SoundShift: Exploring Sound Manipulations for Accessible Mixed-Reality Awareness

Ruei-Che Chang
rueiche@umich.edu
University of Michigan
Ann Arbor, MI, USA

Chia-Sheng Hung
yoyung0809@gmail.com
National Taiwan University
Taipei, Taiwan

Bing-Yu Chen
robin@ntu.edu.tw
National Taiwan University
Taipei, Taiwan

Dhruv Jain
profdj@umich.edu
University of Michigan
Ann Arbor, MI, USA

Anhong Guo
anhong@umich.edu
University of Michigan
Ann Arbor, MI, USA

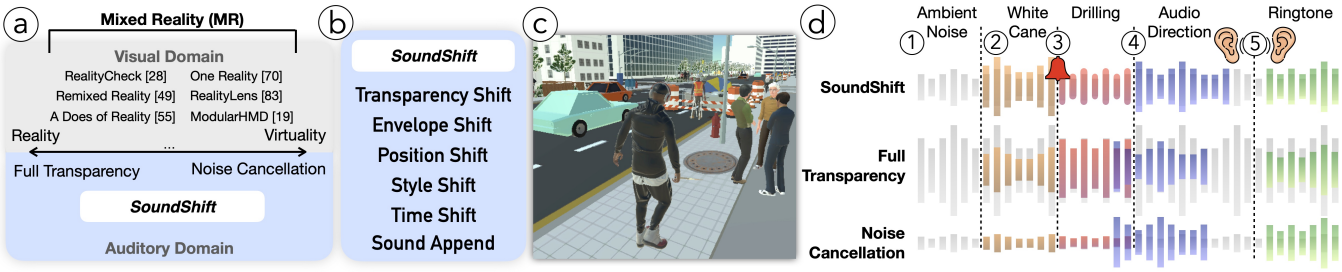


Figure 1: We present *SoundShift*, a concept to manipulate sounds to improve mixed-reality awareness. (a) SoundShift is situated in the auditory Reality-Virtuality Continuum with full transparency and noise cancellation as two ends, and comprises (b) six sound manipulators, which are Transparency Shift, Envelope Shift, Position Shift, Style Shift, Sound Append, and Time Shift. (c) In a scenario, a BVI user navigates a busy street with a white cane and audio directions. They sometimes may receive ringtones and pass by construction sites with drilling noises. (d.1) Transparency Shift makes the auditory transparency half to suppress nuanced noises while retaining real-world awareness. (d.2) Envelope Shift increases the white cane sounds to make them distinctive. (d.3) Sound Append plays an earcon to signal the danger, and Style Shift applies a low-pass filter to make drilling noise less sharp to hear. (d.4) Time Shift delays the audio directions when they conflict with drilling noises. (d.5) Position Shift places ringtone on the right and audio directions on the left to increase distinguishability.

ABSTRACT

Mixed-reality (MR) soundscapes blend real-world sound with virtual audio from hearing devices, presenting intricate auditory information that is hard to discern and differentiate. This is particularly challenging for blind or visually impaired individuals, who rely on sounds and descriptions in their everyday lives. To understand how complex audio information is consumed, we analyzed online forum posts within the blind community, identifying prevailing challenges, needs, and desired solutions. We synthesized the results and propose *SoundShift* for increasing MR sound awareness, which includes six sound manipulations: Transparency Shift, Envelope Shift, Position Shift, Style Shift, Time Shift, and Sound Append. To evaluate the effectiveness of SoundShift, we conducted a user study with 18 blind participants across three simulated MR scenarios, where participants identified specific sounds within intricate soundscapes. We found that SoundShift increased MR sound awareness and minimized cognitive load. Finally, we developed three real-world example applications to demonstrate the practicality of SoundShift.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design theory, concepts and paradigms**; **Mixed / augmented reality**; **Virtual reality**.

KEYWORDS

AR/VR, mixed reality, sound awareness, accessibility

ACM Reference Format:
Ruei-Che Chang, Chia-Sheng Hung, Bing-Yu Chen, Dhruv Jain, and Anhong Guo. 2024. SoundShift: Exploring Sound Manipulations for Accessible Mixed-Reality Awareness. In *Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3643834.3661556

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3661556

1 INTRODUCTION

Mixed reality (MR) is becoming more pervasive nowadays, where real-world (RW) and virtual-reality (VR) visual elements blend and interact with each other in real time, offering users seamless access to both. Visual descriptions [12, 29] or acoustic cues [35, 45, 63] could make visual MR more accessible to people who are blind or visually impaired (BVI) who rely heavily on sounds in their everyday lives. However, there is little discussion on the accessibility of auditory MR, where sounds come from different sources, from RW content to virtual audio presented by hearing devices, which may conflict with each other.

The conflict of sounds could cause confusion and high cognitive load to BVI people and potentially lead to missing crucial information. For instance, imagine walking down a busy street with ambient and crowd noises while receiving navigation instructions and participating in a virtual call. It might be difficult to shift one's focus across different audio applications. Such situations will be increasingly common, as MR applications become more pervasive and have a growing integration of virtual sounds in our lives such as virtual meetings, text-to-speech applications, broadcast, and music/entertainment. Furthermore, the necessity of having descriptions or acoustic cues for non-visually accessing MR would also impose another audio layer on BVI people, and the higher quality and fidelity of synthesized voice or virtual sounds may be hard to discern from RW counterparts. In this work, we aim to investigate the following question: *How to manipulate sounds to enhance sound awareness in a complex MR audio environment for BVI people?*

To understand the current practices in consuming complex audio information, we first conducted a content analysis from active online forums within the BVI community, where we collected posts and comments about the prevailing scenarios, challenges, and potential solutions for consuming complex audio information. We found that many scenarios require the consumption of RW[1] and VR[2] sounds, such as navigating a busy street with audio directions, playing an instrument following music tutorials, and consuming screen reader feedback and other audio applications. On the other hand, BVI people expressed their desires to manipulate sounds and proposed ad-hoc solutions, such as distributing sound sources to different devices to allow the consumption of multiple sounds in parallel, adjusting sound characteristics to make sounds distinctive, or customizing existing sound libraries of applications. We synthesized our findings and proposed *SoundShift* to enhance users' perception and awareness of sounds in mixed reality environments. This approach incorporates six sound manipulation techniques: Transparency Shift, Envelope Shift, Position Shift, Style Shift, Time Shift, and Sound Append.

To understand how SoundShift manipulations can affect the perception of MR sounds, we conducted a user study with eighteen BVI participants who experienced the three simulated scenarios and identified sounds, including a *RW-Focused* scenario when navigating a busy street, a *VR-Focused* scenario when focusing on audiobook, and a *Fully-Mixed* scenario when attending a hybrid conference. In each scenario, we applied pre-defined sound manipulations on real-world and virtual sounds to enhance the perception of the mixed-reality soundscape. We compared SoundShift with the other two ends of the auditory Reality-Virtuality continuum [43, 56, 57]: full acoustic transparency that enhances the presence of the real world, and noise cancellation that enhances the immersion of the virtual reality (Figure 1a).

Our results showed that sound manipulations significantly improved BVI people's ability to perceive and manage sound information compared to full transparency and noise cancellation in our three simulated scenarios. The conditions also had varying effects on participants' performance across the three scenarios. SoundShift also significantly reduced participants' cognitive load in perceiving and managing sounds compared to full transparency and noise cancellation. Additionally, participants shared ideas for how they would further customize the sound manipulations in each scenario. Our evaluations demonstrate that SoundShift effectively improved MR sound awareness for BVI people.

To demonstrate the generalizability and practicality of SoundShift, we further developed three real-world example applications based on our content analysis on online posts and user customization mentioned in our study: *(i)* an audio-adaptive online meeting web application that addresses the conflict between screen reader sounds and meeting conversations, *(ii)* a mixed-reality content-aware image exploration application that provides stylized and spatialized audio feedback based on real-world and virtual content, and *(iii)* a mobile navigation application that analyzes real-world and virtual sound events to identify opportune moments for delivering audio directions.

In summary, our work contributes:

- The concept of SoundShift to make MR sound awareness accessible for BVI people, through six sound manipulators derived from our content analysis on BVI forums.
- An instantiation of the six sound manipulators and three simulated scenarios across the Reality-Virtuality continuum in Unity.
- Results from a user study with eighteen BVI people to provide empirical evidence of significantly enhanced sound awareness through SoundShift manipulations.
- Three real-world example applications to demonstrate the generalizability and practicality of SoundShift.

2 RELATED WORK

Our work was motivated and situated from prior work on mixed reality, MR accessibility, and soundscape personalization.

2.1 Blending Real and Virtual World

Several prior works have explored blending Reality and Virtuality [57] (Figure 1a) to achieve novel interactions by leveraging otherwise unavailable benefits from the other.

In Augmented Virtuality (AV), which blends RW elements into virtual experiences, A Dose of Reality [55] highlighted the benefits of incorporating RW components, like using a visible physical keyboard in VR to improve typing. RealityLens [83] introduced techniques for users to personalize integrating RW visual regions into VR. ModularHMD [19] promoted RW awareness via peripheral views through HMD configurations. RealityCheck [28] delved into

[1]RW sounds refer to sounds from RW environments, such as crowd, speaker, television.
[2]VR sounds refer to sounds rendered through users' hearing devices.

visual blending methods for fusing physical and virtual details. Further research has also pinpointed key aspects for fluidly embedding RW digital data (e.g., smartphone alerts) into VR, examining optimal intervention times [14], notification display methods [22, 85], and proper positions for notifications [31].

As for Augmented Reality (AR) [57], One Reality [70] provided a conceptual framework that embodied the incremental levels of mixed-reality interactions from physical to virtual worlds. Remixed Reality [49] characterized the manipulations of time and space that allow users to experience live 3D reconstruction as a novel form of MR. Given the growing works exploring blended interactions, VRception [24] was thus developed as a toolkit to facilitate the rapid prototyping of MR interactions.

Beyond visual blending, tangible real-world objects serve as tactile proxies for objects in VR, which create enhanced haptic experiences [17, 30, 73]. While numerous prior works have investigated blending RW and VR in visual and haptic domains, (un)blending sounds in MR is still under-explored, which is essential to creating accessible awareness for BVI people.

2.2 Soundscape Formation and Personalization

Augmented Audio Reality (AAR) integrates virtual sounds into RW soundscapes and is prevalent in our everyday lives [40]. Larsson et al. [43] posited that AR or AV in the concept of the Reality-Virtuality continuum can also adapt to auditory domains. McGill et al. [56] found that acoustically transparent headphones increase the sense of presence of reality while noise-canceling headphones diminish it. Advancements in noise-canceling technologies (also known as soundscaping technologies [26]), have transformed the headphone user experience and emerged as a significant research field [36]. However, Haas et al. [25] described the limitations of soundscaping technologies: *"current personal audio technology is not designed in a way that it allows users to handle their social context satisfactorily. Furthermore, information acquisition is made more difficult and users often have to make a choice between the surrounding acoustic environment and their own content and media."* Though prior research [9, 16, 54, 72] or the current commercial headphones support adapting the audio based on users' environment, users still expressed desires to steer and personalize their soundscape in a fine-grained manner [25] (e.g., blocking certain unwanted sounds, masking environmental sounds). This is particularly crucial for BVI individuals, who may have different vision levels and reliance on auditory cues, resulting in varying needs for curating soundscapes in different contexts [71]. Recently, advancements in AI within the auditory domain have enhanced AAR and opened avenues for tailoring soundscaping technologies to users' needs. For example, several works have facilitated intelligent sound extraction [13, 80, 81], enabling users to selectively filter or emphasize specific sounds. Furthermore, sounds can be intelligently adapted to the user's activities, offering seamless experiences. This includes altering music based on driving contexts to improve in-car music experiences [37], or integrating audio notifications with ongoing music to reduce disruption and annoyance [6].

Inspired by these needs and trends in personalizing soundscapes, we explore more fine-grained sound manipulations tailored specifically for BVI individuals, who, as experts in audio technologies, may have a unique reliance and strategies on manipulating sounds. We seek to understand how these fine-grained manipulations could enhance sound awareness for BVI people in various MR scenarios, and how BVI people would perceive and customize these sound manipulations.

2.3 Accessibility of Mixed Reality

In recent years, the World Wide Web Consortium (W3C) has established guidelines for MR accessibility [82] to encourage MR developers to consider the diverse needs of people with different abilities as a primary concern rather than an "afterthought" once the technology has matured [58]. Efforts to enhance MR for people with disabilities include visual [48] or haptic [33] alternatives for people who are deaf or hard of hearing (DHH), and simplified interaction techniques for people with motor impairments [41, 59].

BVI people, on the other hand, often face challenges in fully experiencing MR. Tools like SeeingVR [87] have been developed to enhance visual awareness for people with visual impairments and ensure accessibility to VR content, and Herskovitz et al. [29] concluded a design space for AR tasks and made them accessible via corresponding verbal feedback. Furthermore, various hardware devices have been developed to enhance the haptic experiences, such as cane simulations in VR for accessible navigation and a better sense of immersion [18, 44, 61, 74, 86]. As for audio domains in MR scenarios such as navigation, acoustic maps [34, 62, 64] or spatial audio [4, 5, 34, 62, 64, 79] convey the area information to BVI individuals in an acoustic form that helps them construct mental maps in VR. Also, VRBubble [35] employed sound representations to bolster BVI peripheral awareness during social interactions, and OmniScribe [12] made the immersion of 360° videos accessible to BVI people by rendering traditional audio descriptions spatially based on the orientation of BVI users.

From the trend that prior works proposed different auditory solutions to address certain accessibility challenges, it is expected that an accessible mixed reality for BVI people would entail much more complex audio information. Our work, therefore, aims to help increase BVI people's awareness of complex MR sounds by exploring effective sound manipulations.

3 UNDERSTANDING PRACTICES OF CONSUMING COMPLEX SOUNDS

In this work, we aim to investigate: *How to manipulate sounds to enhance sound awareness in a complex MR audio environment for BVI people?* To answer this question, we first need to understand the current challenges, needs, and practices that BVI people consume complex audio information.

3.1 Method

We conducted a content analysis from active online forums within the BVI community. First, two researchers communicated synchronously over Zoom and reviewed posts and their comments in online forums, including AppleVis [2], and the Blind and Visually Impaired Community on Reddit [3]. We started by filtering the posts by keywords such as "sound", "audio", "headphone", "challenge", and "scenario." We reviewed posts from 2023 backward until we had collected over 1000 posts, ultimately reaching back to 2021.

Then, we eliminated off-topic posts, such as social activity recruitment, debugging devices, and casual conversations. Ultimately, we collected 100 posts and comments highly associated with the difficulties or needs of sound consumption. Then, we used thematic analysis [15] to analyze the data. Together, we reviewed and coded the posts and comments in an online spreadsheet and had discussions to reach an agreement on the major themes described below.

3.2 Findings

We organized our findings by reporting the scenarios of consuming complex audio information, the challenges, and potential solutions mentioned in the posts. In the below sections, RW sounds are those sourced from real-world objects, whereas VR sounds are auditory outputs generated through hearing devices.

3.2.1 Everyday Scenarios to Consume Complex Audio Information. RW or VR environment can create a complex ever-changing soundscape for BVI people. For instance, one stated their daily walk journey: *"My normal daily 6 mile walk typically takes me through a range of environments - from busy roads with high levels of traffic noise, along quieter residential roads, [...] to playing parks."* Examples of complex VR soundscapes include when the screen reader overlaps with screen navigation feedback: *"Is there a way to turn off the navigation sounds when using voiceover? I can customize these on my iPhone, so when I swipe I don't always get that clunking sound when encountering each item as I swipe"*, or when using a desktop interface while engaging in a virtual meeting: *"When I am in a meeting, with a braille display and my watch, I do NOT want to keep hearing the click-click sounds VoiceOver makes when flicking right or left."* Several scenarios also required simultaneous consumption of RW and VR sounds, such as audio directions while navigating the street, screen reader feedback in a noisy environment, or playing instruments with music on the hearing device.

3.2.2 Retaining Real-World Awareness in Noisy Environments. As mentioned, consuming both RW and VR sounds is common and inevitably creates several challenges for BVI people. RW noisy environments can overwhelm BVI people due to the difficulty in discerning important sounds out of noises, as one consumed Siri's sound effects: *"[...] they are difficult to hear in noisy environments even if you don't have a hearing problem."* Although noise-cancellation headphones can block out noises, BVI people still want to retain awareness of their surroundings: *"I want to be able to hear clearly what is around me at the same time I'm listening through them."* From the online forums, bone-conduction headphones are frequently noted as a potential alternative by BVI people; however, there is a risk that VR sounds may be overshadowed by their RW counterparts. Instead, one suggested dynamically adjusting virtual sound volume in accordance with the prevailing levels of RW sounds: *"You would always be aware of ambient sound while you were wearing them, and I liked the idea that volume could be set to adjust automatically depending on the noise around you."* This idea echoes prior research on developing audio-adaptive systems for contextualized interactions [16, 54, 88].

3.2.3 Adjusting Sound Characteristics of Important Sounds in Conflict with Each Other. Aside from conflicting with ambient noises, important sounds may conflict with each other and cause distraction or interference. For instance, one comment said: *"When we are listening to music or watching shows we keep having what is going to play up-next, which sometimes cuts off the start of a track in the case of music. [...] I know it flashes up for sighted users but I feel for voiceover users it is an irritant."* Several posts also inquired about the possibility of selective turning the screen reader on/off for specific applications, or dynamically adjusting sound characteristics to make the screen reader distinctive. *"They are difficult to hear in noisy environments [...] The problem here is that the sounds don't cover a wide enough range in the audio spectrum. There should be a variety of frequencies from low frequency to high frequency to make sure they can be heard in any environment."*

3.2.4 Distributing Sound Sources for Better Distinction. Besides adjusting sound characteristics, we also observed BVI people's strategy to distribute sound sources to different places for better perception of multiple audio streams. For example, one suggested routing audio streams to different devices: *"Ability to route music to my smart speakers while keeping VoiceOver on my phone. It's a deal breaker when having a party and you get to hear VoiceOver on your surround sound when grooving to some music."* Moreover, distributing sounds in different ears could be another solution: *"Frequently, I will just use one of the Beats Flex earpieces so that I limit the sound from my phone to one ear. For example, if traffic was on my right side, I might only use the left earpiece to hear directions from my phone."* These findings reveal the promise of spatial audio for conveying information from many sources in MR scenarios beyond conveying directional information in previous works [12, 34, 45, 62, 64] or commercial apps [4, 5].

3.2.5 Customizing or Augmenting Existing Sound Library. To avoid undesired audio presentations, we also found users' desires to customize the existing sound library, as one said: *"It would just be nice to have maybe a different ring or two instead of either the old phone, the car horn which I find obnoxious"* or configuring the screen reader: *"It should be possible to configure VoiceOver to play a special sound to indicate a control type such as a button or play two distinct sounds to quickly indicate if a checkbox is checked or unchecked."* Also, to avoid overwhelming textual information of the screen reader, BVI people desired a proper amount of audio information, where earcons could play a vital role: *"When any notification comes in, VoiceOver announces that there is a notification, then reads the time, and then eventually goes silent ... This seems like the wrong behavior, giving redundant information and is pretty annoying. "* Besides, BVI people imagined customizing the voice of screen readers to their close ones: *"I do love the idea of someone being able to save their voice ... Leads me to wonder what might happen to that voice when the person passes. Would we want to hear our own words spoken in a departed loved one's voice?"* These findings reveal users' needs of customizing sounds for different granularity of audio information.

3.3 Summary

Our findings revealed different scenarios where BVI people encountered complex audio information. These scenarios had different focuses spanning across the real world (e.g., traffic), virtual sounds (e.g., screen reader), or a combination of both (e.g., listening to

music from the hearing device and playing instruments). Furthermore, BVI people proposed their desires and solutions to better consume the complex audio information, including manipulating ambient noises while retaining real-world awareness, adjusting sound characteristics to make important sounds distinctive, distributing sound sources in different devices or locations, and having customizations on earcons or voice feedback in devices. These findings inspired the three mixed-reality scenarios in our study and SoundShift manipulations described in the next section.

4 SOUNDSHIFT MANIPULATIONS FOR ACCESSIBLE MIXED REALITY AWARENESS

SoundShift is a concept to enhance mixed-reality sound awareness, which includes six sound manipulations: Transparency Shift, Envelope Shift, Sound Append, Time Shift, Position Shift, and Style Shift. We hypothesize that SoundShift manipulations can enhance the awareness of sounds for BVI people in mixed-reality environments. In this section, we describe how content analysis results and prior works inspired each sound manipulator.

Transparency Shift *modulates the ambient sounds or noises to shift presence between RW and VR by varying acoustic transparency.* Acoustic transparency is a common description for headphones that blend virtual audio with real-world sounds [56]. Real-world noises can enhance the sense of real-world grounding and presence, as so-called "primitive hearing" [69]. An increased sense of presence was also found when a user wears acoustically transparent headphones than the noise-cancellation ones [56]. This creates opportunities to balance the presence and awareness between RW and VR through the degree of acoustic transparency [23, 43, 60]. The concept is similar to Apple Adaptive Audio [1] by interpolating different auditory transparency and noise cancellation (Figure 1d.1).

Envelope Shift *modifies the dynamic aspects of sounds, such as volume or pitch, affecting their perceived loudness and tonal characteristics over time.* Fundamental characteristics of sounds, such as pitch, volume, and duration, impact sensation and perception in daily listening [8, 21]. Modifying these characteristics, as discussed in section 3.2.3, helps make sounds distinguishable. As shown in Figure 1d.2, Envelope Shift increases the volume of specific sound sources over ambient noises.

Position Shift *controls the locations of sound sources to enhance the sense of immersion.* Spatial audio has been a long-standing research field in VR and an essential component to embody the sense of presence and immersion [7, 11, 51]. In sections 3.2.4, BVI people also distributed sound sources for selective attention, facilitating awareness through different directions. As shown in Figure 1d.5, Position Shift places sound sources at different spatial locations.

Style Shift *transforms the timbre and fidelity of sounds through various filters to modify their aesthetic and emotional impact.* Section 3.2.5 highlights BVI individuals' suggestions for screen reader voice customization beyond the standard synthesis. This includes diverse sound filters (e.g., high, low pass, human, robotic, anime) to cater to varied preferences. As shown in Figure 1d.3, Style Shift uses a low pass filter to soften sharp drilling noises.

Time Shift *adjusts the timing of sounds to prioritize or deprioritize them within a soundscape, controlling auditory focus.* Overlapping sounds are common (sections 3.2.1, 3.2.2, 3.2.3). Time Shift controls sound timing to prevent overlap, such as pausing or delaying virtual audio in MR when it conflicts with RW sounds. As shown in Figure 1d.4, Time Shift delays audio directions during drilling noises.

Sound Append *appends earcons for corresponding sound events.* Earcons [10] signal object updates and deliver key audio information with minimal user attention. In complex MR environments, earcons can signal RW and VR events to reduce cognitive load instead of describing everything using speech. As shown in Figure 1d.3, an earcon indicates potential danger.

5 USER STUDY

Our goals of the user study are to *(i)* explore the effect of the proposed sound manipulations, and to *(ii)* explore user preferences and customizations on sounds in different MR scenarios. We created simulated environments in Unity to control the playback of sounds and their characteristics for simulating the proposed sound manipulations. This method aims to achieve high-fidelity simulations to immerse BVI people in the simulated environments, foster their engagement in the tasks, and help them imagine the future of sound technologies for providing feedback. This also enabled us to iterate and refine our concepts and user requirements based on their experiences and feedback before committing resources to develop fully functional systems. Our method was inspired by how Wizard-of-Oz methods were used in prototyping novel interactions [38, 47], and how user enactment was used to elicit feedback by providing future usage scenarios [65–67].

Prior research indicated that full acoustic transparency (FT) increases RW presence and awareness, while noise cancellation (NC) enhances VR presence and awareness [56] (Figure 1a). We, therefore, posit them as the two ends of auditory Reality and Virtuality Continuum [43] for optimally augmenting sound awareness in their respective realms. Consequently, our study compared SoundShift manipulations with these two established conditions, hypothesizing that participants can achieve the best performance with SS in mixed-reality settings by combining the best of both worlds. Specifically, we aim to understand the following research questions in this study:

- *RQ1:* How do sound manipulations affect participants' performance compared to full transparency and noise cancellation?
- *RQ2:* How do the different scenarios with varying emphases on reality and virtuality affect participants' performance?
- *RQ3:* How do the conditions affect participants' performance differently across the scenarios?
- *RQ4:* How do sound manipulations affect participants' cognitive load compared to full transparency and noise cancellation?
- *RQ5:* How do participants describe their experiences and ways to further customize their soundscape for each scenario?

5.1 Participants

Through word-of-mouth and public recruitment posts, we recruited 18 BVI participants (10 M and 8 F), aged 20 to 41 (mean=29.0), with a deep experience of using sounds in their lives. All had professional Orientation and Mobility (O&M) training and various sound-related experiences in other domains, like playing instruments, being voice actors, participating in orchestras, and broadcasting. Twelve were

blind since birth, while six lost their vision later in life (Table 1). We refer to our BVI participants as B1-B18 in the following sections.

5.2 Simulated Scenarios and Sound Manipulations

To fully explore the space of MR, we present three scenarios with different focuses on sound awareness: *RW-Focused*, *VR-Focused*, and *Fully-Mixed*. We simulated the three scenarios in Unity and used its built-in audio functions to prototype the six manipulators. In each scenario, there were four types of sounds, and participants were asked to complete sound identification tasks by pressing down specific keys on the keyboard upon hearing corresponding sounds. Participants were instructed to engage in the scenarios by prioritizing certain types of sounds relevant to the scenario's objectives, such as concentrating on RW sounds in the *RW-Focused* scenario, or both RW and VR sounds in the *Fully-Mixed* scenario.

5.2.1 *RW-Focused Scenario: Navigating on the street with a white cane and voice navigation guidance.*

Purpose: We aimed to explore, in the RW-focused setting, whether sound manipulations can harmonize both RW and VR sounds to maintain user awareness, while the user is primarily focusing on real-world sounds.

Motivation: From our content analysis, navigating the road is an everyday routine for BVI people (Section 3.2.1), while they sometimes receive audio directions from navigation apps (Section 3.2.1 and 3.2.4). Thus, we simulated this scenario where the user focuses on the road conditions with occasional virtual audio presented, as detailed below.

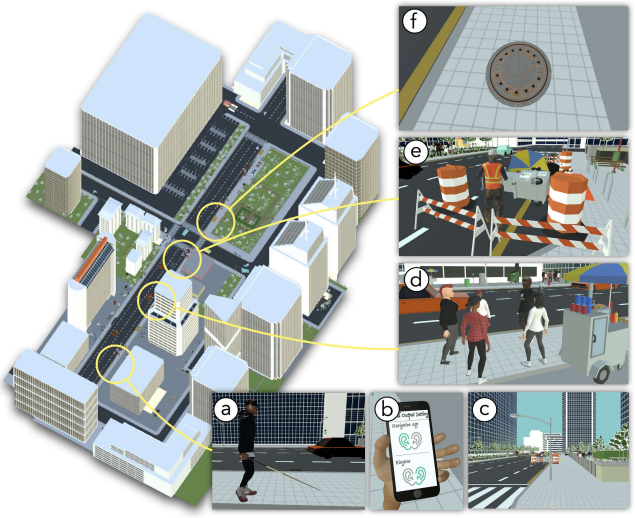


Figure 2: Simulated *RW-Focused* Scenario. (a) The user's avatar in Unity wears a headphone and holds a white cane. (b) The user sets the sound output by placing audio directions on the left, ringtone on the right, and (c) navigates on the street. (d) Several crowds, vendors, and passing cars along the street generate ambient noises, as well as (e) construction sites with drilling noises. (f) While walking, the user might come across random manholes, causing the white cane's sound to change upon contact.

Scenario: This scenario simulates a user navigating with a white cane on a busy street full of crowd noises while using the smartphone navigation app to receive instructions (Figure 2), such as "turn left at the next intersection", and receiving occasional phone ringtones during navigation. In this scenario, the user periodically taps a white cane on the ground to detect surface changes, like manholes, which alter the cane's sound. They must also be alert to drilling noises from nearby construction sites, focusing mainly on real-world sounds for safety.

Sound events and user required actions: For ambient noises, there are crowd noises randomly placed along the street, with occasional car noises to the user's left. Participants were asked to press keys 1, 2, 3, and 4 to identify the white cane on a manhole, drilling, navigation, and ringtone, respectively. The cane tapping sound occurs every 0.5 seconds, changing upon manhole contact for the next four taps. Each sound type has five random instances in our simulation, totaling 20 sound events to be identified.

Sound manipulations:

- Transparency Shift: applies half acoustic transparency (Figure 1d.1).
- Envelope Shift: prioritizes the four sound types by volume: white cane on the manhole, drilling, navigation instructions, and ringtone (Figure 1d.2).
- Position Shift: places navigation instructions on the left and ringtones on the right (Figure 1d.5).
- Style Shift: applies low pass filters to drilling noises to be audibly comfortable (Figure 1d.3).
- Time Shift: delays virtual sounds until after the RW sounds end (Figure 1d.4).
- Sound Append: appends a short earcon when getting close to a construction site (Figure 1d.3).

5.2.2 *VR-Focused Scenario: Consuming an audio handbook while working at the help desk.*

Purpose: We aimed to explore, in the VR-focused setting, whether sound manipulations can harmonize both RW and VR sounds to maintain user awareness, while the user primarily focuses on the virtual audio.

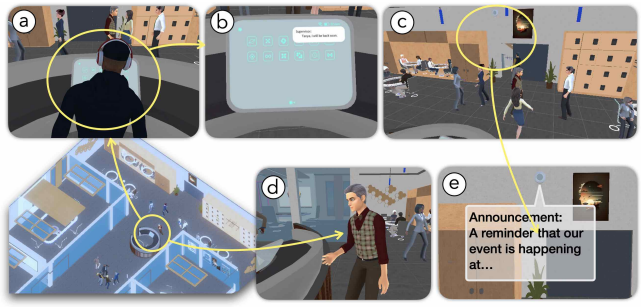


Figure 3: Simulated *VR-Focused* Scenario. (a) The user's avatar in Unity wears headphones, sits and works at the help desk, listens to an audio handbook, and (b) occasionally receives voice notes from a supervisor. (c) In the environment, there are background noises when people walk around, talk to each other, and open/close the sliding door. (d) People sometimes knock on the desk to get the user's attention. (e) The speaker plays occasional public announcements on the front wall.

Table 1: Participant demographics information. O&M refers to Orientation and Mobility.

ID	Age	Gender	Self-Reported Visual Ability	Hearing Experience
B1	20	Female	Blind, since birth. Light perception.	O&M, Orchestra, Instrument
B2	31	Female	Blind, since birth. Light perception.	O&M, Instrument
B3	41	Female	Blind, since birth. Light perception.	O&M, Instrument, Voice Actor, Broadcast
B4	22	Female	Blind, later in life. Light perception.	O&M, Instrument
B5	33	Male	Blind, since birth	O&M, Broadcast
B6	24	Female	Blind, since birth	O&M, Instrument
B7	22	Male	Blind, since birth. Light perception.	O&M, Instrument
B8	32	Male	Blind, later in life	O&M, Instrument
B9	20	Male	Blind, since birth	O&M, Instrument
B10	24	Male	Blind, later in life	O&M, Orchestra, Instrument
B11	33	Female	Blind, later in life	O&M, Instrument
B12	23	Male	Blind, since birth	O&M, Instrument
B13	35	Male	Blind, since birth	O&M, Instrument
B14	31	Male	Blind, since birth	O&M, Instrument
B15	38	Male	Blind, since birth	O&M
B16	29	Female	Blind, since birth. Light perception.	O&M, Instrument, Making audio descriptions
B17	21	Female	Blind, later in life. Light perception.	O&M, Instrument
B18	33	Male	Blind, later in life. Light perception.	O&M

Motivation: From our content analysis, there are examples of focusing on virtual audio in a noisy environment (e.g., screen readers in section 3.2.5). Thus, we simulated this scenario where the user focuses on virtual tasks with occasional RW sounds presented, as detailed below.

Scenario: This scenario depicts a user at a help desk, learning from an audio handbook as a new employee and responding to occasional knocks for attention (Figure 3). They also hear occasional voice notes from a supervisor through a hearing device and public announcements from a front-wall speaker. The user primarily focuses on the audio handbook and the supervisor's voice notes.

Sound events and user required actions: In this environment, ambient noises include random crowd sounds like chatting and footsteps, and occasional sounds of people opening and closing sliding doors. Participants were asked to press keys 1 to 4 to identify new sentences in the audio handbook, supervisor's voice notes, knocking, and public announcements, respectively. The audio handbook has a one-second pause between sentences. Each of the four sound types occurs five times randomly during the simulation, totaling 20 sound events to be identified.

Sound manipulations:

- Transparency Shift: applies full noise cancellation by default and dynamically switches to half acoustic transparency during public announcements.
- Envelope Shift: prioritizes the four sound types by volume: knocking, public announcements, audio handbook, and voice notes.
- Position Shift: places audio handbook on the left and voice notes on the right.
- Time Shift: delays virtual sounds until after the RW sounds end.

5.2.3 *Fully-Mixed Scenario: Attending a hybrid conference.*

Purpose: This setting comprises interwoven RW and VR events. We aimed to explore whether sound manipulations can harmonize both RW and VR sounds to maintain user awareness of both worlds.

Motivation: Our content analysis revealed several instances of simultaneous important RW and VR sounds (section 3.2.1). Also, given the prevalence of hybrid meetings in recent years and the experiences of BVI individuals in virtual settings (Section 3.2.1), we simulated this near-future scenario featuring speakers and occasional sound events from both realities.

Scenario: This scenario envisions a user at a 2050 hybrid conference with attendees participating in-person or through virtual avatars (Figure 4), whose voices are spatially rendered based on their 3D locations. The user hears voices from both RW and VR,

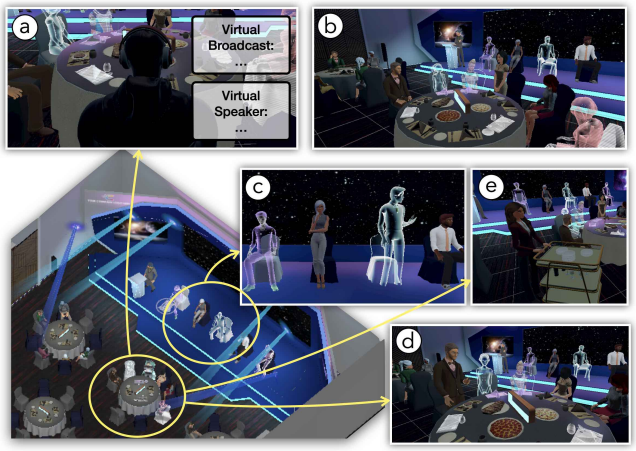


Figure 4: Simulated *Fully-Mixed* Scenario. (a) The user's avatar sits at the dining table and wears headphones to consume the voice of virtual speakers and virtual broadcasts. (b) There are physical and remote virtual speakers on the front stage, (c) where they may stand and speak up at the same time in a panel discussion, (d) similar to the physical and virtual attendees around the table. (e) Waitstaff sometimes comes to clean the table, generating dish clinking sounds.

needing to discern real from virtual speakers for interaction. Additionally, the sound of waitstaff cleaning up, like dish-clinking, demands attention, as the user may need to make room to allow the waitstaff to pass or clear the table. Virtual broadcasts occasionally announce events or notifications, requiring users to check their hearing devices. Here, the user equally pays attention to both RW and VR events.

Sound events and user required actions: In this setting, ambient noises include crowds at different tables. There are six real-world and virtual panelists on the stage, and six attendees around the user's table. Virtual speakers' voices mimic an old-school phone style. To add complexity, each speaker overlaps with another during the panel discussion. Participants were asked to press keys 1 to 4 to identify an RW person speaking, a VR person speaking, dish-clinking, and virtual broadcasts, respectively. There are five instances each for table cleaning and virtual broadcasts. Combined with the twelve individuals speaking, there are 22 sound events in total to be identified. This scenario, along with the other two, is designed to last approximately 90 seconds, ensuring a balanced experience across all three scenarios.

Sound manipulations:

- Transparency Shift: applies half acoustic transparency.
- Envelope Shift: prioritizes the four sound types using volume in the order of virtual/real people's voices, table cleaning, and virtual broadcasts.
- Position Shift: places virtual broadcasts on the right.
- Style Shift: applies a heavier old-school telephone effect to virtual voices.
- Time Shift: delays virtual sounds until after the RW sounds end.
- Sound Append: appends two separate earcons for table cleaning and virtual broadcasts.

5.3 Technical Details on the Unity Implementation

In each trial, the sounds were randomly scheduled and placed in different 3D locations to avoid learning effects on sound profiles across trials. In Unity, RW sounds were spatialized based on their 3D locations. VR sounds were also able to be rendered spatially but mainly on either the left or right ear, or both in our study (Section 5.2), using commodity hearing devices. For Time Shift, the known sound schedules in Unity allowed us to adjust sound characteristics (e.g., volume, pitch, and duration) to differentiate overlapping sounds or reschedule them to avoid conflict.

For Transparency Shift, objects producing sound were labeled as either RW or VR, with noise cancellation effects applied only to RW objects. We also used a headphone-blocking volume level η to simulate the volume level reduced when noise-canceling headphones block the ears. By adjusting the level of transparency τ and the headphone-blocking volume level η, the volume of RW sound sources S_{new} was determined by:

$$\begin{aligned} &S_{\text{new}} = S_{\text{default}} - ((1-\tau) \cdot S_{\text{default}} \cdot \eta) \\ &\tau \in [0,1], \eta \in [0,1] \end{aligned} \tag{1}$$

where we set the default volume of sounds $S_{\text{default}} = 0.5$ to allow room for further manipulations (e.g., increasing volume by

Figure 5: In our study, participants wore headphones, engaged in pre-defined scenarios, and pressed specific keys upon hearing corresponding sounds.

Envelope Shift), as in Unity the volume ranges from 0 to maximum 1. For the headphone-blocking volume-reduced level η, we set $\eta = 0.75$ in our implementation, calculated based on the maximum amount of noise that can be canceled by wearing active noise-cancellation (ANC) headphones (45 dB, as reported by [75]), and the average noise level in real-world environments (60 dB, as reported in [20]). Furthermore, a high pass filter is added to all the RW sound sources and its frequency cutoff C can be determined based on the transparency level τ:

$$C = (1-\tau) \cdot Z \tag{2}$$

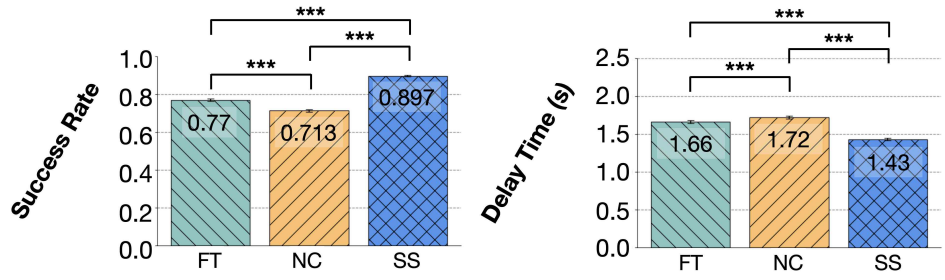
where we set the baseline frequency cutoff to be $Z = 2$kHz, since we wanted to make the half transparency in SoundShift condition (when τ=0.5, C=1kHz) matched to today's ANC technologies that can filter out signals below 1kHz according to a recent report [78]. In our study, the Full Transparency condition was configured as $\tau = 1$, resulting in $C = 0$ kHz and volume $S_{\text{new}} = S_{\text{default}}$. For the Noise Cancellation condition, the settings were $\tau = 0$, $C = 2$ kHz, and volume $S_{\text{new}} = 0.25 \cdot S_{\text{default}}$.

5.4 Apparatus

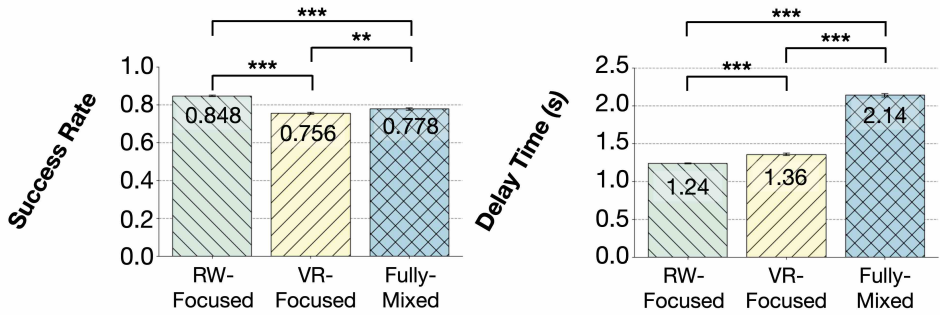
The studies were conducted in person, where participants used Apple AirPods Max provided by us (see Figure 5). We allowed participants to choose whether to use the AirPods Max's active noise cancellation mode, as some BVI individuals in our pilot study found it uncomfortable. However, we avoided the transparency mode to prevent the amplified real-world noises from distracting the participants. The studies took place in a soundproof, enclosed space to minimize distractions from external sounds. Participants completed the study tasks using a wireless keyboard, which, along with the AirPods Max, was connected via Bluetooth to a smartphone or tablet running our Unity application. Users' head movements were tracked using the gyroscope data from the AirPods Max to enable spatial audio.

5.5 Tasks and Procedure

After being welcomed and presented with our study's informed consent and procedure, participants experienced three simulated



(a) Overall results based on *Condition* in response to RQ1.



(b) Overall results based on *Scenario* in response to RQ2.

Figure 6: Results based on *Condition* (RQ1) or *Scenario* (RQ2). **=p<=0.001. ***=p<0.0001.

scenarios (RW-Focused, VR-Focused, Fully-Mixed) under three conditions: Full Transparency (FT), Noise Cancellation (NC), and SoundShift (SS). The order of scenarios and conditions was counterbalanced for the 18 participants. Participants repeated each condition for five trials, which constituted a session.

Before each session, we briefed participants on the scenario, the four sounds to identify, and their key commands (e.g., 1,2,3,4). They were instructed to complete each task "as quickly and as accurately as possible" without sacrificing accuracy for speed and vice versa. They also had a practice session to get familiar with the individual sound events and learn the audio-key mapping. Participants could adjust the volume to their preference and take breaks anytime if needed.

After each session, we verbally described the NASA TLX form [27] to our participants and obtained their responses as workload measures. As mentioned in section 5.2, there are 20 sound events in the *RW-Focused* scenario, 20 for the *VR-Focused* scenario, and 22 for the *Fully-Mixed* scenario. Therefore, for each participant, we collected data from 62 audio events × 3 conditions × 5 trials, resulting in 930 tasks. This amounts to 930 tasks × 18 individuals = 16,740 tasks. However, some data points were excluded due to technical issues, resulting in 16,700 effective tasks. The study, approved by the IRB, compensated each participant with 40 USD. It took about 2 hours, with 80 minutes for task completion and 40 minutes for NASA-TLX responses and other follow-ups.

5.6 Dependent Measures and Data Analysis

For each trial, we recorded task data, focusing on two primary dependent measures: *(i) Success Rate*, calculated as the ratio of correct key presses to total sound events, and *(ii) Delay Time*, considering only correct key presses and measuring the time from the event's onset to the correct key press. We conducted a mixed-methods analysis. We built two separate mixed-effect linear regression models [50] to examine the dependent variables *Success Rate* and *Delay Time*, with fixed effects *Condition* and *Scenario*, and their interaction *Scenario* × *Condition*, taking participant ID as a random intercept. We also transcribed our participants' feedback from each session for further analysis.

5.7 Results

This section reports our study's quantitative and qualitative results to answer each of our research questions.

5.7.1 RQ1: How do sound manipulations affect participants' performance compared to full transparency and noise cancellation? **With *SoundShift*, participants achieved a significantly higher success rate and lower time delay in identifying sounds than FT and NC, but sound manipulations may confuse users' interpretation of the sound content.**

Overall, we found that *Condition* had a significant main effect on both *Success Rate* (F(2,16700)=327.58, p<0.0001) and *Delay Time* (F(2,13242)=88.02, p<0.0001) in the three scenarios (Figure 6a). Post-hoc Tukey's pairwise test revealed that SS (M=0.897) resulted in a significantly higher *Success Rate* than FT (M=0.770) and NC (M=0.713), and FT was also significantly higher than NC, p<0.0001; for *Delay Time*, SS (M=1.43s) was significantly lower than FT (M=1.66s) and NC (M=1.72s), and FT was significantly lower than NC, p<0.0001. It was expected that SS would increase sound awareness in both RW and VR, and all participants commended the benefits of SoundShift: *"SS achieves proper volume without the problem of sounds being too noisy or mixed up"* (B16, *RW-Focused*), *"SS is quite authentic and*

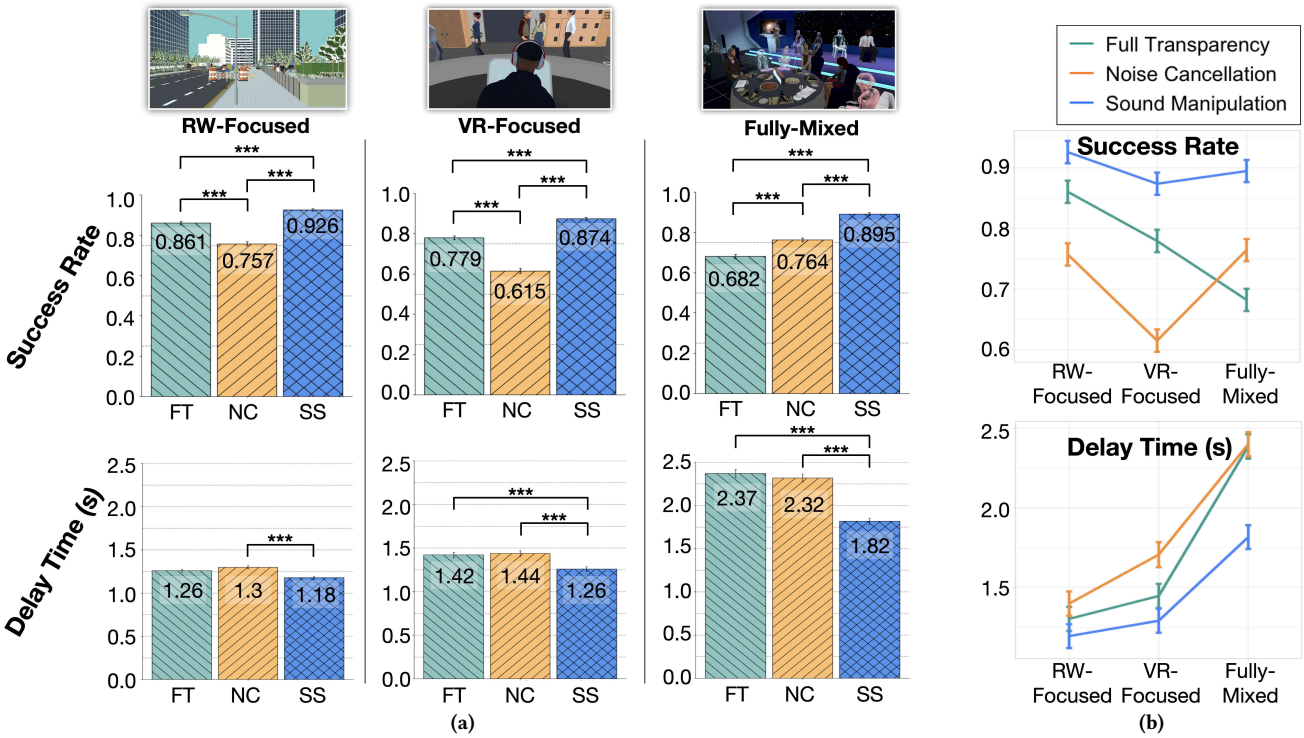


Figure 7: Overall results for RQ3 based on *Condition* and *Scenario*, including (a) *Success Rate* (top, the higher the better) and *Delay Time* (bottom, the lower the better) with error bars showing 95% confidence intervals. ***=p<0.0001. (b) Interaction effects on *Success Rate* and *Delay Time* between *Condition* and *Scenario*.

echos with my custom ... I like to wear headphones on the right ear only when walking" (B12, *RW-Focused*), *"I like SS as I also used to listen to certain things on one side"* (B8, *VR-Focused*), and *"I like it since it makes sounds very apparent, like notification of table cleaning, and the real people's voices are also different from broadcast"* (B3, *Fully-Mixed*). Also, the reason that FT led to better performance than NC was due to the nature of NC that blocks out RW sounds, as remarked by most participants and evidenced by the *Success Rate* of RW (M=0.693) and VR (M=0.895) sounds, and *Delay Time* of RW (M=1.81s) and VR (M=1.51s) sounds. Also, FT was preferable as it retained RW awareness and was more familiar to BVI participants, echoing the results in section 3.2.2 that BVI people emphasized the importance of retaining RW awareness when surveying different headphones.

However, participants also commented that SoundShift may confuse their perception and interpretation of RW sounds, such as *"... the post-processed gentle drilling made me hard to tell the scale of the construction sites"* (B10, *RW-Focused*), *"It is weird ... knocking sounds are far away from me"* (B8,*VR-Focused*), and *"The robot-like filter is distinguishable but distorts the content ... I need to pay a lot of attention"* (B12, *Fully-Mixed*). This highlights a trade-off between enhancing sound awareness and preserving the fidelity of certain sound characteristics when manipulating sounds.

5.7.2 RQ2: How do the different scenarios with varying emphases on reality and virtuality affect participants' performance? **Participants overall achieved a higher success rate and lower time delay in RW-Focused than VR-Focused and Fully-Mixed scenarios due to their familiarity.**

We found that *Scenario* had a significant main effect on both *Success Rate* (F(2,16700)=83.22, p<0.0001) and *Time Delay* (F(2,13242)=810.83, p<0.0001). Post-hoc Tukey's pairwise test revealed that participants had significantly higher *Success Rate* in *RW-Focused* (M=0.848) than in *VR-Focused* (M=0.756) and *Fully-Mixed* scenarios (M=0.778), p<0.0001 (Figure 6b). Furthermore, *Fully-Mixed* was significantly higher than *VR-Focused*, p=0.001. Participants also had significantly lower *Delay Time* in *RW-Focused* (M=1.24s) than in *VR-Focused* (M=1.36s) and *Fully-Mixed* (M=2.14s) scenarios, p<0.0001; *VR-Focused* was significantly lower than *Fully-Mixed*, p<0.0001.

Overall, results suggested that the *RW-Focused* scenario yielded better performance, possibly due to participants' familiarity with the navigation scenario and the RW sounds: *"This is how I get to my home, but mine is more complicated as typically sounds are not evenly distributed and consistent"* (B10), and *"[FT] is helpful as it's very like the real situation ... made me easily engage in"* (B15). Conversely, performance in the *VR-Focused* and *Fully-Mixed* scenarios was less effective, possibly due to their unfamiliarity with the audio content. Specifically, some participants mentioned the audio handbook's sentences in the *VR-Focused* scenario were new and too long to focus on, while the overlapping conversations in the *Fully-Mixed* scenario made the content hard to observe, not to mention further distinguishing its source from RW to VR, as argued by B10: *"I don't think the real or virtual person talking matters ... I cannot even hear clearly on the content when they are totally mixed."* These results

were expected due to our futuristic and complex design of *Fully-Mixed* scenario, where participants' unfamiliarity with the content and scenario affected their performance.

5.7.3 RQ3: How do the conditions affect participants' performance differently across the scenarios? **In all scenarios, SS was more effective and preferable than FT and NC, but still, many participants had varied preferences on sound awareness and found the sound clarity and naturalness of FT and the quietness of NC useful.**

We also found a significant interaction effect (Figure 7b) between *Condition* and *Scenario* for both *Success Rate* ($F(4,16700)=56.7$, $p<0.0001$) and *Delay Time* ($F(4,13242)=23.96$, $p<0.0001$). Specifically, in the *RW-Focused* scenario, post-hoc Tukey's pairwise test revealed that participants had significantly higher *Success Rate* with SS (M=0.926) than FT (M=0.861) and NC (M=0.757), and FT than NC, $p<0.0001$ (Figure 7a); they also had significantly lower *Delay Time* with SS (M=1.18s) than NC (M=1.30s), $p<0.0001$. This was evidenced by most participants (N=14) that SS was preferable and provided a clear sound awareness of both RW and VR sounds than FT and NC, as B18 stated *"I love the first one [SS], because the ambient noise in the second one [FT] is excessively noisy, and the construction sound in the third one [NC] is almost inaudible, which is very dangerous."* On the other hand, four participants liked FT the most as *"... the sound of the white cane is more distinctive. The drilling is more vivid, which makes me more aware of the situation on the road ... It [FT] is helpful because the sound is pretty spatial, I am able to perceive the composition of the environment"* (B14). Participants also indicated that while NC made instructions clearer, it also lost information from RW, which was unfavorable in navigation scenario.

In the *VR-Focused* scenario, post-hoc Tukey's pairwise test revealed that participants achieved a significantly higher *Success Rate* with SS (M=0.874) than FT (M=0.779) and NC (M=0.615), and FT than NC, $p<0.0001$ (Figure 7a); they also had significantly lower *Delay Time* with SS (M=1.26s) than FT (M=1.42s, $p=0.0065$) and NC (M=1.44s, $p<0.0001$). Fourteen participants preferred and performed better in SS for its distinct audio channels, with the audio handbook on the left and the supervisor's voice note on the right: *"All details can be observed. The separate audio channel design is quite good, especially in such a static environment ... it allowed me to distinguish different sounds from different channels ... ease my burden"* (B14). Another three participants liked SS and FT equally as FT retained the fidelity of sounds, as remarked by B8: *"Both SS and FT have aspects I like. FT makes it a real-life scenario, while the information in SS is very clear."* One participant (B13) preferred NC for its quieter environment due to his high sensitivity to sounds.

In the *Fully-Mixed* scenario, post-hoc Tukey's pairwise test revealed that participants performed significantly better with SS (M=0.891) than FT (M=0.68) and NC (M=0.762), and NC than FT, $p<0.0001$ (Figure 7a); they also had significantly lower *Delay Time* with SS (M=1.82s) than FT (M=2.37s, $p<0.0001$) and NC (M=2.32s, $p<0.0001$). Many participants (N=11) preferred SS for its sound clarity, as commented by B15: *"It's helpful and allows me to easily differentiate between broadcasts and human voices. Also, the sounds don't seem to overlap as much."* However, some participants felt more pressured and overwhelmed in SS and preferred FT (N=5), as stated by B10: *"I liked FT which has natural voices, and it has no earcons that might drown out human voices"*, while others preferred NC (N=2), like B12: *"The sound in this condition [NC] is clear and comes gradually, though it doesn't make every sound very clear. The second one [SS] makes sounds too clear ... It might cause me to overlook other important sounds."* In this complex scenario, participants' preferences varied, and the enhanced clarity of specific sounds in SS raised concerns about potentially missing other crucial audio information.

5.7.4 RQ4: How do sound manipulations affect participants' cognitive load compared to full transparency and noise cancellation? ***SoundShift*** **overall reduced the cognitive load.**

We found a significant main effect of *Condition* on the overall workload in NASA-TLX ($F(2,136)=17.01$, $p<0.0001$). Post-hoc Tukey's pairwise test revealed that SS (M=34.8) resulted in a significantly lower workload than both FT (M=45.4) and NC (M=48.2), $p<0.0001$, but there was no significant difference between FT and NC ($p=0.4637$). The reason behind this result might be explained by the previous findings, where SS increased sound awareness via different manipulations; on the other hand, both FT and NC resulted in similar cognitive load, likely due to their inherent limitations, such as the multiple overlapping sounds in FT and the nearly-blocked RW sounds in NC. Interestingly, we found no significant differences between SS and the other two conditions for the overall cognitive load in the *Fully-Mixed* scenario.

This might be due to the result in section 5.7.3 that the stylized voices in SS sometimes made content unclear, or the enhanced clarity of specific sounds in SS might overshadow other crucial audio information.

5.7.5 RQ5: How do participants describe their experiences and ways to further customize their soundscape for each scenario? **Participants suggested retaining natural sounds and customizing sound manipulations based on the content and context of sounds.**

Most participants suggested dynamically increasing the volume of certain sounds or reducing background noises to improve identification performance. They also noted using different filters for certain sounds for better comfort. Several participants proposed various customizations beyond the sound manipulations used in our study.

First, most participants favored the natural presentation of RW sounds to maintain their perception of the real world, (Section 5.7.1). Yet, some participants desired more intelligent sound manipulations based on the RW context. For instance, B1 said in the *RW-Focused* scenario *"I would only mute the voices of strangers but not all people as I want to hear the voices of someone I know"*, and B10 also mentioned *"I think earcon should be played only when the construction site is large enough, as typically, a small scale of construction would not be so noisy and discomforting."* Similarly, most participants also preferred keeping the voices of virtual people in the *Fully-Mixed* scenario unaltered, not imposing another sound filter on them. Instead, they may use other methods, such as sound localization or discerning voice nuances, to distinguish between RW and virtual reality (VR) sources.

Participants also proposed manipulating sounds based on the audio content, where sound characteristics can adaptively change based on the urgency and importance of the content. For instance,

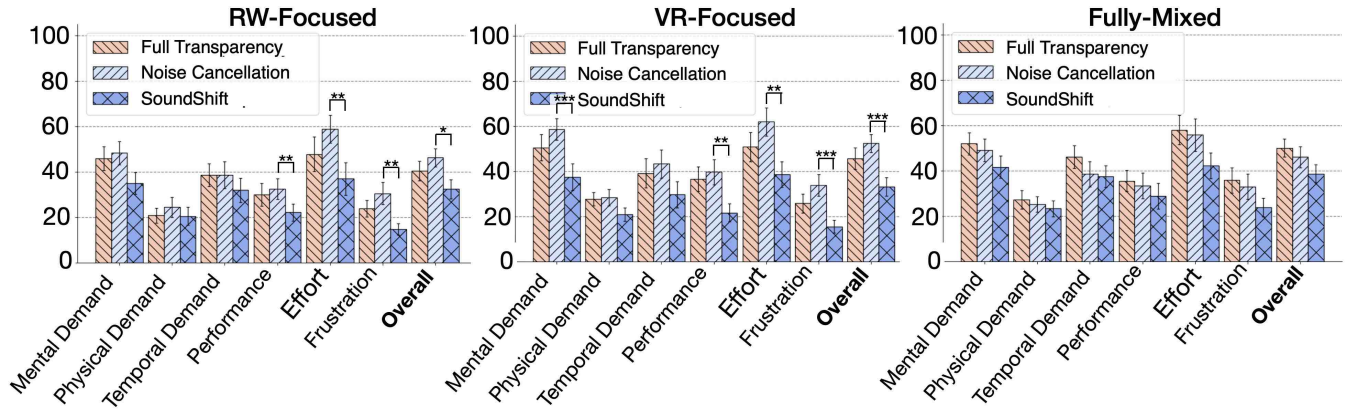


Figure 8: Results of NASA-TLX (RQ4) with error bars showing 95% confidence intervals. $^{***}p<0.0001$, $^{**}p<0.001$ and $^{*}p<0.01$.

B15 stated in the *VR-Focused* scenario that *"I would lower down the public announcements if the content is not relevant to me"*, and some participants also mentioned the volume of the supervisor's voice notes should be proportional to the urgency of the content. Furthermore, B3 mentioned managing the voice font for the virtual broadcast based on the content *"I would change it to a cuter voice such as characters in the game or anime ... however, serious content can be presented in a deep voice."*

In sum, besides our proposed sound manipulations, participants also suggested that sound manipulations could be grounded on context and content to better manage their attention.

6 EXAMPLE APPLICATIONS WITH SOUNDSHIFT

Based on insights from our studies, we developed three proof-of-concept prototypes to showcase the practicality and generalizability of SoundShift manipulations in different MR experiences. Though BVI people did not evaluate these applications, we hoped these applications stimulate further discussions and advancements in sound manipulation for future mixed-reality applications to promote accessibility. Note that these prototypes are functional. We highlight user interactions rather than technical specifics in this section. Please refer to our Video Figure for demonstrations.

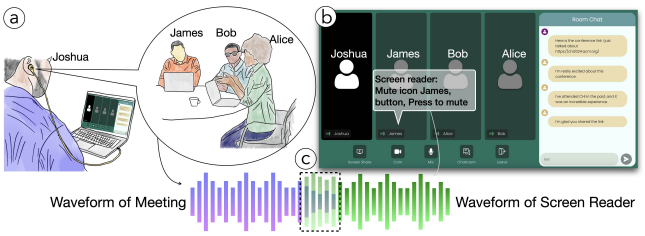


Figure 9: A prototype of an accessible online meeting application and demonstrated waveforms showing how sounds are manipulated. (a) The voices of the three people are assigned to the left, front, and right spatial locations around the BVI user, same as (b) the layout of the attendee panel. (c) When using a screen reader to navigate, the meeting volume is decreased to accentuate the audio feedback of the screen reader.

6.1 Accessible Online Meeting Application

In response to difficulties highlighted in section 3.2.1 about using screen readers during virtual meetings, we developed an audio-adaptive online meeting web application. This app resolves conflicts between screen reader and meeting audio using Time Shift, seamlessly blending these sounds by adjusting their auditory characteristics. For example, it increases the screen reader's volume when someone else speaks but allows users to prioritize meeting audio if it holds higher priority. To address the lack of spatial awareness common in virtual meetings, Position Shift arranges attendees' voices across the left-right audio spectrum, corresponding to their positions on the attendee panel. This feature lets users easily map the voice's location to the attendee's location on the interface for further interaction (e.g., sending private messages).

6.2 Content-Aware MR Image Exploration

Mixed-reality images are increasingly pervasive, such as in news or research articles, which entail rich visual information. However, the spatial information is not adapted from the visual to the audio domain. In response, our image touch exploration system built on [46] spatializes the screen reader's voice feedback by mapping the touch location on the image to the sound feedback's spatial location (Figure 10) for enhancing spatial understanding. Moreover, MR images, which combine digital and physical visuals easily distinguished visually but inaccessible to BVI people. To address this, during image exploration, Transparency Shift renders acoustic transparency for real-world (RW) content and occlusive opacity for virtual reality (VR) content, such as crowd noises in the office for RW content and bird chirping for VR content (Figure 10), which signals the reality and virtuality of content without additional text descriptions. Moreover, Style Shift alters the voice of description to a cartoonish tone to match the digital world's style and enhance immersion.

6.3 Context-Aware Outdoor Navigation

Navigation is indispensable to BVI people, and many commercial apps [4, 5] have been developed to facilitate BVI people's independence and autonomy. However, occasional events may distract users from navigation. We developed a mobile navigation application that analyzes both real-world and virtual sounds to identify opportune

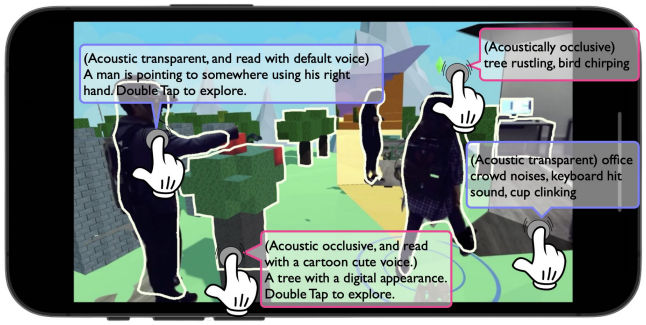


Figure 10: Using ImageExplorer [46] with sound manipulations. The example image was provided by Wang et al. [84], which includes a combination of real and virtual content. The voice corresponding to the touch locations is spatialized from left to right. The voice of VR content is read with a cartoon-ish voice to fit the digital cartoon-style environment, while the voice of RW content remains as default. Noise cancellation mode is turned on when exploring VR content, and shifts to transparency mode when exploring RW content.

Figure 11: Illustration of our context-aware navigation app. (a) The crowd passing by the BVI person generates sounds that are detected by the app, and at the same time, (b) an audio direction is about to happen but will not be played due to its conflict with crowd sounds. (c) After the crowd leaves, the audio direction is played.

moments for delivering audio directions. When a certain RW event is happening and detected, the audio directions will be delayed by Time Shift until the RW event ends (Figure 11). If a certain RW sound is detected but overshadowed by the audio directions, this app can provide post-hoc verbal descriptions to notify users of the detected sound.

7 DISCUSSION AND FUTURE WORK

We have contributed the concept of *SoundShift* with six sound manipulators to increase MR sound awareness for BVI people, a user study to prove the effectiveness of sound manipulations on different MR scenarios, and three example applications to demonstrate its feasibility for practical use. In this section, we discuss our work's limitations and potential improvements, implications for future MR soundscape design, and the generalizability of our findings to broader communities.

7.1 From Simulation to Practical Applications

In our work, we created Unity simulations for real-time manipulation of sound characteristics and developed three proof-of-concept applications to illustrate the generalizability and practicality of SoundShift. We encountered several challenges; for instance, the audio streams in most existing platforms were at the operation system level, which made them hard to access for application-level purposes. Furthermore, implementing these sound manipulations in the real world necessitates several requirements, such as supporting the recognition of diverse categories of sounds, extracting sounds of interest from multiple overlapping ones, and rendering the extracted sounds in their original or higher quality. And all of these components should be achieved in real time to ensure a seamless user experience, which is still challenging in sound research. These are why our three example applications mostly revolve around manipulating virtual audio.

In recent years, researchers have attempted to tackle these challenges. For instance, Ubicoustics [42] supports the recognition of many sound activities through commercially available microphones, Jain et al. [32] developed ProtoSound for people to customize their sound recognition model, and Veluri et al. [80] approached real-time target sound extraction. Also, to increase the accuracy of sound identification or extraction, we could approach the users' sound context with other sensing modalities. For example, in the *RW-Focused* scenario, one's smartwatch could provide gyroscope data to detect white cane taps. Additionally, everyday objects often produce sounds linked with visual elements, detectable through cameras, such as playing an instrument [77]. These approaches can provide a preliminary context for the targeted sound identification or extraction. Despite a long way ahead in real-time manipulation of overarching sounds, recent research endeavors have illuminated promises to apply SoundShift to practical applications.

7.2 Towards Sound-Aware Description Manipulations

As the introduction outlines, incorporating visual descriptions is another essential auditory element to facilitate MR experiences for BVI people. Unlike diegetic sound effects (e.g., knocking, drilling, dog barking), descriptions provide specific information and semantic meanings crucial for BVI users. In section 5.7.5, participants suggested making virtual broadcasts or voice notes discernible amidst other sounds only if the content is deemed critical, as a strategy to mitigate information overload. It is thus promising to adjust the description content or provide opportune sentence breaks to prevent it from overlapping with other MR auditory elements. For instance, the system can offer comprehensive descriptions during quieter moments, while providing succinct yet informative descriptions when time is limited, to create a harmonized user experience. This is similar to creating audio descriptions with dynamic time constraints in a video, such as Rescribe [68] to shorten the descriptions by removing less important words. Overall, sound manipulation in MR environments should extend beyond adjusting sound characteristics, temporal, and spatial aspects. It should also consider the content and audio context, making it more intelligent, adaptive, and user-centric.

7.3 Customizable Sound Manipulations

Though each manipulator in our Unity implementation can manipulate sounds in real-time (e.g., overlap of several sounds) to provide a seamless MR soundscape, we did not consider the varied preferences of our participants. As described in section 5.7.5, participants proposed several ideas on manipulating sounds based on their current context and content of audio information (e.g., sounds, text descriptions), which also echos with our formative results in section 3.2.5 that BVI people desired to augment or customize the sound library of existing applications. It is thus promising to integrate methods that gauge the importance of audio information, such as content analysis or sentiment analysis. Based on the gauged importance, users could personalize sound representation (e.g., earcons, descriptions, diegetic sounds) and presentation (e.g., high/low volume, sound locations) to minimize distraction. Additionally, participant preferences for sound manipulation, influenced by personal experiences and memories, highlight the need for adaptable soundscapes; for instance, the sound of the sliding door in the *VR-Focused* scenario made B1 recall her bad memory of consuming videos about insects, and she wanted to disable it selectively. Research on visual descriptions also emphasized individual customization over a one-size-fits-all approach [39, 52, 53, 76]. Echoing this notion, it is, therefore, worth exploring customizable sound properties and how to design end-user interfaces to enable user customization.

7.4 Generalizing Results to Broader Groups with Different Sensory Modalities

Though our content analysis from online posts gave us broad insights into different scenarios and needs on the everyday consumption of complex sounds, it lacked depth in understanding the full context behind the posts. Future research could dive deeper into relevant topics, such as understanding diverse contexts of consuming complex sounds, the sound technologies BVI people currently use to help with everyday tasks, or their strategies for handling sounds in more futuristic and complex mixed-reality scenarios.

Furthermore, the simulations in our main study may not fully capture the multisensory experience that BVI individuals rely on, such as a combination of haptic feedback, smell, and other contextual cues, to enhance their real-world awareness. That was also why several participants in the *RW-Focused* scenario emphasized the importance of haptic feedback from the white cane besides its auditory feedback. Nonetheless, while other sensory feedback can augment and address the limitations of sound, they are all of significance and cannot replace one another. Also, none of our participants consumed visual feedback despite some having residual visual ability. Instead, they used hearing only to perform the tasks. Future work could investigate how people strategize their sensory modalities for balancing cognitive load and performance in mixed reality, which may enable manipulators across modalities (e.g., split conflicting audio information into visual and audio ones) to create more sensory-adaptive and accessible mixed reality experiences.

8 CONCLUSION

We have presented the concept of SoundShift to make MR sound awareness accessible for BVI people, through six sound manipulators derived from our content analysis on BVI forums, including Transparency Shift, Envelope Shift, Position Shift, Style Shift, Time Shift, and Sound Append. We instantiated the six sound manipulators and three simulated scenarios across the Reality-Virtuality continuum in Unity. We then conducted a user study with eighteen BVI people and found empirical evidence that the six sound manipulations significantly enhanced users' sound awareness and reduced cognitive load. We also found varied preferences and comments across participants on manipulating sounds, which spurred several discussions and promises of future work. Finally, we implemented three proof-of-concept applications to demonstrate the generalizability and practicality of SoundShift, including an accessible online meeting app, an immersive image understanding system, and a context-aware navigation app.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their suggestions and all participants of our study.

REFERENCES

[1] 2023. AirPods redefine the personal audio experience. https://www.apple.com/newsroom/2023/06/airpods-redefine-the-personal-audio-experience/
[2] 2023. AppleVis. https://www.applevis.com/
[3] 2023. Blind and Visually Impaired Community. https://www.reddit.com/r/Blind/
[4] 2023. BlindSquare. https://www.blindsquare.com/
[5] 2023. Microsoft Soundscape. https://www.microsoft.com/en-us/research/product/soundscape/
[6] David Lindlbauer. Alexander Wang, Yi Fei Cheng. 2024. MARingBA: Music-Adaptive Ringtones for Blended Audio Notification Delivery. In *In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Honolulu, HI,. https://doi.org/10.1145/3613904.3642376
[7] Cédric R André, Jean-Jacques Embrechts, Jacques G Verly, Marc Rébillat, and Brian FG Katz. 2012. Sound for 3D cinema and the sense of presence. Georgia Institute of Technology.
[8] Greg Ballou. 1987. Handbook for Sound Engineers: The New Audio Cyclopedia, Howard W. *Sams and Co., Indianapolis* (1987).
[9] Benjamin B Bederson. 1995. Audio augmented reality: a prototype automated tour guide. In *Conference companion on Human factors in computing systems*. 210–211.
[10] Meera M Blattner, Denise A Sumikawa, and Robert M Greenberg. 1989. Earcons and icons: Their structure and common design principles. *Human–Computer Interaction* 4, 1 (1989), 11–44.
[11] Willem-Paul Brinkman, Allart RD Hoekstra, and René van EGMOND. 2015. The effect of 3D audio and other audio techniques on virtual reality experience. *Annual Review of Cybertherapy and Telemedicine 2015* (2015), 44–48.
[12] Ruei-Che Chang, Chao-Hsien Ting, Chia-Sheng Hung, Wan-Chen Lee, Liang-Jin Chen, Yu-Tzu Chao, Bing-Yu Chen, and Anhong Guo. 2022. OmniScribe: Authoring Immersive Audio Descriptions for 360° Videos. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 15, 14 pages. https://doi.org/10.1145/3526113.3545613
[13] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M. Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services* (Portland, Oregon) *(MobiSys '22)*. Association for Computing Machinery, New York, NY, USA, 384–396. https://doi.org/10.1145/3498361.3538933
[14] Kuan-Wen Chen, Yung-Ju Chang, and Liwei Chan. 2022. Predicting Opportune Moments to Deliver Notifications in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 186, 18 pages. https://doi.org/10.1145/3491102.3517529
[15] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.
[16] Laurence Cliffe, James Mansell, Chris Greenhalgh, and Adrian Hazzard. 2021. Materialising contexts: virtual soundscapes for real-world exploration. *Personal and Ubiquitous Computing* 25 (2021), 623–636.
[17] Florian Daiber, Donald Degraen, André Zenner, Tanja Döring, Frank Steinicke, Oscar Javier Ariza Nunez, and Adalberto L. Simeone. 2021. Everyday Proxy Objects for Virtual Reality. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association

for Computing Machinery, New York, NY, USA, Article 101, 6 pages. https://doi.org/10.1145/3411763.3441343

[18] Edoardo D'Atri, Carlo Maria Medaglia, Alexandru Serbanati, Ugo Biader Ceipidor, Emanuele Panizzi, and Alessandro D'Atri. 2007. A system to aid blind people in the mobility: A usability test and its results. In *Second International Conference on Systems (ICONS'07)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 35–35.

[19] Isamu Endo, Kazuki Takashima, Maakito Inoue, Kazuyuki Fujita, Kiyoshi Kiyokawa, and Yoshifumi Kitamura. 2021. ModularHMD: A Reconfigurable Mobile Head-Mounted Display Enabling Ad-Hoc Peripheral Interactions with the Real World. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 100–117. https://doi.org/10.1145/3472749.3474738

[20] Centers for Disease Control and Prevention. 2023. What Noises Cause Hearing Loss? https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html

[21] William W Gaver and Donald A Norman. 1988. *Everyday listening and auditory icons*. Ph. D. Dissertation. University of California, San Diego, Department of Cognitive Science and Psychology.

[22] Sarthak Ghosh, Lauren Winston, Nishant Panchal, Philippe Kimura-Thollander, Jeff Hotnog, Douglas Cheong, Gabriel Reyes, and Gregory D. Abowd. 2018. NotifiVR: Exploring Interruptions and Notifications in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1447–1456. https://doi.org/10.1109/TVCG.2018.2793698

[23] Robert H Gilkey and Janet M Weisenberger. 1995. The sense of presence for the suddenly deafened adult: Implications for virtual environments. *Presence: Teleoperators & Virtual Environments* 4, 4 (1995), 357–363.

[24] Uwe Gruenefeld, Jonas Auda, Florian Mathis, Stefan Schneegass, Mohamed Khamis, Jan Gugenheimer, and Sven Mayer. 2022. VRception: Rapid Prototyping of Cross-Reality Systems in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 611, 15 pages. https://doi.org/10.1145/3491102.3501821

[25] Gabriel Haas, Evgeny Stemasov, and Enrico Rukzio. 2018. Can't You Hear Me? Investigating Personal Soundscape Curation. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia* (Cairo, Egypt) *(MUM '18)*. Association for Computing Machinery, New York, NY, USA, 59–69. https://doi.org/10.1145/3282894.3282897

[26] Mack Hagood. 2011. Quiet comfort: Noise, otherness, and the mobile production of personal space. *American Quarterly* 63, 3 (2011), 573–589.

[27] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[28] Jeremy Hartmann, Christian Holz, Eyal Ofek, and Andrew D. Wilson. 2019. RealityCheck: Blending Virtual Environments with Situated Physical Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300577

[29] Jaylin Herskovitz, Jason Wu, Samuel White, Amy Pavel, Gabriel Reyes, Anhong Guo, and Jeffrey P. Bigham. 2020. Making Mobile Augmented Reality Applications Accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. https://doi.org/10.1145/3373625.3417006

[30] Anuruddha Hettiarachchi and Daniel Wigdor. 2016. Annexing Reality: Enabling Opportunistic Use of Everyday Objects as Tangible Proxies in Augmented Reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1957–1967. https://doi.org/10.1145/2858036.2858134

[31] Ching-Yu Hsieh, Yi-Shyuan Chiang, Hung-Yu Chiu, and Yung-Ju Chang. 2020. Bridging the Virtual and Real Worlds: A Preliminary Study of Messaging Notifications in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376228

[32] Dhruv Jain, Khoa Huynh Anh Nguyen, Steven M. Goodman, Rachel Grossman-Kahn, Hung Ngo, Aditya Kusupati, Ruofei Du, Alex Olwal, Leah Findlater, and Jon E. Froehlich. 2022. ProtoSound: A Personalized and Scalable Sound Recognition System for Deaf and Hard-of-Hearing Users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 305, 16 pages. https://doi.org/10.1145/3491102.3502020

[33] Dhruv Jain, Sasa Junuzovic, Eyal Ofek, Mike Sinclair, John R. Porter, Chris Yoon, Swetha Machanavajhala, and Meredith Ringel Morris. 2021. Towards Sound Accessibility in Virtual Reality. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 80–91. https://doi.org/10.1145/3462244.3479946

[34] Gaurav Jain, Yuanyang Teng, Dong Heon Cho, Yunhao Xing, Maryam Aziz, and Brian A. Smith. 2023. "I Want to Figure Things Out": Supporting Exploration in Navigation for People with Visual Impairments. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 63 (apr 2023), 28 pages. https://doi.org/10.1145/3579496

[35] Tiger F. Ji, Brianna Cochran, and Yuhang Zhao. 2022. VRBubble: Enhancing Peripheral Awareness of Avatars for People with Visual Impairments in Social Virtual Reality. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 17 pages. https://doi.org/10.1145/3517428.3544821

[36] Stine S. Johansen, Niels van Berkel, and Jonas Fritsch. 2022. Characterising Soundscape Research in Human-Computer Interaction. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1394–1417. https://doi.org/10.1145/3532106.3533458

[37] Mohamed Kari, Tobias Grosse-Puppendahl, Alexander Jagaciak, David Bethge, Reinhard Schütte, and Christian Holz. 2021. SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 118–133. https://doi.org/10.1145/3472749.3474739

[38] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. 2000. Suede: A Wizard of Oz Prototyping Tool for Speech User Interfaces. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (San Diego, California, USA) *(UIST '00)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/354401.354406

[39] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. *arXiv preprint arXiv:2205.10646* (2022).

[40] Michael Krzyzaniak, David Frohlich, and Philip J.B. Jackson. 2019. Six types of audio that DEFY reality! A taxonomy of audio augmented reality with examples. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound* (Nottingham, United Kingdom) *(AM '19)*. Association for Computing Machinery, New York, NY, USA, 160–167. https://doi.org/10.1145/3356590.3356615

[41] Rachel L. Franz, Sasa Junuzovic, and Martez Mott. 2021. Nearmi: A Framework for Designing Point of Interest Techniques for VR Users with Limited Mobility. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 5, 14 pages. https://doi.org/10.1145/3441852.3471230

[42] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 213–224. https://doi.org/10.1145/3242587.3242609

[43] Pontus Larsson, Aleksander Väljamäe, Daniel Västfjäll, Ana Tajadura-Jiménez, and Mendel Kleiner. 2010. Auditory-induced presence in mixed reality environments and related technology. In *The engineering of mixed reality systems*. Springer, 143–163.

[44] A. Lecuyer, P. Mobuchon, C. Megard, J. Perret, C. Andriot, and J.-P. Colinot. 2003. HOMERE: a multimodal system for visually impaired people to explore virtual environments. In *IEEE Virtual Reality, 2003. Proceedings.* Institute of Electrical and Electronics Engineers, New York, NY, USA, 251–258. https://doi.org/10.1109/VR.2003.1191147

[45] Cheuk Yin Phipson Lee, Zhuohao Zhang, Jaylin Herskovitz, JooYoung Seo, and Anhong Guo. 2022. CollabAlly: Accessible Collaboration Awareness in Document Editing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 596, 17 pages. https://doi.org/10.1145/3491102.3517635

[46] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 462, 15 pages. https://doi.org/10.1145/3491102.3501966

[47] Min Kyung Lee, Jodi Forlizzi, Paul E Rybski, Frederick Crabbe, Wayne Chung, Josh Finkle, Eric Glaser, and Sara Kiesler. 2009. The snackbot: documenting the design of a robot for long-term human-robot interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 7–14.

[48] Ziming Li, Shannon Connell, Wendy Dannels, and Roshan Peiris. 2022. SoundVizVR: Sound Indicators for Accessible Sounds in Virtual Reality for Deaf or Hard-of-Hearing Users. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 13 pages. https://doi.org/10.1145/3517428.3544817

[49] David Lindlbauer and Andy D. Wilson. 2018. Remixed Reality: Manipulating Space and Time in Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173703

[50] Mary J Lindstrom and Douglas M Bates. 1988. Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.* 83, 404 (1988), 1014–1022.

[51] Matthew Lombard and Theresa Ditton. 1997. At the heart of it all: The concept of presence. *Journal of computer-mediated communication* 3, 2 (1997), JCMC321.

[52] Mariana Lopez, Gavin Kearney, and Krisztián Hofstädter. 2018. Audio Description in the UK: What works, what doesn't, and understanding the need for personalising access. *British journal of visual impairment* 36, 3 (2018), 274–291.

[53] Mariana Lopez, Gavin Kearney, and Krisztián Hofstädter. 2022. Seeing films through sound: Sound design, spatial audio, and accessibility for visually impaired audiences. *British Journal of Visual Impairment* 40, 2 (2022), 117–144.

[54] Aadil Mamuji, Roel Vertegaal, Changuk Sohn, and Daniel Cheng. 2005. Attentive Headphones: Augmenting Conversational Attention with a Real World TiVo. In *Extended Abstracts of CHI*, Vol. 5.

[55] Mark McGill, Daniel Boland, Roderick Murray-Smith, and Stephen Brewster. 2015. A Dose of Reality: Overcoming Usability Challenges in VR Head-Mounted Displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2143–2152. https://doi.org/10.1145/2702123.2702382

[56] Mark McGill, Stephen Brewster, David McGookin, and Graham Wilson. 2020. Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376702

[57] Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems* 77, 12 (1994), 1321–1329.

[58] Martez Mott, Edward Cutrell, Mar Gonzalez Franco, Christian Holz, Eyal Ofek, Richard Stoakley, and Meredith Ringel Morris. 2019. Accessible by Design: An Opportunity for Virtual Reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 451–454. https://doi.org/10.1109/ISMAR-Adjunct.2019.00122

[59] Martez Mott, John Tang, Shaun Kane, Edward Cutrell, and Meredith Ringel Morris. 2020. "I Just Went into It Assuming That I Wouldn't Be Able to Have the Full Experience": Understanding the Accessibility of Virtual Reality for People with Limited Mobility. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 43, 13 pages. https://doi.org/10.1145/3373625.3416998

[60] Craig D Murray, Paul Arnold, and Ben Thornton. 2000. Presence accompanying induced hearing loss: Implications for immersive virtual environments. *Presence* 9, 2 (2000), 137–148.

[61] Vishnu Nair, Jay L Karp, Samuel Silverman, Mohar Kalra, Hollis Lehv, Faizan Jamil, and Brian A. Smith. 2021. NavStick: Making Video Games Blind-Accessible via the Ability to Look Around. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 538–551. https://doi.org/10.1145/3472749.3474768

[62] Vishnu Nair, Shao-en Ma, Ricardo E Gonzalez Penuela, Yicheng He, Karen Lin, Mason Hayes, Hannah Huddleston, Matthew Donnelly, and Brian A Smith. 2022. Uncovering Visually Impaired Gamers' Preferences for Spatial Awareness Tools Within Video Games. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.

[63] Vishnu Nair, Shao-en Ma, Hannah Huddleston, Karen Lin, Mason Hayes, Matthew Donnelly, Ricardo E Gonzalez, Yicheng He, and Brian A. Smith. 2021. Towards a Generalized Acoustic Minimap for Visually Impaired Gamers. In *The Adjunct Publication of the 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 89–91. https://doi.org/10.1145/3474349.3480177

[64] Vishnu Nair, Shao-en Ma, Hannah Huddleston, Karen Lin, Mason Hayes, Matthew Donnelly, Ricardo E Gonzalez, Yicheng He, and Brian A. Smith. 2021. Towards a Generalized Acoustic Minimap for Visually Impaired Gamers. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 89–91. https://doi.org/10.1145/3474349.3480177

[65] William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind K. Dey, and Min Kyung Lee. 2012. A Fieldwork of the Future with User Enactments. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 338–347. https://doi.org/10.1145/2317956.2318008

[66] William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind K. Dey, and Min Kyung Lee. 2012. A Fieldwork of the Future with User Enactments. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 338–347. https://doi.org/10.1145/2317956.2318008

[67] William Odom, John Zimmerman, Jodi Forlizzi, Hajin Choi, Stephanie Meier, and Angela Park. 2014. Unpacking the Thinking and Making behind a User Enactments Project. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 513–522. https://doi.org/10.1145/2598510.2602960

[68] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 747–759. https://doi.org/10.1145/3379337.3415864

[69] Donald A Ramsdell. 1978. The psychology of the hard-of-hearing and the deafened adult. *Hearing and deafness* 4 (1978), 499–510.

[70] Joan Sol Roo and Martin Hachet. 2017. One Reality: Augmenting How the Physical World is Experienced by Combining Multiple Mixed Reality Modalities. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 787–795. https://doi.org/10.1145/3126594.3126638

[71] Monika Rychtarikova. 2015. How do blind people perceive sound and soundscape. *Akustika* 23, 1 (2015), 6–9.

[72] Nitin Sawhney and Chris Schmandt. 2000. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM transactions on Computer-Human interaction (TOCHI)* 7, 3 (2000), 353–383.

[73] Adalberto L. Simeone, Eduardo Velloso, and Hans Gellersen. 2015. Substitutional Reality: Using the Physical Environment to Design Virtual Reality Experiences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3307–3316. https://doi.org/10.1145/2702123.2702389

[74] Alexa F. Siu, Mike Sinclair, Robert Kovacs, Eyal Ofek, Christian Holz, and Edward Cutrell. 2020. Virtual Reality Without Vision: A Haptic and Auditory White Cane to Navigate Complex Virtual Worlds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376353

[75] Advanced Hearing Solutions. 2023. How Does Active Noise Cancelling Work? https://hearlife.org/how-does-active-noise-cancelling-work/

[76] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who Are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 15 pages. https://doi.org/10.1145/3441852.3471233

[77] Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2745–2754.

[78] The New York Times. 2020. What Your Noise-Cancelling Headphones Can and Can't Do. https://www.nytimes.com/wirecutter/blog/what-noise-cancelling-headphones-do/

[79] Rob van Rijswijk and Jeroen Strijbos. 2013. Sounds in Your Pocket: Composing Live Soundscapes with an App. *Leonardo Music Journal* 23 (12 2013), 27–29. https://doi.org/10.1162/LMJ_a_00149 arXiv:https://direct.mit.edu/lmj/article-pdf/doi/10.1162/LMJ_a_00149/1674871/lmj_a_00149.pdf

[80] Bandhav Veluri, Justin Chan, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Real-time target sound extraction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[81] Bandhav Veluri, Malek Itani, Justin Chan, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Semantic Hearing: Programming Acoustic Scenes with Binaural Hearables. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 89, 15 pages. https://doi.org/10.1145/3586183.3606779

[82] World Wide Web Consortium (W3C). 2021. XR Accessibility User Requirements. https://www.w3.org/TR/xaur/

[83] Chiu-Hsuan Wang, Bing-Yu Chen, and Liwei Chan. 2022. RealityLens: A User Interface for Blending Customized Physical World View into Virtual Reality. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 49, 11 pages. https://doi.org/10.1145/3526113.3545686

[84] Chiu-Hsuan Wang, Chia-En Tsai, Seraphina Yong, and Liwei Chan. 2020. Slice of Light: Transparent and Integrative Transition Among Realities in a Multi-HMD-User Environment. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 805–817. https://doi.org/10.1145/3379337.3415868

[85] André Zenner, Marco Speicher, Sören Klingner, Donald Degraen, Florian Daiber, and Antonio Krüger. 2018. Immersive Notification Framework: Adaptive & Plausible Notifications in Virtual Reality. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188505

[86] Yuhang Zhao, Cynthia L. Bennett, Hrvoje Benko, Edward Cutrell, Christian Holz, Meredith Ringel Morris, and Mike Sinclair. 2018. Enabling People with Visual Impairments to Navigate Virtual Reality with a Haptic and Auditory Cane Simulation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173690

[87] Yuhang Zhao, Edward Cutrell, Christian Holz, Meredith Ringel Morris, Eyal Ofek, and Andrew D. Wilson. 2019. SeeingVR: A Set of Tools to Make Virtual Reality More Accessible to People with Low Vision. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300341

[88] Andreas Zimmermann and Andreas Lorenz. 2008. LISTEN: a user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 389–416.