Interactive Auditory Mediated Reality: Towards User-defined Personal Soundscapes

Gabriel Haas *, Evgeny Stemasov *, Michael Rietzler and Enrico Rukzio
Institute of Media Informatics
Ulm University, Ulm
<firstname>.<lastname>@uni-ulm.de

ABSTRACT
Many people utilize audio equipment to escape from noises around them, leading to the desired isolation but also dangerously reduced awareness. Mediation of sounds through smarter headphones (e.g., hearables) could address this by providing non-uniform interaction with sounds while retaining a comfortable, yet informative soundscape. In a week-long event sampling study (n = 12), we found that users mostly desire muting or a distinct "quiet-but-audible" volume for sound sources. A follow-up study (n = 12) compared a reduced interaction granularity with a continuous one in VR. Usability and workload did not differ significantly for the two granularities but a set of four states can be considered sufficient for most scenarios, namely: "muted", "quieter", "louder" and "unchanged", allowing for smoother interaction flows. We provide implications for the design of interactive auditory mediated reality systems enabling users to be safe, comfortable and less isolated from their surroundings, while re-gaining agency over their sense of hearing.

Author Keywords
Auditory Mediated Reality, Auditory Augmented Reality, Hearables, Augmented Hearing, Soundscapes

CCS Concepts
•Human-centered computing → Human computer interaction (HCI): Ubiquitous and mobile computing; •Applied computing → Sound and music computing;

INTRODUCTION
Today’s urban settings are populated by rich acoustic environments, composed of various sound sources [104]. Parts of these soundscapes are aesthetically pleasing to the population. However, there are also bothersome and irrelevant sounds which cause discomfort [120] and are linked to severe health issues [12]. Unlike vision, hearing is not directed and can hardly be averted. To avert it like gaze, one has to essentially close it off entirely – which is what users attempt to replicate with personal audio technology [44, 26]. With the rise of ubiquitous personal audio devices these characteristics of hearing are reflected in the increasing use of headphones and similar devices. The use of personal audio technology is essentially an interaction with the surrounding soundscape on a personal level (i.e., the acoustic effect is only perceptible to the user). However, this interaction is uniform: all sounds in the environment are essentially affected to the same degree and in the same fashion. Therefore, all sounds are dampened or masked virtually alike. Current devices add a static degree of filtering (passive, like physical dampening e.g. earplugs or active, like active noise cancelling headphones) and augmentation via playback of media. Those augmentations can also act as masking sounds, supporting the already present dampening. Putting on headphones therefore drowns out all sound sources in the environment the same way, independently of aesthetics and relevance. While this allows users to curate what they hear [26], it also impacts situational awareness [67] and has distinct safety risks [28] when used in mobile contexts. In the current state of headphone technology, users have to cope with the drawbacks of this augmentation.

Emerging devices, called hearables, claim to be able to alter users’ auditory perception through a permanently worn digital device, providing an auditory mediation of the perceived reality. In theory, they allow to filter particular unwanted sounds and set up a specific hearing profile, comparable to an equalizer for real world sounds. While this is a compelling vision and direction, practical requirements for such devices and interactions have not been determined yet.

In an initial step towards understanding the requirements for interactive auditory mediated reality (IAMR), two studies were conducted: 1) A week-long event sampling study with 12 participants, followed up by semi-structured interviews. 2) An interaction study comparing audio-only and screen-based interaction concepts conducted in virtual reality (VR) with n = 12 participants. In the event sampling study participants recorded and rated 225 sound sources in their environment which they would like to alter the volume of. The semi-structured interviews aimed to further explore the concept of interacting with the surrounding soundscape and corresponding usage scenarios. Based on the insights of this study, we developed and implemented two interaction concepts for real-world sound source manipulation. In the second study those concepts were evaluated in terms of usability and workload.
The results of the first study indicate that besides muting, a distinct, "quiet-but-audible" volume exists, which caters to two requirements at the same time: aesthetics and information acquisition. For the second study, in terms of usability and workload, no significant differences were found for the two vastly different conceptual modalities of audio-only and screen-based interaction. However, we identified a set of four volume states to be sufficient for most scenarios: "muted", "quieter", "louder" and "unchanged". We propose interaction flows allowing better device-independent application of sound.

To summarize, the main contributions of this work are:
1. Usage requirements and patterns of potential users, which include a possible reduction of states, and usage contexts.
2. Specific insights for the system design of IAMR, such as workload of audio-only and screen-based implementations.
3. Implications for the design of future IAMR systems.

**INTERACTIVE AUDITORY MEDIATED REALITY**

The concept of augmented, interactive hearing relies on three fundamental concepts: mediated reality, human augmentation, and hearables. The following sections introduce and define the relevant aspects of these concepts and ground them in previous research and products, concluding with a definition of IAMR which result from the basic concepts discussed. This concept of interacting with surrounding sounds on a per-source level was introduced and discussed in few other works [102, 119, 110]. We aim to build upon these works and further formalize the underlying concepts and nomenclature, along with applicable interaction patterns.

Mediated reality is the result of relaying and/or manipulating human sensation via a device. Usually, this involves real or virtual content being recorded, processed and played back to a user. The term mediated reality was coined by Steve Mann [75, 78]. He describes it as a framework which "combines augmented and diminished reality" [77], allowing deliberate addition as in augmented reality (AR) but also alteration and removal of content from any human sense. As a "smart" device is inserted between world and user, the device receives control over the user’s sensation and perception. Conceptually, it allows any kind of sensation for the user whether they are somehow grounded in reality or not.

Inserting a device between world and user is a necessary prerequisite for mediated reality [62, 41, 75]. This applies to vision, as in AR glasses, as well as in hearing. As in vision, mediating sounds does not only include the addition of content. Through digital mediation, every single aspect of sounds and soundscapes can be altered or added, potentially creating a substitutional virtual world. Sound is recorded with a microphone and relayed to the user via headphones (i.e., via hearables or any other personal audio device). Similarly to visual AR, auditory AR primarily deals with the addition of content to users’ perception. It does not intend to replace the real world, but to enrich it with information [14]. When mapped to two axes (Mediation and Virtuality) as done by Mann in [75], four clusters can be identified (see Figure 1) and presented using exemplary systems.

Augmenting human intellect and abilities has been a recurring topic in general literature and research [36, 84, 76]. Douglas Engelbart was among the first ones to frame this process and concept towards computers as tools or a way to augment human intellect [36]. With the help of devices, physical and cognitive capabilities of users can be restored or expanded. It is even possible to add new capabilities, either through bodily integration/implantation [23], re-mapping senses [51], adding entirely new perceptual dimensions or changing the body’s morphology [64, 63].

Nowadays, augmenting is often conflated with the sole addition and superimposing of information. In a sense, added information or easier access supports and leads to better cognitive abilities and is therefore an augmentation of human abilities. However, removal or control over real and virtual information is equally augmentative. Augmentation can also happen through diminished reality, which removes content from the real world [76] and may make way for increased cognitive performance.

Hearables as the third concept are a device class of wireless earbuds providing functionality like voice assistance and fitness tracking [40, 96]. With hearables, auditory perception can be mediated by passing the real world sounds through a microphone and signal processor before playing back to the user. It allows the device to add to or alter a percept in any fashion and therefore enables personal soundscape interaction as opposed to global soundscape interaction (e.g., adding soundproof barriers along a highway). This concept is already implemented in some products [21, 33] in a rudimentary way (e.g., frequency-based audio-filters).

Mediation, in turn, opens up a large design space for addition [19, 86, 71] and alteration [75, 111]. Additive mediation is known as audio/auditory AR, whereas alteration can be considered modulated reality [75, 78]. This fundamental concept is the prerequisite for any kind of change to users’ abilities and perception. Ultimately, users would gain full agency and control over their sense of hearing by being able to choose and filter what they hear. This type of ability enhancement was proposed for vision in earlier research [75, 77, 76]. Hearing, being an important but less dominant human sense, was focus of research to a lesser extent. Nevertheless, the sheer usage and widespread adoption of personal audio technology indicates that there is a need to control the sense of hearing. This kind of control may go beyond established means.
of restoring hearing with hearing aids. Current coping strategies like wearing headphones are flawed and, in part, dangerous: they drown out relevant and irrelevant sounds alike [44]. Risks arising from the use of personal audio devices include social isolation [50, 27], dangers in traffic [67], and acquired hearing loss [28, 61], all originating from the desire to curate what one is able to hear. Ideally, IAMR is able to alleviate the risks of personal audio while giving users an entire gradient of isolation which is not uniform, but relies on a more selective and fully interactive mediation. To conclude we define the concept of IAMR based on the definition of mediated reality by Steve Mann: IAMR refers to a general framework for user-driven artificial modification of human auditory perception via devices for augmenting, diminishing, and generally altering the sensory input of the auditory system.

A conceptual process for IAMR with a device that provides (auditory) scene understanding and interactive alteration of hearing is depicted in Figure 2. The system segments the soundscape to sound sources, which the users may initially analyse and understand further. Afterwards, risks and requirements can be weighted. With this estimation in mind, users may start to add media or isolate. These two steps are possible with established personal audio devices, like headphones [44]. If conducted with an IAMR system, users have further options for personal human-soundscape interaction: they may start to selectively alter specific sound sources, for instance reducing their perceived volume or removing them from their personal soundscape entirely. The result of the IAMR process is an altered and curated soundscape, which may include any mixture of real-world and virtual sounds.

**SCENARIOS FOR IAMR**

Personal audio technology allows users to choose what they hear on a personal level [44], but is limited to uniform, almost binary interaction with the soundscape. With devices like hearables, the curation of the auditory environment is made possible to all users without the drawbacks of common personal audio devices. Personal soundscape curation then includes not only adding media and drowning out the world, but covers an entire gradient of level and other dimensions. Just like an aesthetic soundmark1 [104, 107] (e.g., a fountain) can be added to a physical soundscape, it can also be added to a virtual one on demand. Notably, the addition of a fountain to the physical environment (i.e., altering the soundscape on a global level) is not easily possible and requires lengthy processes in city planning. With the rise of various machinery, sound levels rose, displacing the sound of nature [95] and silence itself. However, in biological and physiological terms, humans did not adapt as quickly – their sense of hearing is largely the same as before. They were given a tool to drown out sounds on a personal level with devices like headphones. In the following, we present scenarios for IAMR systems, where comfort, media playback and information acquisition are weighted differently, depending on the users’ requirements and context. We consider established devices like headphones to be a precursor to IAMR, which are able to fulfill some, but not all potential user requirements. This is reflected in the presented scenarios, and follows the interaction flow of IAMR seen in Figure 2.

**Scenario I: Open Plan Office**

In today’s work culture, open plan offices are a common occurrence, requiring workers to cope with unwanted sounds [1, 43]. At the same time, communication between colleagues remains a relevant task. If a worker requires auditory comfort, she may isolate herself by using active noise cancellation (ANC) headphones and adding calm background noises. Alternatively, she may selectively remove the sound sources that are irrelevant and disturbing. If she desires media as an addition to her personal soundscape, she may use headphones, potentially drowning out sounds around her. Lastly, if she desires to retain information in her personal soundscape, she may leverage IAMR to exclude specific sound sources (e.g., colleagues not on her team) from the noise filter to be able to get information about their state without specifically devoting attention to them.

**Scenario II: Commuting**

As with open plan offices, commuting is a relevant part of today’s work culture, and is likewise a context where mobile audio devices are common [68]. If a user requires comfort, he may similarly isolate himself by using noise cancellation and drown out remaining background sounds with personal media. Alternatively, earplugs may suffice for ensuring comfortable levels of noise. If he desires to add media to his personal soundscape, the user may consume videos, music, audio books on the way home with the help of a mobile device and headphones. If the goal is to isolate while retaining the informative content of the auditory environment, the user may add speaker announcements to a list of desired sounds to avoid missing the correct station.

**RELATED WORK**

The concept of IAMR, which was also discussed in [102, 119, 110], relies on theoretical and practical advancements in research and industry. Apart from ongoing research in academia, the space around augmented and altered hearing is populated by industry players and crowdfunding platforms. Not all products achieve widespread adoption or reach a state of market-readiness. However, many reflect the vision of human-soundscape interaction and IAMR or parts thereof.

**Human Perception and Sensation**

Users’ perception of (spatial) sound is crucial when attempting to create a method that mediates and alters it. Research on sensation

---

1Equivalent to a landmark for audition
and perception is particularly relevant for filtering methods and helping users understand their soundscape [22]. The field can be separated into research on spatial perception and localization, spatial awareness and segmentation. Additionally, effects of noise and the perception thereof are related to the motivation and development of mediating devices.

Spatial perception and localization are relevant as the interface in world-fixed interactions fully surrounds the user. In 1967, Batteau showed that the auricle of the human ear is relevant for localization of sounds as it introduces delays to the signal propagation [13]. Furthermore, hearing is an omnidirectional sense in humans and therefore able to compensate for missing visual sensations. Sound augmentations like warning tones have been applied in vehicles [114] or general workspaces [29, 42, 4]. On a psychophysical level, humans are already able to segment and cluster the acoustic environments surrounding them [22]. They can also leverage phenomena like the cocktail party effect can also be leveraged to support soundscape understanding.

Interacting with Audio
A minority of previous research incorporated sounds as interaction targets, instead of it being a supporting feedback modality. In this case, the users’ goal is to alter sound and they interact with sound, instead of with the help of sound. Work in this field is related to tasks like music composition, mixing and editing [38, 88, 16]. Adams et al. presented a system that allows exploration and alteration of arbitrary parameters of sound on a touchscreen [2]. In contrast, Mueller et al. presented a system for spatial interaction with sound in the shape of an experimental mixing interface or room [88]. Comparable, mid-air interaction with sound was presented by Alroe et al. [5], Dicke et al. [32] and Gelineck et al. [39]. In contrast, interfaces that allow exploration, search and organization of sound, are not necessarily sound-based or sound-focused [105, 16, 99, 112].

Personal Audio Technology
Apart from industry-driven advances, there is ongoing research concerning personal audio technology and devices, particularly with regards to sociological and societal aspects. Mamiuji et al. presented “attentive headphones” to manage interruptions and mediate communication [74]. A similar approach was used by Danninger et al. in an office scenario [31]. Tyler Bickford observed children’s behaviour around headphones and earbuds like sharing a pair of earbuds or using them as speakers, transforming them to non-personal audio technology [17]. Social behaviour around sound and sharing thereof (e.g., music recommendations) was investigated by McGookin et al. [83, 82] and Hakansson et al. [45] on a larger scale. Michael Bull described modern use and implications of personal audio [25, 26], followed by Liikkanen et al. [68], Haas et al. [44] and Nettamo et al. [89] among others. Technical aspects like performance of noise-cancellation [7, 66], signal processing [10] or new form factors [40, 96, 52] are part of the progress in research. While there are hardware platforms available that originate from and are aimed at academic research [56], much work is primarily industry-driven and is described in more detail below.

Auditory Augmented Reality
Beside the well-established field of visual AR [11, 18], projects focusing on hearing also exist. Core drivers are additive systems attempting to augment and superimpose virtuality over the real soundscape. While the concept envisioned in this work is focused on modulation and alteration instead of addition, the technical side remains the same. Auditory AR is most relevant to the proposed work, as methods and insights from systems can be transferred to the prototypes developed. It can further provide detailed reasoning on directions to prioritize or discard early in the process. Technical fundamentals include the use of hear-through technology [85, 79, 69, 87], bone conduction [69, 81] and the implementation of spatialization [80]. Auditory AR was also a focus in games research, with co-located [60] and location-based [94] games. Application-centred research was conducted in guide [14, 71, 113] and navigation [118] systems.

Products and Industry
Personal audio technology and its advances are nowadays primarily driven by industry [52]. It encompasses new developments in terms of headphones, hearables and earplugs, along with progress in filtering and noise-cancellation. Some headphone-like devices are already equipped with spatial filtering [20, 92], whereas others attempt to blend real and virtual worlds: for example with the Xperia Ear Duo, users are meant to “Stay in tune with the world” [106]. As with other recent products, voice assistants are a selling point [117].

The market of hearables at its core consists of mostly crowd-funded startups. Doppler Labs’ “Here One” comes closest to the vision of interactive hearing, providing presets, equalizers, and blending functionality with their headset [33]. A similar approach is taken by Qbuds [90], with other vendors focusing on fitness tracking [21].

Hearing aids are already an established way to recover and restoratively augment hearing [65, 58]. While they improve auditory perception, it is important to differentiate “audibility and intelligibility” [65], as not all of them are able to restore the ability to perceive speech as before. However, current devices provide situative presets [93], are controllable by the wearer and may even include smart adaptive functions that require no intervention from the user [49]. Some allow the streaming of media to them and have dedicated apps for controls [58, 8]. The intersection between personal audio and hearing aids is the core domain for future developments in IAMR.

SOUNDWALK STUDY
To develop a system that is able to manipulate perceived ambient sounds, we first considered the needs of users. Therefore, we conducted a week-long event sampling study with 12 participants. They were instructed to record and rate sounds that they would like to make louder or

Methods and Apparatus
To gain insights into user requirements in real-world environments, a mobile application was created which allowed users to record and rate sound sources which they would like to make louder or
quieter. As this ability, at most, is currently only part of niche products, users are currently not actively aware of use cases. This is particularly the case for alterations that can not be executed with common personal audio devices, like fine-grained, selective alterations. The methodology is comparable to crowd-sourced research on noise-pollution [73, 98] and is also inspired by Schafer’s concept of “soundwalks” [104]. However, when surveying relevant literature, a core problem became evident: most soundscape research focuses on the soundscape as a whole, not on specific parts of it.

Choosing this method allows involving users in a design process for a future artifact with the requirements and chances being largely unknown to users and developers. Furthermore, it is likely to yield more natural results as users record data whenever they deem it necessary and do not have to remember specific situations. Users would also survey their surroundings primed towards the ability and chance to selectively change sounds, something they usually do not have. Additionally, there is no comparable data set as earlier research with similar methodology [3, 57, 34, 103] had a different framing. When attempting to gather data in-situ, a trade-off need to be made between the amount of data to gather and the number of samples received. The general goal was to enforce brief interactions with the app, nudging participants to collect many entries instead of few, detailed entries.

With the mobile application we provided, participants were able to record a brief sample of the sound they wanted to alter and report at which level they want it to be. The “desired level” ranged from 0 (“inaudible”) to 10 (“as loud as possible”). Furthermore, the application provided a user interface to label and rate those sounds on 7-point scales, derived and adapted from [55]: Naturalness, Pleasantness, Regularity, Proximity and Directionality. We also provided multiple context flags such as working, leisure, social, solitary, public, private, commuting, at home, relaxing and focusing that could be selected via checkboxes.

As users could freely decide to enter data, the application was meant to nudge users gently while not being annoying or overly present. A persistent notification was therefore shown to the user. Tapping it immediately opened the activity for data entry.

Procedure

After the recruiting process, participants were invited for the introductory session consisting of application installation, explaining the target vision of IAMR, and giving instructions for the following week-long data acquisition. To verify the applications’ functionality, participants were asked to record a sound with the examiner present, listen to it, and delete it afterwards. The final meeting, approximately a week after the introduction, consisted of multiple parts. Participants were given the chance to listen to and delete sensitive recordings. This was followed by a semi-structured interview, which required it’s own informed consent. The reward given to participants was partially scaled with their “performance”, which essentially was the number of completely entered data points. The base reward was set to 7 currency, as it was estimated that the initial and final meetings would take roughly 40 to 50 minutes. For each recorded source, the participant would receive 0.4 currency, up to a maximum of 6 currency of added reward.

Results

The participants were aged 27.83 years on average (SD = 2.7, min = 22, max = 31), with exactly half identifying as female, the other half identifying as male. 5 were employed, 6 students (either undergraduate, graduate or PhD) and 1 reported being currently out of work. 7 held a master’s degree, 3 a bachelor’s degree and 2 a high school degree. 10 participants reported having had no diagnosed auditory impairments. The 2 remaining ones reported having had a sudden temporary hearing loss and having hearing loss in a specific frequency band which did not affect them in daily life. In total, 225 samples were collected by the participants.

![Graph showing action classes and the distribution of chosen volumes](image)

(a) Counts of all action classes applied (b) Distribution of the level set by the participants

Figure 3: Action classes and the distribution of chosen volumes they are composed of.

For each source or source cluster, users set a volume level they want it to have. The full distribution can be seen in Figure 3, with estimated densities overlaid over a histogram. The histograms bins are equal to the possible level values users could select. A large cluster is found at the mute value. However, a large amount of sources was made quieter, without fully removing it from the soundscape. Additionally, the level below the “unchanged” value was not used at all. Increased values are more sparse and are mostly found in the middle between the maximum level and the unchanged state.

For all data points (n = 225), the desired level/volume ranged from 0 to 10, covering the entire value range (M = 1.7, SD = 2.7, SE = 0.18) with a median absolute deviation of 1.48. Far more sounds were decreased (n = 193) in volume than increased (n = 32). However, in terms of contribution to a soundscape, decreased volume is vastly different from completely silent sounds. When splitting the group of decreases in “decreased” and “muted”, a more balanced separation arises. The number of increases is unchanged (n = 32), with muting (n = 109) exhibiting a slightly higher sample count than decreasing (n = 84), as can be seen in Figure 3.

The sounds users labelled were overwhelmingly associated with objects. We therefore applied an established coding and classification scheme to them which was introduced by Schafer. It covers natural sounds, human sounds, sounds and society, sounds as indicators, mechanical sounds, and quiet and silence [104]. Distribution of labeled events can be found in Figure 4.

As mentioned before, users had to rate the source or sources they want to alter on 5 scales: pleasantness, naturalness, proximity,
directionality and regularity. The rating on the unpleasantness-scale ranged from 0 to 6, covering the entire value range ($M = 4.3, SD = 1.74, SE = 0.12$) with a median absolute deviation of 1.48. The ratings on the unnaturalness-scale also covered all possible values ($M = 3.62, SD = 2.44, SE = 0.16$) with a median absolute deviation of 1.48. The distance-ratings also ranged from 0 to 6 ($M = 3.65, SD = 1.89, SE = 0.13$) with a median absolute deviation of 2.97. Directionality ranged from 0 to 6, again covering the entire value range ($M = 2.12, SD = 2.12, SE = 0.14$) with a median absolute deviation of 1.48. The rating on the regularity-scale covered all possible values, too ($M = 2.67, SD = 2.04, SE = 0.14$) with a median absolute deviation of 2.97.

**Interviews**

The interviews took 13:19 minutes on average ($min = 6:56 min, max = 28:31 min$), dependent on the participant’s experience and desire to discuss the concept. After selective coding (thematic analysis), specific themes could be identified, which recurred for almost every participant. The following paragraphs summarize them.

**Dangers and Risks** The illusion of an intact auditory perception mediated through such devices was a criticism of the concept. This means that changes made to the auditory perception are no longer conscious for the user and therefore be forgotten. Furthermore, missed information, as already the case with personal audio devices, was linked to this aspect. Additionally, adaptation and overreliance emerged as issues.

**Dynamic Requirements** Participants repeatedly mentioned that besides a changing soundscape, their own requirements may change, depending on mood or context. For a future device, it is reasonable to assume that it can accommodate for some of these changes, but not necessarily all.

**Interaction and Interactivity** With the aforementioned aspects of dangers and changing requirements, a degree of interactivity was required by participants. Generic personal audio devices are already interactive, allowing alteration of media playback, but also physical interaction, like taking them off. This would not be the case for permanently worn hearables. Being able to interact also covers cases, where automation is insufficient or certain requirements change.

**Agency Over One’s Own Perception** Considering that mediating devices may automatically alter and mediate perception itself, users reaffirmed their scepticism against automation. While headphones are an alteration of hearing, this alteration is deliberate and conscious. No other device currently mediates perception in an automated fashion, especially not in a subtractive or destructive fashion.

**Subtlety and Social Aspects** Most participants referred to the alteration of human speech in one way or another. Increased volume for better understanding is considered a noble goal but can be used for malicious eavesdropping. Similarly, muting people, especially without them knowing, is unfavourable on a social level. For some participants, these problems called for very subtle and unobtrusive interactions. For others, this required an enforced, system-side ruleset.

**Limitations and Discussion**

Apart from the sole fact that per-source soundscape curation has its uses and potential, additional requirements can be derived from the interviews and data. Whenever a source is altered, the effect has to be clear to the user. This allows them to weigh advantages and risks appropriately. Furthermore, systems may leverage inherent abilities of users, as humans are already able to segment a soundscape into sources. Leveraging selective attention to define filters would make IAMR open to users. While IAMR may provide more comfort, filtering relevant sounds is a source of danger. Muting has to be treated carefully, especially if auditory comfort can be reached without total removal of sounds.

If given the chance to change the volume of any sound source, users mostly reduce or mute and rarely increase the level. This is a chance for simplification of all future interfaces, as continuous volume control may be reduced to few “states”, like muting. Additionally, it can be assumed that there is a level between “muted” and “unaltered”, which retains awareness and completeness of the soundscape, but caters to users’ desire for silence and acoustical comfort – another chance for simplification. Simplification is ultimately a procedure to reduce interactions needed and therefore reduce the time required to interact and make it applicable for a wider range of devices.

Users identified sound sources predominantly as objects, rarely as textures and never by their acoustical properties. This makes frequency-based interfaces like equalizers unfeasible for everyday use. Therefore, interfaces should rely on understandable labels, as these may already describe the level of granularity required for alterations.

**INTERACTION STUDY**

Based on the insights of the first study, a more concise evaluation was designed covering concrete aspects of future IAMR systems. A study in which participants actually interact with a system and are able to perceive the effects is likely to yield valuable feedback concerning the concept and vision, independently of the specific implementation. The first study conducted yielded a set of requirements and suggestions by users, as induced with their active examination of their everyday soundscapes. Most suggested screen-based interaction, but also required subtlety for it. Additionally, due to the way the desired volume was distributed and with the requirement for brief interactions, it was hypothesized that the granularity of interaction can be reduced or compressed, without a loss in usability. Based on the interviews and the above findings, the evaluation was set to study two interaction concepts (screen-based and head-tilt-based) and two manipulation granularities (discretized and continuous). The **screen-based** mode represent interactions with touchscreen devices like smartphones or tablets, whereas the **head-tilt-based** mode represented spatial...
interaction. Tactile interfaces such as a mixing desk were not considered as they are not applicable for mobile scenarios.

**Application Design**

Conducting an evaluation of interactions in VR is valid, as long as these interactions are either meant for VR, or are reproduced faithfully. Additionally simulated auditory scenes are a relevant topic in research, especially concerning their faithful reproduction [15, 37, 100, 109] and usage for analysis [101, 35, 72, 54]. Interaction with a small screen, like a smartphone or even a smartwatch, is challenging to reproduce faithfully in VR. It requires the users’ fingers to have a dexterity close to their real abilities. Furthermore, small text, as used on such devices, requires high resolutions to remain legible. Generally speaking, each interaction tested in VR suffers from some degree of indirection. This applies to (touch-)screen-based interaction in particular, but also influences gestural interfaces to some degree. However, an evaluation in VR has distinct benefits other methods can not provide such as full control over the world, which was important for this investigation.

**Apparatus**

To execute the study in VR, an appropriate environment was developed and designed. A top view of the scenery with sound sources is shown in Figure 5. It was meant to be compelling and visually consistent. Therefore, a pre-made asset was used. To avoid overloading the user with a large amount of sounds, yet encourage users to interact with the system, all sounds were introduced following a scripted timeline (Figure 5). To ensure proper spatialization and room effects, the Resonance Audio library was used. It supports spatialization by interaural time and level differences, head-related transfer functions (HRTFs), frequency-dependent effects, reverberation and occlusion. To ensure believable reflections, the reverberation baking functionality was used, which mapped visual materials to acoustic materials and properties.

**Variables and Study Design**

The potential design space around IAMR covers a large set of dimensions, ranging from device use over interaction modality to treatment of specific sound source types. The first variable, Concept / Paradigm was chosen to cover two extremes in terms of modality: screen-based (close to visual-only) compared to head-tilt-based (close to audio-only). The second variable is Granularity / Complexity, which again covers two extremes in terms of manipulation. A simplified granularity provides minimal discretized steps or states for alteration. This subset of volume consists of a muted state, a quieter state, an unchanged state and a louder state, as derived from the first study. In contrast to this reduction, a method of continuous alteration allows for a potentially infinite amount of states.

**Screen-based**

Screen-based modes are meant to represent interaction via a surrogate device, like a smartphone or a tablet. For each sound source that is present in the scene, an element is added to a tablet-like interface the user is holding with his hand. As all elements are listed, interaction happens on an ordered aggregate of sources (Figure 6). Whenever a new source is added to the soundscape, it is placed above all others in the list, along with a visual highlight (blinking). The continuous mode (Figure 6, d), consists of 3 interactive modules per source. A slider covers the volume range from 0 (mute) over 0.5 (unchanged) to 1 (maximal), with the numeric value hidden from the user, as done by most other interfaces for sound. Additionally, 2 shortcuts to specific states are present, as suggested by users in the first study: reset the volume to it’s original state and mute the source. The on-screen discrete or simplified mode (Figure 6, e) operates on state. Each source offers 4 buttons, representing a muted, a quieter, an unchanged and a louder state.

**Head-Based**

The two modes for head-based interaction are implemented as audio-only systems. While this is not necessarily optimal, this depicts a fixed extreme point in the possible design space from which other prototypes may be derived. The design itself was derived from earlier prototypes mapping audio from or to sight [46, 47, 115] and interaction techniques relying on head roll or yaw [30, 53, 108, 91]. Sources are targeted with head rotation and a sphere-cast along the users’ head orientation. Alteration is mapped to the head’s tilt or roll, with a knob-like metaphor. This virtual knob then either has a continuous mapping, or has an ordered set of states covering the heads tilt range. Figure 6 a-c depicts the angle ranges and the functions mapped to them. The discrete head-based mode functioned similarly, but discretized input to a set of four states, as can be seen in Figure 6 b. The edit mode is reached in a similar fashion as the screen-based variant. To ensure comparability, both modes were enriched with a set of supporting functionality. For the head-based modes, the “labels” of targeted sources were read out to the user via text to speech (TTS). Similarly, the discrete states of each mode (e.g., “mute”) were also read out to the user.

**Procedure**

The study took place in a separate room of our institution. Participants received 8 currency and a bar of chocolate as compensation with the entire study taking roughly 60 minutes. For each of the 4 study iterations, the participant received a brief explanation of the mode they are about to use. The 4 iterations were balanced in a Latin Square. The examiner relied on 4 cards, depicting the mode, with the help of which he showed the core part of the mode. Additionally, any VR-specifics were explained and demonstrated such as the activation gestures and which buttons are to be pressed for them. Participants wore Sony WH-1000XM2 ANC headphones. The VR-headset used was an Oculus Rift CV1. Users first entered a training scene, where 4 sound sources were placed in a neutral environment. The 4 sources were visually represented by grey cubes. 3 of the sources emitted irregular tones played on a piano, while the remaining one constantly played a 440 Hz sine wave. Whenever participants felt comfortable, the examiner loaded the real environment. The participants then received 5 minutes to configure the soundscape to their liking, while discovering the method. Additionally, they were encouraged to think aloud.

Each iteration was followed by a questionnaire, containing the NASA TLX [48], system usability scale (SUS) [24] and selected single-item questions on 5-point Likert scales. Additionally, users were provided with 3 comment fields to express likes, dislikes and general comments concerning the method they just used. After finishing all 4 trials, users received a final questionnaire with rankings, general questions and demographic data.
The 12 participants were aged 28.58 years on average ($SD=2.55$, $min=25$, $max=32$), with 5 identifying as female, the remaining 7 identified as male. 10 participants reported having had no diagnosed auditory impairments. One participant reported having had a temporary sudden hearing loss and one reported having hearing deficiency in a specific frequency band which does not constrain him. Their self-assessed, estimated average usage time of personal audio devices per week was 24.33 hours on average ($SD=17.3$, $min=2$, $max=56$).

**Results**

The following sections describe and evaluate the quantitative and qualitative results gathered in the study.

**Load and Usability**

To assess the users’ mental load and usability we used two standardized questionnaires. The NASA TLX [48] was meant to assess task load, while the system usability scale (SUS) [24] was meant to assess the usability of the concept. As mentioned before, users were asked and instructed to abstract from the specific implementation and from constraints imposed by the VR environment.

Core takeaways from the SUS scales are that granularity has no great influence on the perceived usability of the concepts presented to the users. While the median value was the same (both $Mdn=71$), the continuous granularity exhibits a larger variance in values, especially towards the lower end of the scale. This may be interpreted as an indicator that a metaphor of a continuous slider is overly complex for the use-case of sound alteration in most everyday scenarios. When comparing the SUS scores by paradigm, screen-based modes yield better results, with the interquartile range remaining above the marked score of 68. This can be ascribed to multiple aspects: aggregating sources (on a screen) makes them easy to interact with and it simplifies the task of finding them. The localisation task, especially when relying on audition only, adds another step of indirection, while being able to spatially understand and segment a soundscape becomes less relevant.

Results of the TLX are in line with the answers provided to the SUS questionnaire. High effort needed for the head-tilt modes can be primarily traced back to the difficulties users had localizing and selecting sources. This is likewise reflected in the performance item. Physical demand was less differentiated between the modes. Furthermore, the simplified head tilt was slightly less mentally demanding. This may be traced back to the the fact that dealing
with an invisible state was easier in this mode: a selection of a source did not necessarily immediately alter the sound. For all other items, comparing granularity yields minimal differences. This indicates that reducing interaction complexity by reducing manipulation complexity is a viable path for practical systems.

As the SUS questionnaire yields a comparable score and each sub-scale is actually related to the system’s operation, it was used for further quantitative analysis. When comparing each combination, differences are visible, primarily between modes and not between granularities. When assessing all gathered SUS scores, the mean score is found at 70.05 while the scores cover the range from 55 to 85 (SD = 7.21, SE = 1.04). The Shapiro-Wilk normality test indicates that the data does not significantly deviate from a normal distribution, with $W = 0.97$ and $p = 0.26$. Ratings gathered from the "Head Tilt with continuous change" mode had a mean SUS score of 65.63, ranging from 55.0 to 82.5 (SD = 8.19, SE = 2.37). For "Head Tilt with discrete change", ratings had a mean SUS score of 67.71 and ranged from 57.5 to 82.5 (SD = 6.44, SE = 1.86). "On-Screen with continuous change" had a mean SUS score of 72.08, which ranged from 57.5 to 80 (SD = 5.92, SE = 1.71). Similarly, "On-Screen with discrete change" had a mean score of 74.79, ranging from 67.5 to 85 (SD = 4.70, SE = 1.36).

To determine whether the differences in mean and range are significant, a Friedman rank sum test was conducted. The Friedman rank sum test (Friedman’s ANOVA) determined that there were significant differences between the scores of the conditions, with $\chi^2(3) = 15.7$ and $p = 0.001306 < 0.005$. As a follow up, multiple comparisons were executed, which indicated significant differences in the SUS score between the On-Screen with discrete change and the Head Tilt with continuous change condition and between the On-Screen with discrete change and the Head Tilt with discrete change condition. Additionally, pairwise comparisons using the Nemenyi test with $q$ approximation were executed [97], yielding similar results with $p = 0.0016 < 0.005$ and $p = 0.0142 < 0.05$ respectively.

Preferences by Context

Four contexts were chosen to cover most application scenarios in an abstract fashion. A private context reduces social interaction and conveys a sense of security and safety, whereas a public context may have social implications and requirements. A static context exhibits no or limited changes in the soundscape and does not involve movement of the user. A dynamic context, in contrast, may involve movement and change of the environmental soundscape and the users’ themselves. Having experienced all concepts and granularities, users were asked to rank the desired usage for each of these contexts, with 2 ranked items being mandatory.

A clear pattern over all 4 rankings is the preference of screen-based interfaces. In a private context, continuous change is preferred over discrete (5 and 7 mentions respectively). This is reversed for the public context, where the simplified version has nearly twice as many votes (8 and 3 respectively). A similar pattern emerges when comparing the static and dynamic contexts, where the static context has 6 mentions of the continuous change, compared to 4 of discrete change. The dynamic context has 8 mentions for the discrete change and only 4 for the continuous mode. This can be traced back to the desired brevity and speed of the interaction. With a simplified user interface, a desired state can be reached quickly, albeit not as precisely as with a fine grained control. This aspect of speed and precision explains the differences found in the rankings. While a fine-grained alteration is necessary to retain, a simplified may suffice in many contexts, where neither the time nor the desire to configure sounds precisely is present.

Source Ordering and Priority

Users were additionally asked which criteria they would employ to order and prioritise sources on aggregated interfaces (i.e., screen-based lists). The criterion most commonly ranked at the first position, was volume (6x). This is in line with the dimension for acoustic comfort but also indicator for safety and urgency [6]. It was followed by distance (3x), an indicator for the feasibility of spatial filtering, as it is again an indicator for relevance and potential danger. The second rank is lead by annoyance (4x), a subjective criterion relevant for acoustic comfort and followed by relevance (3x). The third and last mandatory rank is lead by relevance (4x), followed by distance and pleasantness (3x each). The remaining criteria, including recency, as implemented for the screen interfaces in the study, were mentioned sparsely across all ranks. Notably, these preferences may be transferable to spatial user interfaces where occluding and intersecting sources require some sort of arbitration when they are being targeted and selected. Likewise, these features may serve as additional input for automated or partially automated systems if consistent rules can be derived and modelled.

Interaction Concept and Granularity

Most importantly, 100% of the participants agreed with the statement that 4 states are sufficient for most use-cases. This again confirms the potential of reduced manipulation granularity for optimisation for faster and dynamic contexts and environments. Additionally, 67% of the users tended towards agreement when offered that a fine-grained alteration is still possible, with 33% remaining undecided.

Furthermore, users predominantly agreed with the notion that they would configure their entire surrounding (67%) while disagreeing with the idea that they would alter only few, select sources (67%). The latter is in line with the process of soundscape interaction via personal audio devices, like headphones. The concept itself was met with acceptance, with 83% disagreeing with the statement that they are not interested in changing the sounds around them.

Limitations

The study design focused on two specific dimensions future systems can have and did not incorporate nuances or mixtures of these extremes. Users were additionally put in an environment they were not familiar with. This impacted their ability to locate sounds, as contextual information was not available to them. In a real home-scenario, occurrences and mappings of sounds would likely be known to the users, simplifying their process of locating and understanding. Despite these limitations, it was possible to receive valuable insights on a conceptual level, invaluable for future iterations.

Discussion

The core result of the evaluation is that a reduction of alteration complexity is not only accepted by users, but also preferred and considered more appropriate for dynamic scenarios. This insight is likely applicable independently of modality or device, as it is
based on the interaction between human and soundscape and less in the interaction between human and device. Another takeaway is the aspect that an audio-only system, despite its theoretical advantages, is hard to develop and requires thorough testing and iterations. Despite the lower performance of both head-tilt prototypes, the idea of a natural and rather subtle mapping was accepted and understood by the participants. It is also reasonable to assume that visual AR can serve as supporting or core modality, if it is desired to retain the spatial aspect of interaction with a soundscape.

If a spatial reference is needed or desired, metaphors aggregating sources along with spatial information may be a reasonable direction for future evaluations.

**IMPLICATIONS FOR DESIGN**

The two studies yielded a large set of insights for future systems that build upon the concept of IAMR. Currently, personal soundscape curation is made possible via personal audio technology. Yet, their functionality is limited and the way they let users interact with the soundscape is uniform and almost binary. Haying gained more selective control over sounds opens up possibilities of human-soundscape interaction on a personal level. Based on both conducted studies and the quantitative and qualitative insights gained, we present a set of implications for such systems.

**Reclaiming and Retaining Agency** The concept of IAMR is meant to support users to re-gain some agency over their sense of hearing, without suffering from the disadvantages of established personal audio technology. An ever-increasing amount of noise in urban soundscapes can be regarded as a coercion to hear. Automating this process may be conceptually a similar coercion of the user to hear or miss specific noises. Likewise, automated selective alterations to the users’ acoustic environment may convey an illusion of an intact soundscape, whereas manual alterations may not. Interactivity, either based on manual mixing or the definition or selection of presets, is therefore an appropriate path to agency.

**Soundscape Understanding Precedes Soundscape Alteration** Any system aiming to allow users to alter their auditory perception has likewise to support the users in their task of understanding their acoustic environment. This is particularly relevant for dynamic environments or environments unknown to the users. Apart from filtering sounds alone, systems should provide labelling and other functions to the users to aid them in understanding the current soundscape and the consequences of their alterations. This improves the judgment of risks, but also the expression of the users’ requirements towards the soundscape.

**Spatial Interaction vs. Aggregated Interaction** Conceptually, interacting with sound sources in space may help users to segment and understand the soundscape, leveraging spatial and selective hearing. However, this introduces some degree of reliance on visual information. In the case of occluded sources or sources without a precise location, spatial interaction may either be inefficient or introduce additional issues. Interacting on an aggregated set of representations for sources therefore is applicable but may require the use of visual interfaces, like smartphones or AR-headsets.

**Appropriate Interaction Fidelity** Real-world sound does not have to be treated equal to media sounds. Treating a soundscape as a sound mixing system used in the industry is rarely appropriate and deters users from soundscape interaction and alteration. A large set of use-cases for interactive hearing can be covered by a greatly reduced set of states for volume: “quiet-but-audible” for aesthetic requirements and information acquisition alike, lightly increased volume for better understanding and muted volume for the removal of irrelevant sources from the soundscape. Long configuration is not desirable for dynamic environments. This reduced interaction fidelity is likely to suffice, while allowing for more brief and unobtrusive interactions.

**Meaningful Sound Source Granularities and Decompositions** It is also important to consider the granularity of sound sources to be interacted with. One can consider “traffic” to be a sound source. Alternatively, one can decompose it into its parts like cars, pedestrians. Likewise, it is possible to group sparsely occurring sounds into ambience as an interaction target, like the voices of different bystanders into “speech” as an aggregate. These aggregations or decompositions can be defined by semantics, spatial relation to the user or the relevance to the user.

**Harmless and Harmful Use** By providing heightened or even superhuman senses, issues of acceptance and privacy arise. Distinguishing sounds and their sources through selective playback in situations where it would not be possible otherwise is problematic. Furthermore, it is not intended that the system allows for eavesdropping on people beyond users’ natural ability for selective hearing. As it is the case with many intrusive technologies, acceptance must be discussed and weighted thoroughly.

**CONCLUSION AND FUTURE WORK**

Given appropriate tools, users may regain agency over their auditory perception. The intersection and interface between user and soundscape is currently non-interactive. For interactive mediated hearing, requirements are different from current soundscape design approaches.

The first study was able to uncover real-world requirements for augmented hearing. It yielded a large set of sources users wanted to alter, with various mechanical sounds irrelevant to users. Nearly all were referenced by object and not by high-level clustering or physical properties. Similarly, sources were made louder, to increase aesthetic qualities or to acquire additional information. Core insight however was that users reduce volume instead of muting sources to account for both acoustical comfort and information acquisition. The second study showed that a reduced manipulation granularity had no significant impact on usability and acceptance. In general, the development of specific systems and the evaluation of their effects in real-world scenarios is a goal to progress IAMR systems.

This work focused on one dimension of sound: volume. While it is the most graspable one to users, spatial position, frequency and various other factors are relevant for future research. Additionally, more specific interactions and visualizations can be implemented and evaluated with the proposed guidelines. Audio-only, visual-only and mixed-modality systems can be compared on a more refined level, based on the results gathered in this work.

**ACKNOWLEDGEMENTS**

This work has been supported by the DFG project "Empirical Assessment of Presence and Immersion in Augmented and Virtual Realities" (RU 1605/4-1).
REFERENCES


