Association and Angle of Displacement on the "Ventriloquism" Effect". In: *Perceptual and Motor Skills* 37.3 (1973). PMID: 4764534, pp. 967–979. DOI: [10.1177/003151257303700360](https://doi.org/10.1177/003151257303700360). eprint:<https://doi.org/10.1177/003151257303700360>. URL: <https://doi.org/10.1177/003151257303700360>.
- [331] Christopher C. Berger et al. "Generic HRTFs May be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity". In: *Frontiers in Neuroscience* 12 (2018). ISSN: 1662-453X. DOI: [10.3389/fnins.2018.00021](https://doi.org/10.3389/fnins.2018.00021). URL: <https://www.frontiersin.org/articles/10.3389/fnins.2018.00021>.
- [332] Nikunj Raghuvanshi and John Snyder. "Parametric Directional Coding for Precomputed Sound Propagation". In: *ACM Trans. Graph.* 37.4 (July 2018). ISSN: 0730-0301. DOI: [10.1145/3197517.3201339](https://doi.org/10.1145/3197517.3201339). URL: <https://doi.org/10.1145/3197517.3201339>.
- [333] Michel Callon und Bruno Latour. "Unscrewing the big Leviathan: how actors macro-structure reality and how sociologists help them to do so". In: (1981). Ed. by K. Knorr-Cetina and A.V Cicourel, pp. 277–303. URL: <http://www.bruno-latour.fr/articles/article/09-LEVIATHAN-GB.pdf>.
- [334] Marian Dörk, Rob Comber, and Martyn Dade-Robertson. "Monadic exploration". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2014. DOI: [10.1145/2556288.2557083](https://doi.org/10.1145/2556288.2557083). URL: <https://doi.org/10.1145/2556288.2557083>.
- [335] Giorgio Franceschelli and Mirco Musolesi. *Reinforcement Learning for Generative AI: State of the Art, Opportunities and Open Research Challenges*. 2023. DOI: [10.48550/ARXIV.2308.00031](https://arxiv.org/abs/2308.00031). URL: <https://arxiv.org/abs/2308.00031>.
- [336] Edward T. Hall et al. "Proxemics [and Comments and Replies]". In: *Current Anthropology* 9.2/3 (Apr. 1968). Publisher: [University of Chicago Press, Wenner-Gren Foundation for Anthropological Research], pp. 83–108. ISSN: 0011-3204. DOI: [10.1086/200975](https://www.jstor.org/stable/2740724). URL: <https://www.jstor.org/stable/2740724> (visited on 11/17/2021).
- [337] Rose Connolly et al. "The Impact of Navigation on Proxemics in an Immersive Virtual Environment with Conversational Agents". In: *IEEE Transactions on Visualization and Computer Graphics* 31.5 (May 2025), 2787–2797. ISSN: 2160-9306. DOI: [10.1109/tvcg.2025.3550231](https://doi.org/10.1109/tvcg.2025.3550231). URL: <http://dx.doi.org/10.1109/TVCG.2025.3550231>.
- [338] Mitsuhiko Kimoto et al. "Effects of appearance and gender on pre-touch proxemics in virtual reality". In: *Frontiers in Psychology* 14 (June 2023). ISSN: 1664-1078. DOI: [10.3389/fpsyg.2023.1195059](https://doi.org/10.3389/fpsyg.2023.1195059). URL: <http://dx.doi.org/10.3389/fpsyg.2023.1195059>.
- [339] Jonathan Mumm and Bilge Mutlu. "Human-robot proxemics: physical and psychological distancing in human-robot interaction". In: *Proceedings of the 6th international conference on Human-robot interaction*. HRI'11. ACM, Mar. 2011, 331–338. DOI: [10.1145/1957656.1957786](https://doi.org/10.1145/1957656.1957786). URL: <http://dx.doi.org/10.1145/1957656.1957786>.
- [340] Tomáš Spurný et al. "Dynamic Proxemic Model for Human–Robot Interactions Using the Golden Ratio". In: *Applied Sciences* 15.15 (July 2025), p. 8130. ISSN: 2076-3417. DOI: [10.3390/app15158130](https://doi.org/10.3390/app15158130). URL: <http://dx.doi.org/10.3390/app15158130>.

- [341] Luca Longo et al. "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions". In: *Information Fusion* 106 (June 2024), p. 102301. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2024.102301](https://doi.org/10.1016/j.inffus.2024.102301). URL: <http://dx.doi.org/10.1016/j.inffus.2024.102301>.
- [342] B. J. Fogg. "Persuasive technology: using computers to change what we think and do". In: *Ubiquity* 2002.December (2002), 5:2. DOI: [10.1145/764008.763957](https://doi.org/10.1145/764008.763957). URL: <https://dl.acm.org/doi/10.1145/764008.763957> (visited on 09/04/2025).
- [343] Makuochi Nkwo and Rita Orji. "Persuasion in ecommerce: a comparative analysis of western and indigenous African ecommerce". In: *Proceedings of the Second African Conference for Human Computer Interaction: Thriving Communities*. 2018, pp. 1–5.
- [344] Makuochi Nkwo, Rita Orji, and John Ugah. "Persuasion for promoting clean and sustainable environment". In: *Proceedings of the Second African Conference for Human Computer Interaction: Thriving Communities*. 2018, pp. 1–5.
- [345] Rita Orji and Karyn Moffatt. "Persuasive technology for health and wellness: State-of-the-art and emerging trends". In: *Health informatics journal* 24.1 (2018), pp. 66–91.
- [346] Jo Vermeulen et al. "Crossing the bridge over Norman's Gulf of Execution: revealing feedforward's true identity". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. ACM, Apr. 2013, 1931–1940. DOI: [10.1145/2470654.2466255](https://doi.org/10.1145/2470654.2466255). URL: <http://dx.doi.org/10.1145/2470654.2466255>.
- [347] Don Norman. *The design of everyday things*. en. 2nd ed. London, England: Basic Books, Nov. 2013.
- [348] Jorge L. Armony and Joseph E. LeDoux. *Emotional responses to auditory stimuli*. Oxford University Press, Jan. 2010. DOI: [10.1093/oxfordhb/9780199233281.013.0019](https://doi.org/10.1093/oxfordhb/9780199233281.013.0019). URL: <http://dx.doi.org/10.1093/oxfordhb/9780199233281.013.0019>.
- [349] F Alton Everest and Ken Pohlmann. *Master handbook of acoustics*. 5th ed. McGraw-Hill Education Tab, May 2009.
- [350] Tomoko Shimoda, Rei Nishijima, and Sadako Yoshimura. "Examining Modifications of the Noise Generated by Nurses' Footsteps and the Use of Pass-By Objects: A Pilot Study". In: *Open Journal of Nursing* 10.02 (2020), 184–193. ISSN: 2162-5344. DOI: [10.4236/ojn.2020.102012](https://doi.org/10.4236/ojn.2020.102012). URL: <http://dx.doi.org/10.4236/ojn.2020.102012>.
- [351] Matthew Lombard and Theresa Ditton. "At the heart of it all: The concept of presence". In: *Journal of Computer-Mediated Communication* 3.2 (1997), JCMC321.
- [352] Bob G Witmer and Michael J Singer. "Measuring presence in virtual environments: A presence questionnaire". In: *Presence: Teleoperators & Virtual Environments* 7.3 (1998), pp. 225–240.
- [353] Jane Lessiter et al. "A cross-media presence questionnaire: The ITC-Sense of Presence Inventory". In: *Presence: Teleoperators & Virtual Environments* 10.3 (2001), pp. 282–297.
- [354] Wijnand IJsselstein, Yvonne de Kort, and Karolien Poels. *The Game Experience Questionnaire*. Retrieved from <https://pure.tue.nl/ws/portalfiles/portal/21666971/GXQ.pdf>. 2013.

- [355] Mohammed Ibrahim Alhojailan. "Thematic Analysis: a Critical Review of its Process and Evaluation". In: *West East Journal of Social Sciences* 1.1 (Dec. 2012), 39–47.
- [356] John C. Gower. "A general coefficient of similarity and some of its properties". In: *Biometrics* 27.4 (1971), pp. 857–871.
- [357] Robert L. Thorndike. "Who Belongs in the Family?" In: *Psychometrika* 18.4 (Dec. 1953), 267–276. ISSN: 1860-0980. DOI: [10.1007/bf02289263](https://doi.org/10.1007/bf02289263). URL: <http://dx.doi.org/10.1007/BF02289263>.
- [358] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), 53–65. ISSN: 0377-0427. DOI: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). URL: [http://dx.doi.org/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org/10.1016/0377-0427(87)90125-7).
- [359] Alessandro Giuseppe Privitera and Michele Geronazzo. "Designing sonic interactions in intelligent reality with egocentric audio technologies". In: *The Routledge Handbook of Sound Design*. Focal Press, Sept. 2024, 219–251. ISBN: 9781003325567. DOI: [10.4324/9781003325567-16](https://doi.org/10.4324/9781003325567-16). URL: <http://dx.doi.org/10.4324/9781003325567-16>.
- [360] Tanh Quang Tran et al. "Classifying Presence Scores: Insights and Analysis from Two Decades of the Igroup Presence Questionnaire (IPQ)". In: *ACM Trans. Comput.-Hum. Interact.* (2024). Just Accepted. ISSN: 1073-0516. DOI: [10.1145/3689046](https://doi.org/10.1145/3689046). URL: <https://dl.acm.org/doi/10.1145/3689046> (visited on 08/23/2024).
- [361] Tiina Lautamo et al. "Psychometric properties of the Assessment Tool for Perceived Agency (ATPA-22) – utility for the rehabilitation of young adults not in education, employment or training (NEETs)". In: *Scandinavian Journal of Occupational Therapy* 28.2 (Feb. 2021). Publisher: Taylor & Francis, pp. 97–109. ISSN: 1103-8128. DOI: [10.1080/11038128.2020.1782983](https://doi.org/10.1080/11038128.2020.1782983). URL: <https://www.tandfonline.com/doi/abs/10.1080/11038128.2020.1782983> (visited on 09/08/2025).
- [362] Claire Glanois et al. "A survey on interpretable reinforcement learning". en. In: *Machine Learning* 113.8 (Aug. 2024), pp. 5847–5890. ISSN: 1573-0565. DOI: [10.1007/s10994-024-06543-w](https://doi.org/10.1007/s10994-024-06543-w). URL: <https://doi.org/10.1007/s10994-024-06543-w> (visited on 09/09/2025).
- [363] Scott Fujimoto, Herke Hoof, and David Meger. "Addressing Function Approximation Error in Actor-Critic Methods". en. In: *Proceedings of the 35th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2018, pp. 1587–1596. URL: <https://proceedings.mlr.press/v80/fujimoto18a.html> (visited on 09/09/2025).
- [364] Sašo Džeroski, Luc De Raedt, and Kurt Driessens. "Relational Reinforcement Learning". en. In: *Machine Learning* 43.1 (Apr. 2001), pp. 7–52. ISSN: 1573-0565. DOI: [10.1023/A:1007694015589](https://doi.org/10.1023/A:1007694015589). URL: <https://doi.org/10.1023/A:1007694015589> (visited on 09/09/2025).
- [365] Andrei A. Rusu et al. "Policy Distillation". In: Nov. 2015. URL: <https://www.semanticscholar.org/paper/Policy-Distillation-Rusu-Colmenarejo/1c4927af526d5c28f7c2cfa492ece192d80a61d4> (visited on 09/09/2025).

- [366] W. Bradley Knox and Peter Stone. “Interactively shaping agents via human reinforcement: the TAMER framework”. In: *Proceedings of the fifth international conference on Knowledge capture*. K-CAP '09. New York, NY, USA: Association for Computing Machinery, Sept. 2009, pp. 9–16. ISBN: 978-1-60558-658-8. DOI: [10 . 1145 / 1597735 . 1597738](https://doi.org/10.1145/1597735.1597738). URL: [https : / / dl . acm . org / doi / 10 . 1145 / 1597735 . 1597738](https://dl.acm.org/doi/10.1145/1597735.1597738) (visited on 09/09/2025).