SonicVista: Towards Creating Awareness of Distant Scenes through Sonification

CHITRALEKHA GUPTA, School of Computing, National University of Singapore, Singapore
SHREYAS SRIDHAR, School of Computing, National University of Singapore, Singapore
DENYS J.C. MATTHIES, Technical University of Applied Sciences Lübeck, Fraunhofer IMTE, Germany
CHRISTOPHE JOUFFRAIS, IPAL, CNRS, Singapore
SURANGA NANAYAKKARA, School of Computing, National University of Singapore, Singapore

Fig. 1. In contrast to a sighted person who is aware of their vista-space based on visual perception, a PVI has limited awareness of the vista-space that is surrounding them. The figure illustrates how a sighted person gathers information versus our proposed vision of creating awareness of the vista space for a PVI through scene sonification.

Spatial awareness, particularly awareness of distant environmental scenes known as *vista-space*, is crucial and contributes to the cognitive and aesthetic needs of People with Visual Impairments (PVI). In this work, through a formative study with PVIs, we establish the need for vista-space awareness amongst people with visual impairments, and the possible scenarios where this awareness would be helpful. We investigate the potential of existing sonification techniques as well as AI-based audio generative models to design sounds that can create awareness of vista-space scenes. Our first user study, consisting of a listening test with sighted participants as well as PVIs, suggests that current AI generative models for audio have the potential to produce sounds that are comparable to existing sonification techniques in communicating sonic objects and scenes in terms of their intuitiveness, and learnability. Furthermore, through a wizard-of-oz study with PVIs, we demonstrate the utility of AI-generated sounds as well as scene audio recordings as auditory icons to provide vista-scene awareness, in the contexts of navigation and leisure. This is the first step towards addressing the need for vista-space awareness and experience in PVIs.

Authors' Contact Information: Chitralekha Gupta, chitralekha@ahlab.org, School of Computing, National University of Singapore, Singapore; Shreyas Sridhar, shreyas@ahlab.org, School of Computing, National University of Singapore, Singapore; Denys J.C. Matthies, denys.matthies@th-luebeck.de, Technical University of Applied Sciences Lübeck, Fraunhofer IMTE, Germany; Christophe Jouffrais, christophe.jouffrais@cnrs.fr, IPAL, CNRS, Singapore; Suranga Nanayakkara, suranga@ahlab.org, School of Computing, National University of Singapore, Singapore.

ACM 2474-9567/2024/5-ART76
https://doi.org/10.1145/3659609

CCS Concepts: • **Human-centered computing → Empirical studies in accessibility**; **Auditory feedback**.

Additional Key Words and Phrases: People with Visual Impairments, Vista Space, Space Awareness, Sonification, Generative Models for Audio

ACM Reference Format:
Chitralekha Gupta, Shreyas Sridhar, Denys J.C. Matthies, Christophe Jouffrais, and Suranga Nanayakkara. 2024. SonicVista: Towards Creating Awareness of Distant Scenes through Sonification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 76 (June 2024), 32 pages. https://doi.org/10.1145/3659609

1 INTRODUCTION

According to the classification of psychological spaces provided by Montello et al. [61], a *Vista* is the far-field space that is projectively as large or larger than a human body but can be visually apprehended from a single place, without appreciable locomotion, which includes spaces in town squares, small valleys, and horizons. Sighted people are passively aware of scenes in their vista space [61], which are beyond an average hearing range, but within their visual range. While this spatial awareness is often taken for granted, it contributes to the cognitive and aesthetic needs of people (Figure 3). However, awareness of these distant scenes is not available to People with Visual Impairments (PVI)[1] (Figure 1). In this work, we study the usability of different types of sounds (audio recordings, handcrafted sounds, and AI-generated sounds) for vista space awareness for PVI.

Existing technologies for spatial awareness for PVI sonify individual objects in a scene using various kinds of sounds [34] such as *earcons* that include musical notes with varying pitch and loudness (e.g. beeps and tones) [73], sped-up speech icons called *spearcons* [16], and specially designed auditory icons that have semantic connections to the objects [15]. Although such handcrafted sounds are shown to be intuitive and learnable for communicating and detecting obstacles or objects in the immediate vicinity [15, 36, 41, 75], there is a lack of studies on sounds that could provide distant or vista space awareness to the PVI. Designing handcrafted auditory icons involves methods to model the interaction between different objects [26]. Such physical models are object-dependent and thus are not scalable and are less flexible when the number and types of objects increase, or when the vista-space scene description changes. With generative AI models for audio synthesis, we have an opportunity to generate soundscapes based on scenes in the vista space. However, to the best of our knowledge, the usability of AI-generated sounds for the purpose of vista space awareness is yet to be evaluated. Additionally, the sounds used in currently existing technologies such as *spearcons* are often associated with a high cognitive load [18, 41], which further hinders the usage of such sounds for a long duration of time and for multi-tasking scenarios.

In this work, we address the knowledge gap of the need for effective sonification mechanisms to enable awareness of scenes in the vista-space for PVIs. Specifically, we address the following objectives:

- Objective 1: Investigating the relevance and importance of vista-space awareness for PVIs
- Objective 2: Comparing the performance of existing sonification methods and AI audio generative models in producing intuitive and pleasant sounds for objects and scenes
- Objective 3: Evaluating the utility of sounds such as audio recordings and audio from generative models in enabling awareness of vista-space scenes to PVI

To understand the perspective of PVIs on vista space awareness, we conducted a formative study with seven participants with visual impairments. In order to investigate sonification mechanisms, we recruited 24 sighted participants for a listening test to evaluate and compare sounds designed by three methods - (a) handcrafted auditory icons or audio recordings, (b) AI model 1 - AudioLDM [57], and (c) AI model 2 - Im2Wav [79]. The evaluation was based on the intuitiveness of the sound being mapped to an object or a scene, the learnability of the sound-to-object/scene mapping, and the pleasantness and comfort level for long-term usage. We also

[1]We use 'PVI' throughout the paper. This abbreviation is meant for ease of reference and consistent with prior work [6, 7, 50, 52, 89]. We hope readers do not misunderstand this being disrespectful.

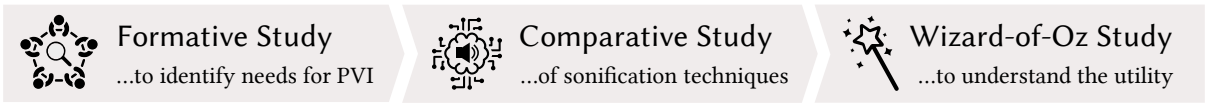


Fig. 2. In this paper, we conducted a three-phase study to address the problem of vista-scene awareness for people with visual impairments (PVI). In the first phase, we performed a formative study using semi-structured interviews to elicit the needs and preferences of PVI. In the second phase, we conducted a comparative study to evaluate the performance in terms of intuitiveness and learnability of different types of sounds, including handcrafted and AI-generated sounds. In the third phase, we carried out a wizard-of-oz study to investigate how the sounds affect the perception and understanding of vista scenes.

recruited seven participants with visual impairments to evaluate a subset of these sounds that performed well with the sighted participants to validate the findings. Finally, we conducted a follow-up user study to test the utility of such sounds in a wizard-of-oz experimental setup, where the same seven PVIs as the formative study experienced these sounds spatially rendered through headphones to provide vista-space awareness while sitting as well as walking in a controlled laboratory environment.

To the best of our knowledge, this is the first work to address the need for PVI to be aware of scenes in the vista-space. We contribute with: (1) an understanding of the importance of vista-space awareness for PVIs; (2) an empirical evaluation of the intuitiveness & learnability of handcrafted and AI-generated sounds, pointing out challenges and opportunities, and (3) an empirical evaluation to understand how sonified vista-space can create awareness for PVI.

2 LITERATURE REVIEW

In this section, first, we briefly discuss the assistive systems for PVIs in Ubicomp literature and outline the gap that our work tries to address. Then we position the need for spatial awareness of vista scenes in the theoretical framework of Maslow's hierarchy of needs. Next, we discuss the current sonification techniques used for spatial cognition in studies and applications for PVI, and finally we motivate the exploration of generative AI models for audio generation in the context of vista scene sonification for PVIs.

2.1 Assistive Systems for PVI in Ubicomp Research

The research of assistive technologies is a common theme in ubicomp, which has a rich history of diverse concepts and artifacts. PVI are often the target users, as they have a specific need for assistive technologies. Assistive tech can be useful in many scenarios, such as for leisure activity. For instance, Rector et al. [72] explored audio interfaces for PVI to experience art installations. Other applications include support for grocery shopping, such as explored by Boldu et al. [6, 7]. For this, mounting a camera onto the finger seems to be a common strategy [6, 82] although this physical setup is prone to obstruct the PVI in their daily routines. However, most extensive research is on the application of pathfinding, which supports PVI indoors [28] and outdoors [42]. A common research focus is on the exploration of using multi-modal feedback. Xu et al. [91] empirically evidenced that both, audio and haptic cues are both valid means for PVI showing similar accuracy for path finding. Vibrotactile actuators were investigated at several body positions in a wearable setup [91] as well as in form of mobile artifacts, such as at a handle of a bag [49], watch [85], bracelet [74] and a cane [10, 67]. Although sounding contradictory, also visual cues can be utilized to guide PVI as long as there is a residual gaze, as demonstrated by Yang et al. [92]. The closest concept to ours is FootNotes [29], where GPS-based geo-locations of several points of interest can be marked on a map, and a mobile phone app connected to a 3D audio-capable headset reads out the points of interest when nearing them. In LiSee [11], a custom headset was developed that provides audio feedback to guide users to reach nearby objects based on a video-processing approach. In this paper, we augment existing research

by transforming scenes and objects in the vista space into auditory representations, addressing the challenges of accessibility of the vista space for individuals with visual impairments.

2.2 Positioning vista-space awareness in Maslow's hierarchy of needs

According to Maslow [58, 59], human needs and motivations can be arranged in the form of a pyramidal hierarchy, ranging from *deficiency* needs at the bottom to *growth* needs at the top. Although this theory has faced criticism [23], it remains a valid framework that can be utilized to comprehend and address the diverse needs and motivations of individuals across different contexts and situations. Watkinson et al. [88] suggested using Maslow's hierarchy of needs as a theoretical model for addressing the needs of patients with visual impairment. The *deficiency needs* include fundamental needs such as *physiological needs* (e.g. eating, drinking, and sleeping), *safety needs* (e.g. being alert of dangers and being able to avoid obstacles and safely navigate), and *love and belonging needs* (e.g. compassion and empathy). The *growth needs* refer to needs that lead to realizing a person's full potential, such as *cognitive needs* (e.g. knowledge and understanding, curiosity, exploration), *esteem needs* (e.g. independence, dignity, achievement), *aesthetic needs* (e.g. appreciation and search for beauty, balance, form), and *self-actualization* (e.g. realizing personal potential, self-fulfillment). Based on these definitions, we classify the currently available technologies and studies targeted at PVI into the broad categories of deficiency and growth needs, as shown in Figure 3. For a long time, a large number of studies for PVI have focused on developing techniques for obstacle avoidance and object detection [6, 7, 15, 43, 69, 80], danger alert [41, 70, 83], and wayfinding [39, 54, 68, 75] tasks, that can be classified under deficiency (safety) needs. More recently, studies have also started looking at cognitive and aesthetic needs of PVI, such as access to VR [40, 95], makeup [55], social media [3], emoji [84], screen content [56], face recognition [96], photography [33], and sports [94]. The experience of the vista-space scenes, for example after hiking to the top of a hill, is readily available to people with sight, but is unavailable to PVI. Therefore, vista-space awareness is less functional, and more experiential, which could have the potential of improving the quality of life and sense of independence of PVIs, as such, it could be positioned as a part of the cognitive, aesthetic and esteem needs in the pyramid of needs for PVI.

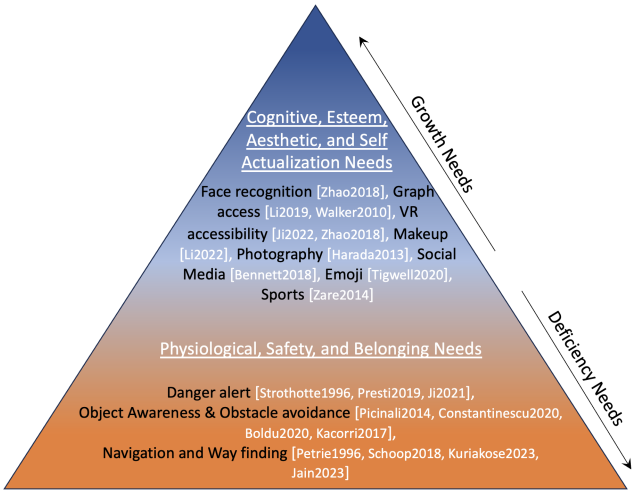


Fig. 3. Broadly classifying technologies and studies for PVI into the hierarchy of needs by Maslow.

Table 1. A summary of the existing sonification methods.

Sonification Method	Advantages	Disadvantages
Spearcons	• easy to learn [16] • provides direct information • acoustically unique representation due to spoken words[87]	• might interfere with the voice instructions of another app or system being used [15] • could be slow if a lot of information needs to be conveyed over a short time [15, 24] • language dependent
Earcons	• helps speed up a speech-based interface [9, 87], • helps the user know what the content of a menu item is, not just where it is in the menu hierarchy [9] • helpful to denote items that have no iconic sound [87] • arbitrary mapping, hence any set of concepts can be represented [9]	• arbitrary mapping, hence hard to learn, needing explicit training [16] • potentially limited transfer of training when moving between two systems employing different earcon "languages" [87]
Auditory Icons	• intuitive, as there is a resemblance to the concept it represents (i.e. natural auditory associations) [2, 25] • apt to convey objects of interest [2] • language independent, when used alone • can be short, thus being able to convey a lot of information in a short time [15]	• easier to learn than earcons, but still needs a training phase [16] • needs expert knowledge to design object-specific sounds, therefore, not easily scalable • difficult to design icons for objects and concepts that have no natural sounds

2.3 Current sonification techniques for spatial cognition for PVI

To create a mental representation of a space, peripheral information plays an important role in updating the relative positions of surrounding landmarks when one is moving [27]. For a person with visual impairment, the amount of information naturally provided by the environment is reduced, making the comprehension of the location and nature of available landmarks difficult [60]. Virtual audio 3D rendering (spatial audio) has been long employed for research in spatial cognition with PVI [1, 47, 48], as it provides the ability to play back individual sounds at specific positions, or to create complex spatial audio scenes without having to manipulate physical devices. Thus, spatial audio has been used to study various low and high-level cognitive processes such as localization, spatial configurations, and architectural navigation for PVI. In this work, we use spatial audio to investigate the usability of different types of sounds for vista-space awareness for PVI.

Different kinds of sounds are used to map objects or events to audio, such as auditory icons, earcons, and spearcons [16, 18]. Auditory icons [25] are brief sounds that have semantic connections to the objects, functions, and actions they represent. They take advantage of the user's prior knowledge and natural auditory associations with sound sources and causes. On the other hand, earcons are abstract, synthetic, and mostly musical tones or sound patterns, whose attributes reflect the structure of a hierarchy of information (eg. menus). Earcons are especially helpful for items with no clear iconic representation, such as the arrival of an email. Spearcons are sounds obtained by speeding up speech sounds—usually to the point where they are no longer recognizable as speech—while conserving their original pitch [87]. Each spearcon is unique and non-arbitrary due to the specific underlying speech phrase, which has been found to require less learning compared to earcons [65] and auditory icons [16]. The advantages and disadvantages of these methods are well-studied in the literature [2, 16, 87]. We provide a summary of these studies in Table 1.

Sonification is widely employed in aiding people with visual impairments, offering accessibility to visual content like graphs [56, 86], maps exploration [19], navigation support [39, 54, 68, 75], and identification of features and obstacles [15, 69]. However, there is a gap in the literature regarding sonification for the use-case of communicating far-field or vista-space scenes. Constantinescu et al.'s study [15] is pertinent, utilizing auditory icons designed using parametric modeling [25] for object representation. We build upon this work by assessing the usability of auditory icons for vista-space awareness, incorporating far-field scenes. Additionally, parametric sound modeling, which involves sound designing for individual objects, lacks scalability due to the manual crafting of the model parameters, demanding expert knowledge. This challenge intensifies when there are simultaneous and multiple events in a scene. Furthermore, we believe that experiencing the aesthetics of a vista scene involves communicating not only the information about the individual objects in the scene but also the interaction between the objects comprising the scene. Hence, our study assesses the usability of AI-based generative models for synthesizing auditory icons to communicate both objects and vista-scenes efficiently.

2.4 Generative models for AI-sounds

The state of the art in the field of generative models for audio is progressing rapidly. Traditionally, audio generation has been achieved through signal-processing techniques [46]. In recent years, generative models have achieved high-quality audio synthesis through various modeling approaches including generative adversarial networks (GANs), variational autoencoders (VAEs), transformer-based, and diffusion-based models. These models are either unconditional [51] or conditioned on different input prompt modalities such as labels [13, 45], text [57], speech [17], audio examples [44], images [38, 79], musical parameters such as pitch and instruments [21, 22, 64] for musical sounds, and other parameter labels like wind-strength [32, 90] for environmental sounds. All of these models are trained and validated to generate high-quality domain-specific audio (eg. speech, music, and environmental sounds). Here, due to the focus of our research, we consider the recent models that show a high performance in generating environmental sounds.

We considered audio-generative models that use natural language text and images as input prompts. These models have the potential to provide higher flexibility and finer-grained scene descriptions for the audio to be generated than what the models conditioned on simple discrete labels [12] can provide. For the task of text-to-audio generation, diffusion models [35, 81] have achieved state-of-the-art synthesis quality, outperforming GAN-based, and autoregressive models (eg. [53, 57, 93]. AudioLDM [57] uses latent diffusion models (LDMs) for text-to-audio generation on a continuous latent representation of audio, instead of learning discrete representations, based on the text embedding obtained by a pre-trained text encoder CLAP (Contrastive Language Audio Pretraining) [20]. With pre-trained CLAP to connect audio and text, the audio embedding and the text embedding share a joint space, and both contain cross-modal information. AudioLDM is shown to outperform the diffusion model-based DiffSound [93], the autoregressive model AudioGen [53], as well as GAN models in terms of overall audio quality as well as its relevance to the input prompts. For the task of image-guided audio generation, Im2Wav [79] consists of a transformer-based audio language model conditioned on image representations obtained from a pre-trained CLIP model [71]. Im2Wav is shown to outperform the other baselines in terms of the generated audio quality as well as the image-to-audio correspondence.

Although the general audio quality of AI-generated sounds has been evaluated, their utility in specific use cases or contexts is still an open question. In this work, we particularly evaluate the sounds generated from AudioLDM [57] (text-to-audio) and Im2Wav [79] (image-to-audio) for the task of designing audio feedback for spatial awareness. And conduct user studies with both sighted and PVI participants to analyze the utility of sounds for vista-space awareness.

Table 2. Demographic of the PVI participants along with details of their medical condition.

P	Age	Gender	VI	Onset of VI	Medical Details
1	49	F	partially blind	at 14 yrs	Tunnel vision (<5% visual acuity)
2	49	M	legally blind	since birth	<2% visual acuity
3	41	M	legally blind	at 16 yrs	<2% visual acuity
4	30	M	legally blind	at 26 yrs	<2% visual acuity
5	26	M	partially blind	since birth	Retinoblastoma, can see colors and shapes but blurry
6	59	M	partially blind	since birth	Glaucoma, highly shortsighted
7	63	F	partially blind	at 19 yrs	<5% visual acuity

3 FORMATIVE STUDY

Our formative study aimed to investigate the relevance and importance of vista-space awareness in individuals with visual impairments. In this section, we discuss the procedure and our qualitative findings from this study.

3.1 Procedure and Participants

The formative study included semi-structured interviews with participants with visual impairments using open-ended questions. In addition, we asked questions to understand the current assistive technology usage amongst these participants, as well as build our understanding of the needs and desires of the PVI community, where future technology development could help. We recruited seven participants with various degrees of visual impairments through local organizations for the visually impaired. Detailed demographics of the participants are in Table 2. We obtained ethics approval from the Institute Review Board (IRB) before conducting all the user studies in this work. Our questionnaire for the formative study is provided in Appendix A.

3.2 Qualitative Findings

We analyzed the qualitative comments from the participants using reflexive thematic analysis. We went through the comments, coded them using a semantic coding strategy to capture the commonly occurring semantic observations in the comments, and recursively combined and refined the codes to finally arrive at 16 codes, which could be then organized into three themes. The count plot of these codes along with their grouping into themes is shown in Figure 4. The themes and their take-aways are discussed here.

3.2.1 The need for vista-space awareness differs depending on PVI's condition. When we brainstormed with the participants on what more ways they think technology could assist them, one of the points raised by *P*4, who acquired visual impairments at a later stage in their life, was that they missed being aware of the scenes around them. He mentioned, "*Let's say I'm in a park or somewhere where maybe there can be a mode (of a device) that describes to me the scenery surrounding me, (and, something that tells me that) there are multiple trees or a lake [...] I think that will be really good also because it kind of gives you an idea of the place and also lets you enjoy the scenery.*" *P*3, who has been blind since the age of 16, mentioned "*I hike a lot, so when I walk up to a hilltop, I would try to draw a picture of the landscape in my mind, try to imagine.*" *P*1 who has tunnel vision since the age of 14, said that she likes to hike a lot, but has to depend on her guide to tell her what the scenes are around her after she climbs a hill. *P*5 and *P*7 who had partial vision condition mentioned they would want to enjoy the scenes on their leisurely walks. Moreover, *P*4 mentioned that only a few of his friends would provide him with scene descriptions, and in general, he feels uncomfortable asking others to describe the scenes for him. However, *P*2, who has been blind since birth mentioned, "*I would not relate much to scenes*

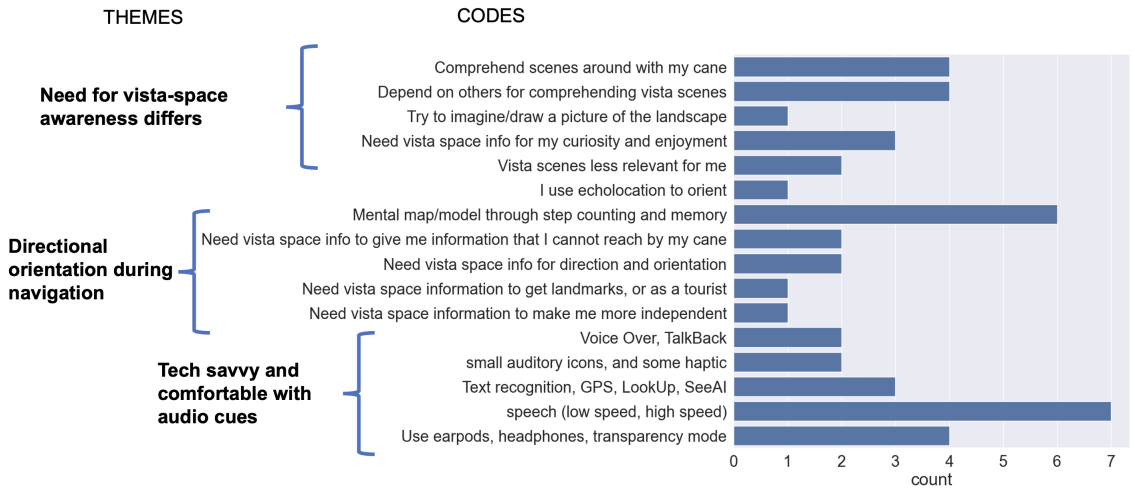


Fig. 4. Themes and codes derived through thematic analysis of the qualitative inputs from PVIs during the formative study.

and landscapes as I would have less imagination compared to someone who acquired blindness at a later stage in their life". Thus, the need for experiencing vista scenes during leisure activities was particularly relevant to the participants who acquired visual impairments later in their lives and possibly had memories of landscapes from earlier in their lives. All four participants who became visually impaired later in their lives indicated this need in different ways such as asking their companion to describe the vista scene, trying to imagine the landscape, and so on. Two out of the three participants with a visual impairment condition from birth said that the experience of the vista scenes was less relevant to them. Therefore, apart from basic safety and obstacle detection needs, based on the formative study, we identified that there is a desire in PVIs who acquired visual impairment condition later in their lives to have the ability to sense and enjoy the scenes surrounding them at a distance, i.e. vista-space, which falls under cognitive and aesthetic needs in the hierarchy of needs (Section 2.2).

3.2.2 Directional orientation during navigation. Another important point raised by all the participants was that they often have a basic mental map of their surroundings based on their step count, or important landmarks that they have identified when they navigate a path that they know very well. However, if they stumble because of a new obstacle (eg. a new potted plant) in their path or simply drift from their route unknowingly, they often need help to re-orient themselves. In such cases, having information about key landmarks around them would help. P1 mentioned, "*The regions that I cannot reach by my cane, but it's something that you can tell me or give an idea, for example, if I go to college, (the device) can tell me there's a cafe on my left, or a big shopping mall in another direction*". *P*5 mentioned "*if you're walking around in the city, maybe there's a landmark that, you can see from everywhere, that helps you orient yourself. I feel something to tell me about those would be really helpful for orientation.*" *P*4 mentioned "*So if there's like a system that tells me about the landmarks, like if I am heading in the direction of my landmark, or have gone off-course, would be very helpful.*" They mentioned that navigation apps such as Google Maps are not always helpful, as there is a margin of error (2-5 meters) in localizing them which often leads to confusion while walking. Most of the participants mentioned that they think vista space information would be helpful for

various directional orientation reasons during navigation such as to get information beyond the reach of their cane, to find out about the landmarks around them, and to help them with general direction and re-orientation.

3.2.3 Tech-savvy and comfortable with audio cues. We found that all the participants use a white cane and one participant also uses a guide dog. In terms of assistive technologies, all the participants use smartphones with integrated screen readers (called VoiceOver[2] in iPhones and TalkBack[3] in Android phones) with adjustable speed. Some of the participants use mobile apps for navigation and money counting, such as LookOut[4]. All the apps that they use provide them with audio feedback, mostly spearcons and a few auditory icons such as beeps, and all the participants were used to and comfortable having audio feedback. All the participants use headphones or earpods in conjunction with the assistive apps. Five participants also mentioned that they use headphones in transparent mode or use only one earpod to be aware of the sounds in their surroundings. Our findings concur with the literature, stating that individuals with visual impairments are tech-savvy and eager to adopt new technologies [66], which can enhance their independence and quality of life [63].

3.3 Focus of this work

This formative study evolved our initial understanding of the relevance of vista-space awareness to PVIs. The need for experiencing vista scenes during leisure activities was relevant to PVIs who became visually impaired later in their lives, and the need for vista scene information for various directional orientation reasons during navigation was consistently observed across all the participants. Thus, together with PVIs, the idea was sparked to map sounds to distant objects and scenes, i.e. "SonicVista", which is aimed to effectively provide vista-space awareness to PVI.

Based on the formative study, we defined that the focus of this work is to evaluate the effectiveness of different types of sounds in communicating scenes and objects, and to assess the utility of such sounds for vista-space awareness. We wanted to find out if AI-generated sounds can be used as an alternate to handcrafted audio recordings for communicating vista-scenes. For this, we first conducted a sound evaluation study comparing two kinds of AI sounds and handcrafted recordings (comparative study) with sighted participants. This study helped us derive a subset of generally intuitive and high quality sounds. This subset was then verified for their intuitiveness and quality with PVI participants. In the final study, we analyzed the utility of these sounds for vista-scene sonification with PVI participants through a Wizard-of-Oz setup (wizard-of-oz study).

Comparative study: *Evaluation of Sounds.* First, we studied the usability of different sound types. We designed three sets of sounds by three different sound generation methods to represent a set of objects and scenes, and we conducted a listening study with sighted people to evaluate whether the sounds could be adequately mapped to the intended objects/scenes and are pleasant to listen to. We addressed objective 2 through this study.

Wizard-of-Oz Study: *Utility of Sonification for Vista-Space Awareness.* In the second user evaluation, we perform a controlled experiment with PVIs to understand the utility of these sounds and if they effectively help to create vista space awareness. We address objectives 1 and 3 through this study.

4 COMPARATIVE STUDY: EVALUATION OF SOUNDS

In this study, our main aim was to study the possibility of using handcrafted auditory icons, as well as the benefits and challenges of the current AI-generated sounds for communicating objects and scene information.

[2] https://shorturl.at/vGIT6
[3] https://support.google.com/accessibility/android/answer/6283677?hl=en
[4] https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en&gl=US

The motivation of this study was to understand the general quality and characteristics of these sounds and whether these sounds can reliably represent what they were intended to represent. At this stage, we wanted to evaluate the quality of the AI-generated sounds compared to audio recordings for objects and scenes of interest. Hence, we recruited a relatively large easily accessible group of sighted participants for the listening tests in this study to derive insights that pertained to this general goal. We acknowledge that sighted participants are a non-representative group when building tools for blind individuals due to differences in their perception and strategies when using a tool [4, 76]. Therefore, all the findings and insights derived from our study with sighted participants cannot be directly applicable to PVI participants. However, we justify the use of sighted participants for the sound evaluation study for two reasons:

(1) This is a preliminary study to evaluate the general perception of the different kinds of sounds (eg. AI-based) that have not been studied before. A preliminary study of a new solution is also suggested by Sears at al. [76] to be a case where using a non-representative group may be appropriate.
(2) This study helps us derive a smaller subset of well-performing sounds that can then be further verified with PVI participants and used in contexts that are relevant to PVIs (i.e. the Wizard-of-Oz utility study). This further helps us optimize the limited time we have with the small number of PVI participants.

4.1 Objects/Scenes & Sound Types

Based on computer vision literature on object and scene recognition [5, 62, 77], we define an object as a single entity in focus in the field of view, such as a bench or a car. We define a scene as multiple objects being simultaneously present in the field of view within a context having spatial, functional, and semantic relationships between each other. For example, the scene of a canteen will have chairs, tables, cutlery, food, and people. We curated 28 sounds representing 22 objects and 6 scenes. Out of the 22 objects, 19 objects were chosen to be the same as that in [15][5]. The list of objects and scenes and the sources of audio are provided on our webpage[9]. This audio set consists of objects that may or may not be a part of a vista scene. Moreover, the purpose of the sounds designed by [15] was to provide a quick indication of all the objects in a scene, which was why the sounds were designed to be short. This differs from our purpose, which is to provide a summarized experience of the vista scene. Nevertheless, we opted to incorporate this audio set as a starting point in the sound evaluation study to contextualize our work within the existing literature. In addition to these 19 objects, we included 3 more items to the list of objects - *birds*, *bells*, and *ducks*, as signature objects of far-field scenes of a nature park, a bell tower, and a pond respectively. To understand whether sounds can represent the composite nature of scenes that consist of multiple objects and events occurring simultaneously, we also included 6 commonly occurring scenes - *park*, *street*, *beach*, *canteen*, *kids playing*, and *mall*. For example, a *park* scene may have birds chirping, the sound of a breeze, and the rustling of leaves; a *street* scene might have different types of vehicles, and possibly traffic lights, and so on. Thus, we split this list of 28 objects/scenes into three categories - 12 *sonic* objects (objects that make sound naturally), 10 *non-sonic* objects (objects that do not make any sound by themselves), and 6 *scenes*.

We created two sets of AI-generated sounds corresponding to the above list. **AI Model 1 (AudioLDM)**[6], that consisted of sounds generated using text-to-audio generative model, AudioLDM [57]. **AI Model 2 (Im2Wav)**[7], that consisted of sounds generated using image-to-audio generative model, Im2Wav [79]. We obtained multiple license-free images from Unsplash[8] corresponding to each of the objects and scenes, and used these images as the input to the Im2Wav model to generate the corresponding audio. And our baseline comprised of the handcrafted auditory icons of the 19 objects from [15] plus audio recordings from FreeSound[9] for the rest of the objects and

[5]There were 18 objects in [15] and one additional object *building* was available in the audio set shared by the authors with us.
[6]AudioLDM codebase and model: https://github.com/haoheliu/AudioLDM
[7]Im2Wav codebase and model: https://github.com/RoySheffer/im2wav
[8]https://unsplash.com/
[9]https://freesound.org/

scenes. The complete list of objects and scenes, along with their audio files for handcrafted, AudioLDM, and Im2Wav, as well as the text and image prompts for the two AI models are provided on our webpage[10].

4.2 Participants

We recruited a total of 24 sighted participants (11 males/ 13 females). Their ages ranged from 22 to 40 years (mean age was 28 years) and were either students or staff at the university. The participants listened to the sounds through noise-canceling headphones[11].

4.3 Data Preparation

We used a total of 84 sounds, i.e., 28 sounds from each of the three sonification mechanisms: audio recordings/handcrafted sounds, AI Model 1 (AudioLDM), and AI Model 2 (Im2Wav) with sighted participants. For each sonification mechanism, there were 12 sounds representing sonic objects, 10 sounds representing non-sonic objects, and 6 sounds representing scenes. To conduct the study for every participant within a stipulated time (approx. one hour) while also preventing fatigue, we exposed each participant to only half of the sounds. We ensured a balanced distribution of sounds across the 24 participants so as to gather a uniform number of ratings across all the sounds in our dataset. Specifically, we presented different subsets of 14 sounds out of the 28 sounds for each sound type (i.e. handcrafted, AI Model 1, and AI Model 2) during each test (intuitiveness and learnability) to different participants, such that upon concluding the study for all the participants, we had at least 12 scores for intuitiveness and learnability for every sound in our data set. We, also, randomized the order of sounds across sound types as well as objects/scenes for each participant. All of these permutations were designed to counterbalance the effects of learning and familiarity bias. Moreover, for every sound, we had 24 self-reported ratings of intuitiveness, pleasantness, and comfort level collected from all the participants. We presented a subset of 8 sounds from this data set to the VI participants, as further explained in Section 4.6.2.

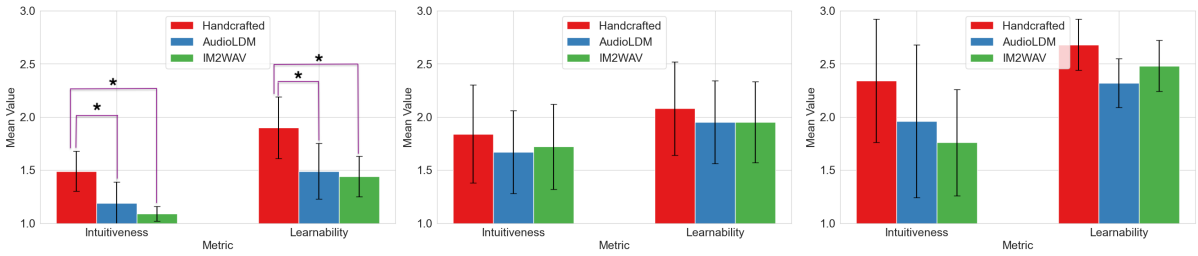
4.4 Procedure

Each participant was given a list of all the objects and scenes in our dataset. Following the exact procedure for the intuitiveness test from [15], we played the sounds one at a time, and the participant was asked to identify the object or scene from either a list of 22 objects or a list of 6 scenes. We recorded their answer, and the time they took to guess after the sound was played. If the participants answered quickly, without taking time to think, the sound is intuitive. Following this, participants were given access to all 28 sounds along with their assigned labels through a web-page. They could listen to the sounds as many times as they wanted and were asked to rate each sound on intuitiveness, pleasantness, and comfort level for long-term usage on a Likert-scale of 1 to 5. Finally, we conducted a learnability test [15], which tested the ability of the participants to learn and remember the objects or scenes associated with the sounds. This entire process was repeated for the three sound types. The whole study took about 1 hour to complete.

4.5 Evaluation Metrics

We calculated an *intuitiveness score* [15] based on the response times and success rates of the participants in guessing which of the objects or scenes a sound belongs to. Also, we calculated a *learnability score* [15] by assessing how easily the participants remembered the object classes associated with the sounds. The values of these scores ranged from 1 to 3, as follows: 1 - if the participant did not guess or remember the label of the sound correctly, 2 - if the participant guessed or remembered the sound label correctly after thinking for longer than 3

[10] Our anonymized webpage with audio files for your reference https://spatial-awareness-project.github.io/spatial-awareness-project/
[11] https://www.sony.com.sg/electronics/headband-headphones/wh-1000xm4?showModelPrices=true



(a) Non-Sonic Objects (b) Sonic Objects (c) Scenes

Fig. 5. Overall performance of the comparative study metrics - Intuitiveness Score and Learnability Score (out of 3.0) for the three sound types [Mean ± 95% Confidence Interval], obtained from the sighted participants. Values with a $*$ indicate p-value<0.05 when compared against both of the other two sound types using Wilcoxon signed rank test.

seconds after the sound was played, and 3 - if the participant guessed or remembered the sound label correctly in less than 3 seconds after the sound was played. To assess the statistical significance between any two lists of scores, we used the Wilcoxon signed rank test assuming related paired samples, and for group comparisons, we used the Friedman test.

4.6 Results

4.6.1 Ratings from the Sighted Participants. The intuitiveness and learnability scores (out of 3) of this study are shown in Figure 5. For non-sonic objects, the mean values of intuitiveness scores of handcrafted, AudioLDM, and Im2Wav sounds were 1.49, 1.19, and 1.08, respectively. A Wilcoxon signed-rank test showed that for this type of object, the intuitiveness score of handcrafted sounds was significantly higher than AudioLDM ($W = 2$, $Z = -2.599$, $p = 0.009$), as well as Im2Wav ($W = 0$, $Z = -2.8031$, $p = 0.005$). Similarly, the learnability score of handcrafted sounds for non-sonic objects was also significantly higher than AudioLDM ($W = 8$, $Z = -1.9876$, $p = 0.047$), as well as Im2Wav ($W = 7$, $Z = -2.0896$, $p = .037$) sounds. Since AI models are predominantly trained on sonic objects and scenes and not on non-sonic objects, the sounds generated by these models for non-sonic objects are understandably less intuitive or learnable than the hand-designed ones. However, for sonic objects and scenes, the performance of all three sound types has no significant difference for both intuitiveness and learnability scores. However, the Friedman test showed no significant difference in the ranks of the data values between the three sound types for sonic objects (intuitiveness: $\chi^2 = 1.625$, $df = 2$, $p = 0.443$; learnability: $\chi^2 = 4.04$, $df = 2$, $p = 0.133$) and scenes (intuitiveness: $\chi^2 = 0.333$, $df = 2$, $p = 0.846$, learnability: $\chi^2 = 2.33$, $df = 2$, $p = 0.311$). This implies that the current generative models have the potential to generate intuitive and learnable sounds for objects that inherently make sounds as well as scenes that have multiple events. The self-report ratings on intuitiveness, pleasantness, and comfort-level for sustained usage are shown in Figure 6. For scenes, the handcrafted sounds, which are audio recordings of the scenes, are rated significantly higher than the sounds from the two AI models for all three criteria ($p < 0.05$), where the test of significance was done using Wilcoxon signed-rank test. A possible reason could be that the AI models are predominantly trained on the sounds of individual objects and events, but not on a collection of multiple events or objects occurring at the same time that comprises a scene or a soundscape.

A notable exception amongst the non-sonic objects was the handcrafted *door* sound, which got incorrectly guessed as the sound of *stairs* several times because the discrete events in the designed auditory icon representing either the closing of a door or knocking on a door could be confused with walking on wooden stairs. The sound

of *door* generated by AudioLDM, on the other hand, represented the creaking of a door which is a distinct feature of a door and thus, was less confusing and more intuitive as well as learnable.

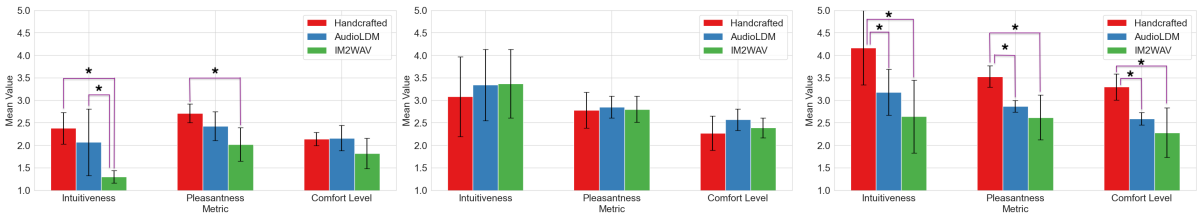
Notable high-performing sonic objects were the sounds of *bells*, *birds*, *dog*, and *ducks*, which were highly intuitive as well as learnable (scores $\geq$ 2.0) across all three sound types. The handcrafted *traffic light* sound was significantly more intuitive and learnable than the other two types, because the handcrafted sound was the signature pedestrian traffic light sound ("green man"), whereas the two AI models generated traffic noises or police/ambulance sirens for traffic light prompts, which was indeed not intuitive and got confused with vehicle sounds such as car and truck. Amongst the sonic objects, short handcrafted sounds were rated lower on pleasantness than the corresponding AI-generated sounds, while the opposite was true for longer handcrafted sounds (e.g. bells, birds, ducks, and traffic light).

Amongst the scenes, the handcrafted sound (i.e. audio recordings for scenes) of the *canteen* scene, which had a distinct sound of cutlery as well as people chatting, was highly intuitive and learnable. On the other hand, the AudioLDM *canteen* sound had chatting people and rhythmic music along with a faint sound of cutlery, while the Im2Wav *canteen* sound mostly had chatting people, so both of these sounds were confused with the scene of a *mall*. Another notable instance was the Im2Wav sound of the *beach* scene which was highly intuitive and learnable, possibly because of the distinct water gurgling sound and the amplitude modulation corresponding to waves crashing on the shore and receding. On the other hand, both the AudioLDM and the handcrafted audio recording of the *beach* scene were highly confused with the street scene. Both these sounds had a constant noise or a slightly increasing noise amplitude, which can be found at a beach on a windy day, but which is likely to be also present on a street.

More details on intuitiveness and learnability scores and self-reported ratings can be found in the Appendix.

4.6.2 Extension of the results to PVIs. To inform the next study (Wizard-of-Oz) that is designed to evaluate the utility of sounds in relevant contexts for PVI participants only, we needed to first investigate if some of our findings from the sound evaluation study with the sighted participants could be extended to PVI. For this, we did two validation tests.

In the first test, we compared the mean intuitiveness and learnability scores reported by our sighted participants to the corresponding scores reported by [15] in their study with seven PVIs for the 18 objects that were common between the two experiments. This comparison is only an approximate indicator, as the number of objects, the number of participants, as well as the purpose of the two experiments, are different. The normalized Euclidean distance between two series of values ranges between 0 to 1, where a smaller distance means the values are similar, and cosine similarity between two vectors also ranges between 0 to 1, where a higher cosine of the



(a) Non-Sonic Objects (b) Sonic Objects (c) Scenes

Fig. 6. Overall performance of the ratings obtained from the Self-Report web survey from the sighted participants in the comparative study for the criteria - Intuitiveness, Pleasantness, and Comfort-level (out of 5.0) for the three sound types [Mean ± 95% Confidence Interval]. Values with a $*$ indicate p-value< 0.05 when compared against both of the other two sound types using Wilcoxon signed rank test.

angle between the two vectors means they point in a similar direction. We found that the normalized Euclidean distance between PVI and sighted participants' scores for intuitiveness and learnability was small (0.02 and 0.03 respectively) and the cosine similarity between them was high (0.97 for both) which confirms that the sighted participants perform similarly to PVI participants when they map sounds to objects.

In the second test, we conducted a shortened version of this study with the seven PVIs that we recruited. The PVI participants were the same as those who participated in the Formative study earlier (Table 2). Based on a linear combination of the scores and ratings, we chose a subset of four high-scoring objects and scenes, i.e. *canteen* and *beach* as vista scenes, and *birds* and *bells* as signature objects of a vista scene, from the study with the sighted participants to verify that this small subset of sounds is similarly intuitive for PVI. Since there were only four objects/scenes, for the intuitiveness test, we did not present the list of objects to the participants and instead asked them to tell us what object or scene was the most relatable to the sound they heard. So as long as the meaning was the same as the label assigned by us, we considered the guess to be correct (eg. *restaurant*, *food stall*, and *eatery* are considered as correct guess for our label *canteen*). Moreover, although we present the learnability scores, learning only four sounds was easy, therefore the task was not comparable to the one with the sighted participants.

All four handcrafted sounds were audio recordings, while for comparison, we chose the highest scoring amongst the two AI-generated sounds, i.e. AudioLDM sounds for *bells*, *birds*, and *canteen*, and Im2Wav sound for *beach*. We found that the chosen sounds were similarly intuitive for the PVIs for both the handcrafted sounds as well as the AI-generated sounds for these four objects/scenes. The mean ratings are provided in Table 3. The normalized Euclidean distance between the ratings of sighted and PVI was small (< 0.02 across all the ratings) and the cosine similarity between them was high (> 0.97 across all ratings) for both the handcrafted and the AI-generated sounds. This test indicates that the chosen subset of sounds are intuitive for PVIs, and thus can be used for the next scene sonification study.

5 WIZARD-OF-OZ STUDY: UTILITY OF SONIFICATION FOR VISTA-SPACE AWARENESS

In this study, we aimed to evaluate the utility of sonification of the vista space in providing passive awareness of vista scenes to PVI. For this purpose, we chose a subset of objects and scenes to evaluate the utility of auditory icons (audio recordings and AI sounds). Additionally, we used spearcons [16, 18] as an alternate form of feedback, because the formative study suggested that all the participants were used to and comfortable with audio feedback in the form of spearcons, such as screen readers with adjustable speed. Therefore, spoken words corresponding to the objects and scenes with speed adjusted to the comfort level of the individual participant were used for comparison.

Table 3. Comparison of sound evaluation metrics between participants with visual impairments (PVI) and sighted participants (SP) for a subset of objects and scenes. Sound Type acronyms are Handcr.: Handcrafted, AI: either AudioLDM or Im2Wav. For *bells*, *birds*, and *canteen*, the AI model is AudioLDM, and for *beach*, the AI model is Im2Wav. The intuitiveness and learnability scores are out of 3.0, and the self-reported intuitiveness, pleasantness, and comfort ratings are out of 5.0.

	Intuitiveness Scores				**Learnability Scores**				**Reported Intuitiveness**				**Pleasantness Rating**				**Comfort-Level Rating**			
Sound Type →	Handcr.		AI		Handcr.		AI		Handcr.		AI		Handcr.		AI		Handcr.		AI	
Participants →	PVI	SP	PVI	SP	PVI	SP	PVI	SP	PVI	SP	PVI	SP	PVI	SP	PVI	SP	PVI	SP	PVI	SP
bells	2.86	2.64	2.57	2.64	3.00	2.91	3.00	2.82	4.86	4.96	3.00	4.75	3.57	3.62	2.43	2.67	2.71	2.67	1.86	2.17
birds	3.00	2.92	2.43	2.83	3.00	3.00	3.00	2.92	4.86	5.00	4.57	5.00	4.43	4.04	4.14	3.58	4.43	3.67	3.71	3.29
beach	1.29	1.58	1.71	2.25	3.00	2.55	3.00	2.75	2.29	2.83	3.43	3.25	3.43	3.42	3.57	2.75	3.14	3.46	3.14	2.58
canteen	2.43	2.73	1.43	1.09	3.00	2.82	3.00	2.09	4.14	4.75	3.14	2.54	3.00	3.42	2.57	2.75	2.57	3.33	2.14	2.46

5.1 Sounds

We have two sound types for this user study - auditory icons and spearcons. For auditory icons, we chose a subset of sounds that performed well in the sound evaluation study verified by the PVI participants, that consisted of a mix of AI-generated and handcrafted sound. Specifically, we chose two high-performing AI-generated sounds - Im2Wav sound of *beach* and AudioLDM sound of *birds*, and two high-performing handcrafted recordings - *bells* and *canteen*, where *birds* and *bells* were under the sonic objects category and *beach* and *canteen* were under the scenes category. For this wizard-of-oz utility study, we considered these objects and scenes as either a vista scene by itself or a signature object that represents a vista scene. For example, the *birds* sound was considered to be a signature object of a scene with a nature park or a garden, while the *bells* sound was considered to be a signature object of a scene consisting of a clock tower, church or a bell tower. The sound of *canteen* could represent a scene with some kind of a food center, or a restaurant, while the *beach* sound could represent a scene with a sea, a beach, or a water body in general with some wind. The spearcons were the corresponding spoken words created with an online text-to-speech engine[12], and sped up without changing the pitch of the voice in Audacity[13]. During the study, we adjusted and selected the speed of the spearcons that the participant said they were comfortable hearing. We chose spearcons as our baseline because this kind of audio feedback is the most familiar to PVI, as derived from the formative study (Section 3).

5.2 Experimental setup

As a follow-up session of the formative study (Section 3) held previously, we recruited the same seven PVI participants (Table 2) and conducted two Wizard-of-Oz spatial awareness listening experiments. We created a virtual environment in Unity[14] that helped us simulate the vista-space through spatialized audio. Our purpose with these simulated environments was to understand what kind of experience would these sounds give if the system worked perfectly in converting vista scenes into spatial sounds. However, the participants were only asked to wear headphones, and at no point during any of these experiments did they wear any head-mounted VR device. Most of the participants in the formative study indicated that they preferred using the transparent mode of their headphones or just one earpod to be aware of their surroundings while listening to audio playback. Bone conduction headphones provide the feature of audio playback without blocking ears and do not interfere with normal hearing. This feature was helpful in this study as it involved alternating between listening to audio playback and interacting with the study coordinator. Moreover, previous studies have found bone conduction headphones to be a preferred method of audio playback in studies involving PVIs because of this feature [7]. Therefore, we chose to use bone conduction headphones[15] for the wizard-of-oz study. In the first experiment, the sounds of the four objects/scenes were spatially rendered within a 180-degree field of view in front of the seated participant through bone-conduction headphones (See Figure 7 (a)). We call this the *central vista-space awareness test*. During the second experiment, which we call the *peripheral vista-space awareness test*, the participant walked a short distance (guided by a sighted person), while the sounds of the four objects/scenes (in the peripheral vista-space) were spatially rendered at different points along the walk, either to the left or the right of the participant (See Figure 7 (b)). Both experiments were conducted in a controlled laboratory environment without any obstacles, as shown in Figures 7(c) and (d).

[12] https://voicemaker.in/
[13] https://www.audacityteam.org/
[14] https://unity.com/
[15] Bone conduction headphones used in this study: https://shokz.com/products/openrun

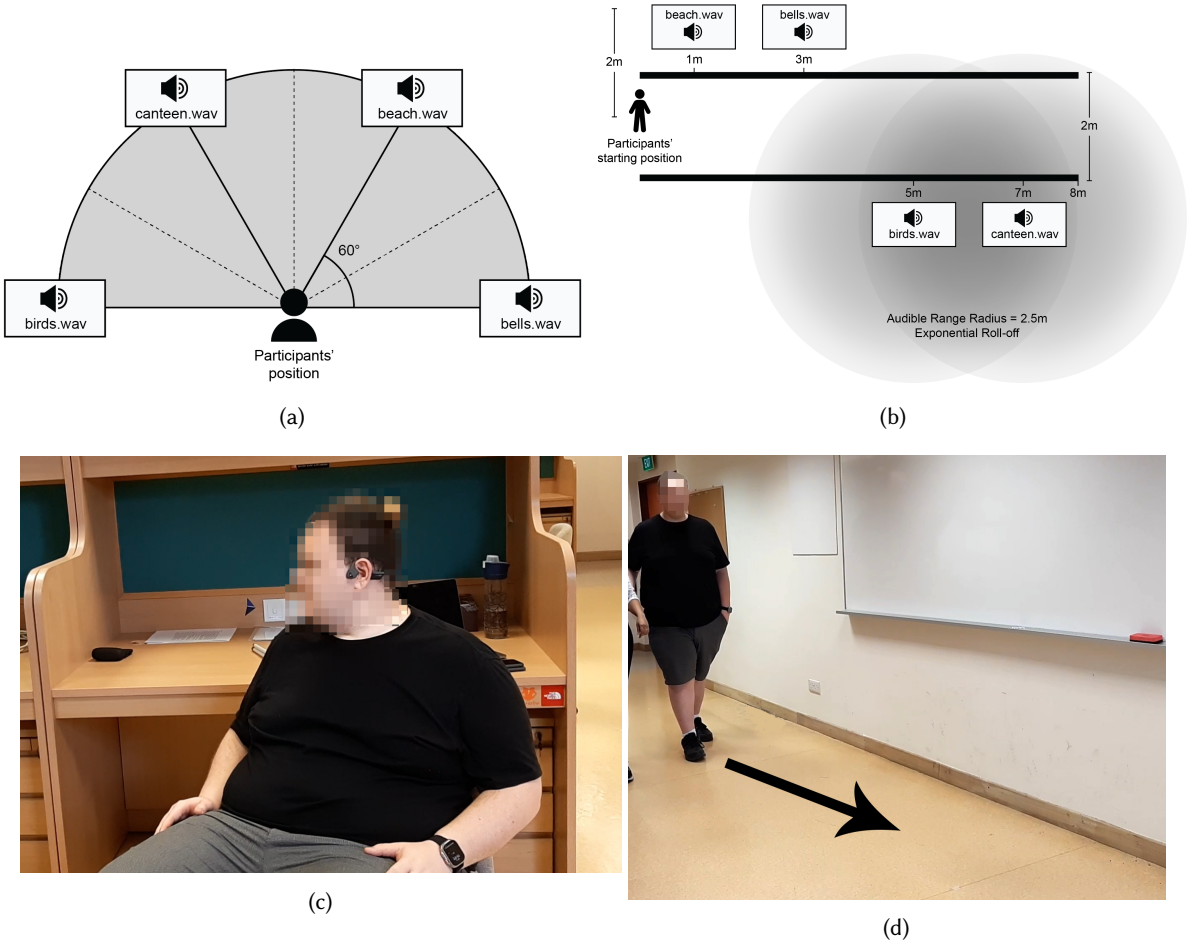


Fig. 7. Overview of the experimental setup for Wizard-of-Oz Study: (a) Central Vista-Space Awareness Test, (b) Peripheral Vista-Space Awareness Test, (c) a PVI participant experiencing spatially rendered sounds through bone conduction headphones while turning their head in the central vista test, and (d) a PVI participant experiencing spatially rendered sounds through bone conduction headphones while walking in the peripheral vista test.

We conducted semi-structured interviews with participants after each experiment to qualitatively assess their experience. The questionnaire of the interview is provided in Appendix B. We have also made our Unity codebase for the two test setups available on GitHub[16].

5.2.1 Central vista-space awareness test. The simulated environment for this test consisted of 4 audio sources placed in an arc around the user at a distance of 5 meters. The sources were placed at angles of -90°, -30°, 30° and 90° from the user's forward-facing direction (See Figure 7 (a)). Each set of sources representing a sound type could be played separately, and the volume of each of the sources was dynamically set. The volume of a particular source would increase linearly as the participant turned to face the audio source and decrease as they turned away. This setup was tested with auditory icons as well as with spearcons.

[16]Unity Codebase to be published on GitHub after paper acceptance.

5.2.2 Peripheral vista-space awareness test. For this test, the simulated environment consisted of the 4 sounds placed along an 8-meter-long, 2-meter-wide passageway. Audio sources were placed at a distance of 1, 3, 5, and 7 meters away from the starting position of the participant and 2 meters to the left or right of the center of the passageway (See Figure 7(b)). Each sound source had an audible range/radius of 2.5 meters, with an exponential roll-off function, so that multiple sounds did not interfere with each other along the walk and there was a gradual change in volume as they walked across the corridor. The volume of a particular source would also increase if the participant turned to face the audio source and decrease as they turned away, similar to the central vista setup. The participants, wearing the bone-conduction headset, walked with the guidance of a sighted person along this passageway. They were also asked to pay attention to the environment they were walking in. At the end of the walk, the participants were asked to describe the sounds they heard and the directions they heard them from. This process was done twice, i.e. with auditory icons and spearcons.

5.3 Quantitative Findings

Across the two setups, we calculated the average percentage of objects and scenes correctly recognized by each participant (100 × number of correctly recognized items / total number of items) as well as the accuracy percentage in detecting the correct direction of those sounds (100 × number of correct directions of the correctly recognized items / total number of items). We found that for the auditory icons, the average recognition accuracy was 93.75% while for spearcons, it was 84.38%. This implies that auditory icons have a higher sound-to-object mapping in an active situation (observing the surroundings from one location, or while walking a short distance) compared to spearcons. As mentioned by participant *P*3, spoken words are often missed, which could be the cause of their lower accuracy, whereas it is hard to miss the auditory icons. However, the average accuracy of detecting the directions of the correctly recognized objects and scenes was 100% for both the sound types, which implies that as long as the audio is correctly recognized, spatializing that audio could provide direction, which could, therefore, help in orientation.

5.4 Qualitative Findings

We analyzed the open-ended qualitative thoughts and comments from the participants in the semi-structured interview after the two experiments using the bottom-up approach of the reflexive thematic analysis [8, 14]. We went through the comments, coded them using a semantic coding strategy to capture the commonly occurring semantic observations in the comments, and recursively combined and refined the codes (for example, the codes "Use at leisure" and "Use to enjoy scenes" were combined into one code) to finally arrive at 14 codes, which could be then organized into four themes - Cognition and Aesthetics, Utility Attributes, Orientation and Navigation, and Wishlist. The count plot of these codes along with their grouping into themes is shown in Figure 8. The themes and their take-aways are discussed here.

5.4.1 Cognition and Aesthetics: ***"Makes me feel immersed and included in my environment."***
Auditory icons were considered to be more engaging, pleasant, immersive, and enjoyable than spearcons in experiencing vista scenes during leisure activities. The sonified vista space through auditory icons was positively received by all the participants for relaxed settings, such as when sitting on a bench on top of a hill or on a leisurely walk and scanning the place around by turning their heads. To experience and imagine the aesthetics of the scenes surrounding them, the auditory icons were clearly preferred over spearcons. *P*5 mentioned "*These sounds (auditory icons) help me create the image in my head, gives me a more vivid idea of my surroundings [...] kind of paints a picture of what's around me [...] I would definitely prefer these sounds over speech for experiencing my surroundings*". Similarly, *P*4 said, "*(the auditory icons) make me feel really included and makes me feel immersed in that environment*", and P2 mentioned, "*I could sense how the environment would change with these sounds*". Both *P*3 and *P*5 mentioned they would listen to the auditory icons when they are curious to know and feel the scenes

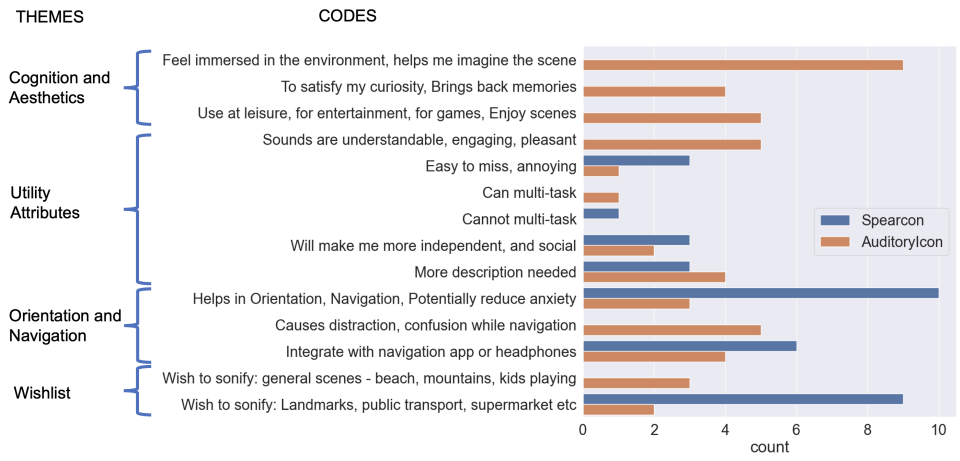


Fig. 8. Themes and codes derived through thematic analysis of the qualitative inputs from PVIs during the Wizard-of-Oz Study.

around, as *P*5 mentioned "*when I want to feel the sensation of what it is to look at a scene from far away.*". Similarly, *P*3 said, "*These sounds are really useful to sense the environment around me. I walk around and know what's around me.*". Moreover, these auditory icons evoked memories in the participants who became blind recently and helped them to imagine the landscape. *P*4 said, "*It really brings back a lot of memories from when I was sighted. Those bird sounds brought me back to that mountain (where) I spent a couple of hours back then.*", and *P*7 mentioned "*These (sounds) are bringing back joyful memories of events when I was sighted.*" Thus, there is a strong cognitive and aesthetic desire to have an immersive sense of vista-space scenes, which these auditory icons show a clear potential to fulfill.

Use-cases of Vista-Space Experience through Auditory Icons: The vista-space experience through the auditory icons (soundscapes) was suggested to be useful in leisurely settings, such as on a hill top, or in virtual environments such as games. For example, *P*4 mentioned, "*As a sighted person I could enjoy the beautiful sceneries when I traveled on a holiday. So these sounds I think can really be great. (For example,) When I can just be looking around and the sound of the waves crashing on one side and the sounds of the vast landscapes on the other side would be really great for me to imagine the scene. I see great potential in this.*" He also mentioned, "*I will 100% use it in a relaxed setting when I am at a rooted place like at a picnic, where I could just be enjoying the scenes with my friends or family*". Another interesting use-case of the auditory icons was mentioned by *P*2 that he would be really interested in using these sounds in virtual environments such as the games for the blind that he plays on his phone.

5.4.2 Utility Attributes: For the use-case of enhanced awareness of the vista-space scenes, the most important attributes of the experience were identified as:

Pleasant, Engaging, and Memorable: The participants in general found the experience with the auditory icons to be pleasant and engaging, and pleasantness of the sounds was found to be important for the experience. For example, *P*3 said, "*These sounds are going to be engaging. As you move along, you hear the environment sound changes*". *P*1 mentioned that "*These sounds (auditory icons) are pleasant to listen to*". The synthesized Im2Wav sound for *beach*, although recognizable as beach, had a slightly grainy texture due to the generative model's audio

quality limitations. Though most of the participants did not notice it, *P*5 commented that due to this texture, he found the experience to be "slightly less pleasant, but the content of the sound isn't unpleasant." Moreover, they found the experience to be memorable. For example, *P*5 mentioned, "(The auditory icons) paint a more vivid picture for me of what I'm listening to. So I can recall it even after a while [...] These sounds are so much more memorable than the speech sounds."

Comprehensible: All the participants mentioned that they could recognize the general scenes, but expected more description both from the auditory icons as well as the spearcons, especially in the peripheral awareness test, i.e. in the context of directional orientation during navigation. For example, *P*1 said, '*Though these (auditory icons) sounds, the scenes are understandable [..] But when I hear the bells, but is it from a (clock) tower or is it from a church? [...] Distance and height are also unknown. More description would be good.*" *P*2 said, "*If I hear the beach sound from one direction, I want to know it is 1 km away*". The spearcons, on the other hand, could be hard to listen to, as mentioned by *P*3, that he often misses listening to some words in spoken descriptions, and finds it annoying to hear for a long period of time, therefore, would prefer a combination of auditory icons with spearcons. He said, "*Even if I miss the text (spearcons), I still hear the auditory icon and I won't miss the auditory icons.*"

Sense of Independence: All the participants acknowledged that vista-space sonification would give them a sense of independence, which satisfies their esteem needs. *P*4 mentioned, "*Only 3 out of 20 friends may give me a nice description of the scenes around[...] I don't usually ask them for descriptions because I don't want to disturb them[...] With this, I can get an experience of the scenes, without having to completely depend on my friends.*" *P*3 mentioned, "*I could sense a park and a forest and start a conversation with a sighted person. So it makes me a bit more independent.*"

5.4.3 Orientation and Navigation: ***"This will help me re-align to the direction I want to head to."***

All the participants suggested that sounds that communicate the vista-space will be helpful in wayfinding and self-orientation in the right direction, especially when people are not around to help. In the peripheral vista-space awareness setup, which simulated the navigation scenario, the participants were looking for direct information rather than experience. Therefore, spearcons were preferred over auditory icons for directional orientation because these sounds provided information without having to guess. *P*4 said, "*I could really see myself walking and using it in day-to-day activities, (as it) could be describing the scenes around me while I walk*". With spearcons, *P*1, *P*4, *P*5, and *P*6 suggested that a system integrated with such sounds would be helpful to provide information about their surroundings during both daily activities such as going to the supermarket as well as outdoor activities such as hiking and camping in the forest, where humans are not available to help. *P*3, on the other hand, said he would prefer spearcons played once for direct information, along with the auditory icons being played continuously in the background to reinforce the information.

Three (*P*2, *P*4, and *P*6) out of seven participants mentioned that the auditory icons could be distracting while walking, as they could get mixed with the existing environmental sounds. *P*2 said, "*If I'm searching for a certain location I will use speech more than sound[...] These (auditory icons) will distract me (in the) outdoors. So I will not use it.*" *P*4 said, "*While walking, I don't feel like it is necessary to have more sounds of your environment already*".

5.4.4 Wishlist.

The participants wished for a range of objects and scenes related to daily necessities that they would want sonified, such as traffic, traffic lights, buses, trains, and pavements. Moreover, they want several common landmarks to be sonified, such as bus stops, train stations, police stations, and supermarkets. Some of the participants who lost their sight later in their lives mentioned that they also wished for sonified aesthetic scenes such as mountains, sea beaches, lakes, and kids playing at a distance.

6 DISCUSSION

6.1 Addressing Objective 1: Importance of vista-space awareness for PVIs

Through the formative study and the wizard-of-oz study, the necessity of vista-space awareness and sonification emerged as a strong need in PVIs under two situations - at leisure and during navigation. For the people who became visually impaired at a later stage of their lives, a key purpose of being aware of the vista-space scenes was to regain the ability to sense and enjoy the scenes surrounding them, thus aligning with their cognitive, esteem, and aesthetic needs. This is in line with positioning the need for vista-space awareness in the upper half of Maslow's hierarchy of needs. This is also in line with the vision of assistive augmentation [37] that suggests empowering individuals with disabilities beyond their basic needs. Additionally, during navigation, they wanted to find out more about their surroundings for curiosity and awareness, especially in places and situations where help is not easily available, eg. while hiking. This suggests the importance of vista-space awareness for cognitive and esteem needs in Maslow's hierarchy of needs. Furthermore, all the PVI participants found vista-space awareness to be useful for getting a sense of the direction of the landmarks around them during navigation, which could help them re-orient themselves if they diverge from their path. Directional orientation during navigation could also be considered to be a safety need, thus also suggesting the importance of vista-space awareness in the lower half of Maslow's pyramid of needs.

6.2 Addressing Objective 2: Comparison between handcrafted and AI-based sonification methods

We conducted the sound evaluation study with sighted participants and verified our findings with a small set of PVIs. We found their ratings for the sounds in terms of intuitiveness, learnability, pleasantness, and comfort to be similar. Moreover, we found that there was no significant difference between the scores of intuitiveness and learnability from the sighted participants for the sounds when compared to the corresponding scores reported by [15] with PVIs in their study, implying that the sighted participants performed similarly to PVI when they mapped sounds to objects.

For sonic objects as well as scenes, the performances of the AI model sound types and the handcrafted sounds are comparable in terms of intuitiveness, and learnability. This implies that in contrast to the existing handcrafted auditory icons used in literature [15], the state-of-the-art AI models can provide a more scalable method of generating sounds for sonic objects and scenes, that can be as intuitive and learnable as the handcrafted sounds.

The ratings of pleasantness and comfort-level for long usage of the three sound types for non-sonic and sonic objects were found to be comparable. This implies that the audio quality of the AI-based sounds for objects is similar to that of the handcrafted sounds, which means that the scalability offered by the AI models can be used to produce pleasant sounds for objects. Many of the AudioLDM sounds for the non-sonic as well as sonic objects were rated to be highly comfortable for listening for a long period of usage as a background sound. A reason could be that these sounds are closer to the naturally occurring sounds, whereas the handcrafted sounds that were designed for these objects using parametric modeling [15]sound artificial and thus not quite suitable for listening over longer periods of time.

However, for the scenes, the pleasantness and comfort ratings for the handcrafted sounds were significantly higher than the AI-generated sounds. This degraded audio quality of the AI-generated sounds for scenes could be because the AI models are predominantly trained on datasets consisting of objects and event labels, but not on scene soundscapes that consist of multiple objects or events occurring at the same time. Therefore, further improvements in generative models for audio are needed in the context of scene sonification.

6.2.1 Opportunities and Challenges with the AI generative models. During the comparative study, we found several commonly occurring confusions in mapping sounds to scenes, indicating that there are certain distinct and defining characteristics of a scene (e.g. the sound of cutlery and people chatting for a canteen scene or the

sound of waves crashing and receding along with wind for a beach scene) that need to be captured in a sound for it to be easily recognized as the intended scene. Since these characteristics in recorded sounds are not always guaranteed or controllable, therefore existence of these characteristics in the current AI-generated sounds is also not always guaranteed, even after extensive prompt engineering. A possible reason could be that these models have not been designed for the use-case of generating sounds that most intuitively represent scenes. Im2Wav, for example, has been trained to generate sounds corresponding to only one event or object that is in focus, and not based on the overall scene in the image. A scene with multiple events occurring simultaneously becomes harder to generate. Therefore, automatically identifying the most important audio signature(s) of a given scene (in text or image), and generating a highly intuitive and recognizable sound is an open research problem for the AI community.

Most datasets used for training general-purpose audio generative models are not balanced (with a high percentage of music and speech content), which is disadvantageous in the context of generating environmental sounds. Thus there is a possibility that the quality of the generated audio trained on such datasets is biased in favour of the labels for which there was more training data. There is a need to contextualize these general-purpose generative models for the intended tasks and investigate their performance with respect to the needs of the task.

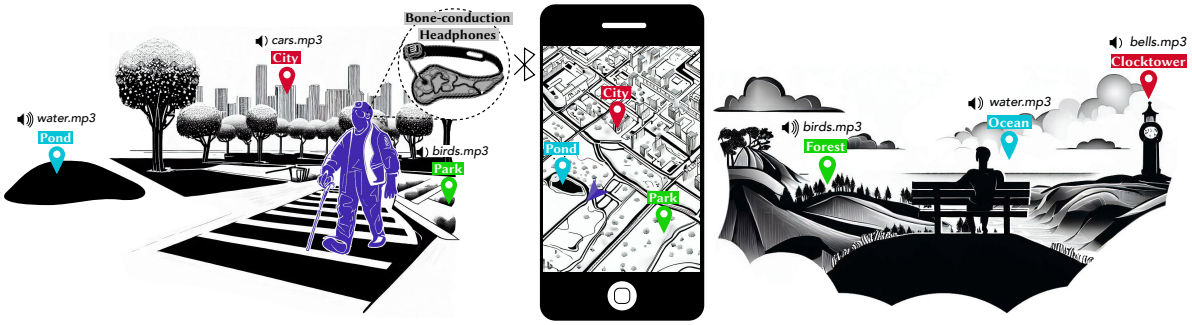
6.3 Addressing Objective 3: Effectiveness of sounds in providing scene awareness in vista-space to PVIs

The effectiveness of the two sonification mechanisms -auditory icons and spearcons, in sonifying the vista scenes for PVIs depended on the situation in which they were employed, i.e. leisure and navigation. The auditory icons were generally preferred in the context of leisure, while spearcons were preferred for direction orientation during navigation.

Through the qualitative feedback during the wizard-of-oz study, auditory icons were considered to be more engaging, pleasant, immersive, and enjoyable than spearcons in experiencing vista scenes during leisure activities (Figure 8). The PVI participants who became visually impaired at a later stage of their life particularly mentioned that they enjoyed the intuitive and immersive experience of the scenes around them provided by the auditory icon soundscapes which met their need to comprehend, imagine, and enjoy the vista scenes. They said that this experience brought back memories of these scenes, and helped them experience the beauty of the scenes without having to depend on others for a description. Also, they said they would use a system that can convert the vista-space scenes into auditory icon soundscapes in leisurely situations in the outdoors, such as when they are seated on top of a hill or a rooftop of a building, or in virtual environments such as games. Moreover, they wished for the sonification of more aesthetic and experiential far-field scenes such as mountains, sea beaches, forests, and lakes.

For the purpose of orientation during navigation, most of the participants preferred spearcons over auditory icons since spearcons provide direct information about the vista space, without having to guess. Most of the participants thought that such a feature would help them during navigation for directional orientation. However, one participant mentioned that he would prefer a combination of auditory icons and spearcons over only spearcons for this purpose. He reasoned that one tends to often miss information in spoken descriptions, they are hard to multitask with, and listening to spearcons for a long time is annoying, whereas auditory icons are hard to miss, pleasant to hear, and one could multitask with them. On the other hand, two participants mentioned that auditory icons could be distracting while walking, and may get mixed with the existing environmental sounds.

Most of the participants also wanted more information from both auditory icons and spearcons for orientation, for example, the distance of the object or scene. All the participants could see themselves using a system that can sonify the vista-space around them while they walk for the purpose of orientation. They wished for the



(a) Directional orientation during navigation (b) Leisure to enjoy the scenes around

Fig. 9. Illustrating the concept of SonicVista: Far objects and scenes are visualized using a mapping of sounds that are arranged in correspondence to the location of PVI, while the volume of the sound indicates the distance from the scene. The user's geolocation as well as the nearby points of interest can be easily fetched from APIs of direction services, such as Google Maps [78]. Bone-conduction headphones, which are often already part of PVI's inventory are then used to display the sounds. PVIs suggested two applications (see a+b).

sonification of various scenes and landmarks that would help with daily necessities such as bus stops, train stations, and supermarkets.

6.4 Key Take-Aways

This is a summary of the key take-aways of this work:

- Vista-scene awareness emerged as a strong requirement in PVIs for cognitive and aesthetic needs. This awareness is specifically needed in two situations - a) at leisure, to have the ability to sense and enjoy scenes around, and b) during navigation for direction orientation.
- For sonic objects and scenes, handcrafted audio recordings and AI-generated sounds perform similarly in terms of intuitiveness and learnability of the sounds, which indicates the potential of using generative models for providing a scalable sound generation method for this purpose.
- Qualitative feedback from PVIs suggested that representative auditory icons could be more engaging, pleasant, immersive, and enjoyable than spearcons, which could be helpful for experiencing vista scenes in the context of leisure. On the other hand, spearcons provided more direct information which could be helpful during navigation.
- Vista scenes, where multiple objects and events occur simultaneously, the direct application of generative models (Im2Wav and AudioLDM) may not always guarantee an intuitive and representative sound, possibly because these models are trained to generate one event or object that is in focus.

6.5 Concept Framework

Based on the findings of our studies, we envision a system that will convert scenes in the vista-space into intuitive and learnable sounds under two usage scenarios, (a) for direction orientation during navigation, and (b) at leisure to enjoy the scenes around (Figure 9). Such a system will consist of several components, including a scene-capturing module, a scene-recognition module, a sonification module, a head-tracking module for spatialization, and the audio output and interaction module. Future work will include the deployment of this system and testing its usability in a real-world scenario. Based on our formative study with PVI, we see the potential of introducing a new technology for vista-space awareness. However, in order to fit into their general assistive technology usage, the vista-space sonification module could be integrated as a feature with the current devices and apps,

which could be switched on upon request. Such a system could therefore be integrated with mobile phones as an app leveraging on the existing hardware and sensors on the phone (camera, GPS, IMU), in conjunction with headphones. The sonification module could also be designed as a middleware with devices for the blind such as Orcam[17] or an additional feature in apps for blind such as LookOut[18] and SeeingAI[19]. The basic functioning of the concept is illustrated in Figure 9.

7 LIMITATIONS AND FUTURE WORK

7.1 Limited number of PVI participants

Due to limited access to PVIs, we had a small number of VI participants, which is a common problem in this research area [76]. However, one could question the generalizability of our findings. We recruited a relatively large group of sighted participants for the listening tests in the sound evaluation study which enabled us to derive statistically significant insights on a large set of sounds (84 sounds), for both AI-generated and audio recordings. However, the motivation of this study was to understand the general quality and characteristics of these sounds and whether these sounds can reliably represent what they were intended to represent. So, the insights drawn from this study also pertained to this general goal and were not specific to the context of their use cases for PVIs. With our limited access to the PVI participants, we tried to optimize the time we had with them. We selected a small subset of sounds that were generally rated as highly intuitive in the sound evaluation study, made sure that this small subset was rated similarly by PVI participants too, and then tested out their use in the contexts relevant to PVI such as spatial awareness during navigation and at leisure. For the wizard-of-oz utility study, although we had a limited number of PVI participants, we conducted a thorough thematic analysis of their qualitative feedback, from which consistent and recurring codes emerged, some of which were common across all the participants. A qualitative assessment with a low number of participants to draw design implications is considered to be a reliable methodology [30, 31], especially when working with a niche population such as PVI [2, 7, 15, 76]. We believe our findings provide useful insights to the research community when aiming to design assistive technologies for PVI. An in-depth sound evaluation study, as well as utility assessment study with a larger number of PVI participants, is needed as part of future work.

7.2 More sophisticated mechanisms for scene sonification needed

For scene sonification, we used the most straightforward method of generating the sounds, the current state-of-the-art image-to-audio and text-to-audio models. However, based on the observations in the sound evaluation study (Section 6.2.1), more sophisticated mechanisms to compose the sounds corresponding to a scene must be designed in the future, such as identifying different sound-making objects within a scene and merging the sounds.

7.3 Need for a real-world application and field study

We acknowledge that the current study provides a preliminary understanding of vista-space awareness through scene sonification through controlled lab environment tests. There is a need to test the ecological validity of sonification designs in a real-world setting. Our findings from the current study will inform future research that will include building a self-contained system that converts vista scenes into sounds through different sonification mechanisms, deploying this system in real-world scenarios, and systematically evaluating its usability, as suggested in Section 6.5. For example, an important insight from this study that could be useful in a future system is that the sounds to sonify vista scenes for orientation during navigation is preferred to be speech-based, while those for leisure should be designed differently so as to improve the experience of the scene. Moreover, the comparable

[17] https://www.orcam.com/en-us/orcam-myeye
[18] https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en&gl=US
[19] https://apps.apple.com/us/app/seeing-ai/id999062298

quality of AI-generated sounds and handcrafted audio recordings indicates that AI-generated sounds could play an important role in building a fully automated system that captures the scene and converts into representative audio, instead of relying on libraries of handcrafted sounds. Building and testing such a self-contained system for scene sonification will have many tasks and challenges of its own, such as handling model response latency, spatialization on the move, etc., which are beyond the scope of the aim of this work, and need to be investigated in future work.

8 CONCLUSION

Through this work, we show the potential of the state-of-the-art generative models for audio in sonifying objects and scenes in the vista space, as they perform similarly to the currently used handcrafted sounds, in terms of intuitiveness and learnability of the sounds. We also show that automatically generating highly intuitive, recognizable, as well as pleasant sounds corresponding to vista-scenes, that consist of multiple event signatures, is still an open research problem. Finally, we show that the idea of sonification of the vista-scenes through such sounds is useful to the PVI community as a cognitive and aesthetic need, and will potentially improve their quality of life. This work opens discussions on the utility of AI models for audio generation in the context of PVIs and their needs. Future work will include improving the performance of the generative models for this context, and building and deploying a system with these sounds to test its usability in a real-world scenario.

ACKNOWLEDGMENTS

We thank all the participants for contributing to this work. We thank NUS School of Computing for providing support through NUS SoC Seed Grant A-8001658-00-00.

REFERENCES

[1] Amandine Afonso-Jaco and Brian FG Katz. 2022. Spatial Knowledge via Auditory Information for Blind Individuals: Spatial Cognition Studies and the Use of Audio-VR. *Sensors* 22, 13 (2022), 4794.

[2] Nida Aziz, Tony Stockman, and Rebecca Stewart. 2019. An investigation into customisable automatically generated auditory route overviews for pre-navigation. (2019).

[3] Cynthia L Bennett, Jane E, Martez E Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How teens with visual impairments take, edit, and share photos on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[4] Cynthia L. Bennett and Daniela K. Rosner. 2019. The Promise of Empathy: Design, Disability, and Knowing the "Other". 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, 1–13. https://doi.org/10.1145/3290605.3300528

[5] Irving Biederman. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review* 94, 2 (1987), 115.

[6] Roger Boldu, Alexandru Dancu, Denys JC Matthies, Thisum Buddhika, Shamane Siriwardhana, and Suranga Nanayakkara. 2018. Fingerreader2. 0: Designing and evaluating a wearable finger-worn camera to assist people with visual impairments while shopping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–19.

[7] Roger Boldu, Denys JC Matthies, Haimo Zhang, and Suranga Nanayakkara. 2020. AiSee: an assistive wearable device to support visually impaired grocery shoppers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–25.

[8] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.

[9] Stephen A Brewster, Peter C Wright, and Alistair DN Edwards. 1993. An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 222–227.

[10] Luis AI Cabrera-Sosa, Luis E Lopez-Garcia, Alejandro Diaz-Sanchez, Jose M Rocha-Perez, Victor H Carbajal-Gomez, and Gregorio Zamora-Mejia. 2020. Design and Implementation of a White-Cane Based on the Vibration-Distance Sensing Technique for Blind People. In *2020 IEEE International Conference on Engineering Veracruz (ICEV)*. IEEE, 1–8.

[11] Kaixin Chen, Yongzhi Huang, Yicong Chen, Haobin Zhong, Lihua Lin, Lu Wang, and Kaishun Wu. 2022. LiSee: A Headphone that Provides All-day Assistance for Blind and Low-vision Users to Reach Surrounding Objects. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–30.

[12] Keunwoo Choi, Jaekwon Im, Laurie Heller, Brian McFee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, and Shinosuke Takamichi. 2023. Foley sound synthesis at the dcase 2023 challenge. *arXiv preprint arXiv:2304.12521* (2023).

[13] Won-Gook Choi and Joon-Hyuk Chang. 2023. *HYU Submission For The DCASE 2023 Task 7: Diffusion Probabilistic Model With Adversarial Training For Foley Sound Synthesis.* Technical Report. Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea, Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea.

[14] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[15] Angela Constantinescu, Karin Müller, Monica Haurilet, Vanessa Petrausch, and Rainer Stiefelhagen. 2020. Bring the environment to life: A sonification module for people with visual impairments to improve situation awareness. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 50–59.

[16] Tilman Dingler, Jeffrey Lindsay, Bruce N Walker, et al. 2008. Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proceedings of the 14th International Conference on Auditory Display, Paris, France*. 1–6.

[17] Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208* (2018).

[18] Gaël Dubus and Roberto Bresin. 2013. A systematic review of mapping strategies for the sonification of physical quantities. *PloS one* 8, 12 (2013), e82491.

[19] Julie Ducasse, Anke M Brock, and Christophe Jouffrais. 2018. Accessible interactive maps for visually impaired users. *Mobility of Visually Impaired People: Fundamentals and ICT Assistive Technologies* (2018), 537–584.

[20] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[21] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. GANSynth: Adversarial Neural Audio Synthesis. In *International Conference on Learning Representations*.

[22] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. 2020. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*.

[23] Rodwan Hashim Mohammed Fallatah, Jawad Syed, Rodwan Hashim Mohammed Fallatah, and Jawad Syed. 2018. A critical review of Maslow's hierarchy of needs. *Employee motivation in Saudi Arabia: An investigation into the higher education sector* (2018), 19–59.

[24] Emerson Foulke. 1968. Listening comprehension as a function of word rate. *Journal of Communication* 18, 3 (1968), 198–206.

[25] William W Gaver. 1987. Auditory icons: Using sound in computer interfaces. *ACM SIGCHI Bulletin* 19, 1 (1987), 74.

[26] William W Gaver. 1993. Synthesizing auditory icons. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 228–235.

[27] James J Gibson. 1978. The ecological approach to the visual perception of pictures. *Leonardo* 11, 3 (1978), 227–235.

[28] Cole Gleason, Dragan Ahmetovic, Saiph Savage, Carlos Toxtli, Carl Posthuma, Chieko Asakawa, Kris M Kitani, and Jeffrey P Bigham. 2018. Crowdsourcing the installation and maintenance of indoor localization infrastructure to support blind navigation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–25.

[29] Cole Gleason, Alexander J Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2018. FootNotes: Geo-referenced audio annotations for nonvisual exploration. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[30] Nielsen Norman Group. 2012. How Many Test Users in a Usability Study? https://www.nngroup.com/articles/how-many-test-users/. [Online; accessed 23-Jan-2024].

[31] Nielsen Norman Group. 2021. How to Conduct Usability Studies for Accessibility. https://media.nngroup.com/media/reports/free/How_to_Conduct_Usability_Studies_for_Accessibility.pdf. [Online; accessed 23-Jan-2024].

[32] Chitralekha Gupta, Purnima Kamath, Yize Wei, Zhuoyao Li, Suranga Nanayakkara, and Lonce Wyse. 2023. Towards Controllable Audio Texture Morphing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[33] Susumu Harada, Daisuke Sato, Dustin W Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. 2013. Accessible photo album: enhancing the photo sharing experience for people with visual impairment. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2127–2136.

[34] Thomas Hermann, Andy Hunt, John G Neuhoff, et al. 2011. *The sonification handbook*. Vol. 1. Logos Verlag Berlin.

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[36] Weijian Hu, Kaiwei Wang, Kailun Yang, Ruiqi Cheng, Yaozu Ye, Lei Sun, and Zhijie Xu. 2020. A comparative study in real-time scene sonification for visually impaired people. *Sensors* 20, 11 (2020), 3222.

[37] Jochen Huber, Roy Shilkrot, Pattie Maes, and Suranga Nanayakkara. 2018. *Assistive augmentation*. Springer.

[38] Vladimir Iashin and Esa Rahtu. 2021. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*.

[39] 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. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–28.

[40] Tiger F Ji, Brianna R Cochran, and Yuhang Zhao. 2022. Demonstration of VRBubble: Enhancing Peripheral Avatar Awareness for People with Visual Impairments in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.

[41] Zihao Ji, Weijian Hu, Ze Wang, Kailun Yang, and Kaiwei Wang. 2021. Seeing through events: Real-time moving object sonification for visually impaired people using event-based camera. *Sensors* 21, 10 (2021), 3558.
[42] Wenqiang Jin, Mingyan Xiao, Huadi Zhu, Shuchisnigdha Deb, Chen Kan, and Ming Li. 2020. Acoussist: An acoustic assisting tool for people with visual impairments to cross uncontrolled streets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–30.
[43] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. 2017. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5839–5849.
[44] Purnima Kamath, Chitralekha Gupta, Lonce Wyse, and Suranga Nanayakkara. 2023. Example-Based Framework for Perceptually Guided Audio Texture Generation. *arXiv preprint arXiv:2308.11859* (2023).
[45] Purnima Kamath, Tasnim Nishat Islam, Chitralekha Gupta, Lonce Wyse, and Suranga Nanayakkara. 2023. *DCASE Task-7: StyleGAN2-Based Foley Sound Synthesis.* Technical Report. National University of Singapore, Singapore and Bangladesh University of Engineering and Technology, Bangladesh and Universitat Pompeu Fabra, Barcelona, Spain.
[46] Kevin Karplus and Alex Strong. 1983. Digital synthesis of plucked-string and drum timbres. *Computer Music Journal* 7, 2 (1983), 43–55.
[47] Brian FG Katz, Florian Dramas, Gaëtan Parseihian, Olivier Gutierrez, Slim Kammoun, Adrien Brilhault, Lucie Brunet, Mathieu Gallay, Bernard Oriola, Malika Auvray, et al. 2012. NAVIG: Guidance system for the visually impaired using virtual augmented reality. *Technology and Disability* 24, 2 (2012), 163–178.
[48] Brian FG Katz, Slim Kammoun, Gaëtan Parseihian, Olivier Gutierrez, Adrien Brilhault, Malika Auvray, Philippe Truillet, Michel Denis, Simon Thorpe, and Christophe Jouffrais. 2012. NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality* 16 (2012), 253–269.
[49] Seita Kayukawa, Tatsuya Ishihara, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2020. Guiding blind pedestrians in public spaces by understanding walking behavior of nearby pedestrians. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–22.
[50] Seita Kayukawa, Daisuke Sato, Masayuki Murata, Tatsuya Ishihara, Akihiro Kosugi, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2022. How Users, Facility Managers, and Bystanders Perceive and Accept a Navigation Robot for Visually Impaired People in Public Buildings. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 546–553.
[51] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).
[52] Julian Kreimeier and Timo Götzelmann. 2019. First steps towards walk-in-place locomotion and haptic feedback in virtual reality for visually impaired. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
[53] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation. In *International Conference on Learning Representations (ICLR)*.
[54] Bineeth Kuriakose, Raju Shrestha, and Frode Eika Sandnes. 2023. DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments. *Expert Systems with Applications* 212 (2023), 118720.
[55] Franklin Mingzhe Li, Franchesca Spektor, Meng Xia, Mina Huh, Peter Cederberg, Yuqi Gong, Kristen Shinohara, and Patrick Carrington. 2022. "It Feels Like Taking a Gamble": Exploring Perceptions, Practices, and Challenges of Using Makeup and Cosmetics for People with Visual Impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
[56] Jingyi Li, Son Kim, Joshua A Miele, Maneesh Agrawala, and Sean Follmer. 2019. Editing spatial layouts through tactile templates for people with visual impairments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[57] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. In *Proceedings of ICML*.
[58] Abraham H Maslow. 1975. *Motivation and personality.* Harper & Row.
[59] Saul McLeod. 2007. Maslow's hierarchy of needs. *Simply psychology* 1 (2007), 1–18.
[60] Susanna Millar. 1988. Models of sensory deprivation: The nature/nurture dichotomy and spatial representation in the blind. *International Journal of Behavioral Development* 11, 1 (1988), 69–87.
[61] Daniel R Montello. 1993. Scale and multiple psychologies of space. In *European conference on spatial information theory*. Springer, 312–321.
[62] M.C. Mozer. 2001. Object Recognition: Theories. In *International Encyclopedia of the Social & Behavioral Sciences*, Neil J. Smelser and Paul B. Baltes (Eds.). Pergamon, Oxford, 10781–10785. https://doi.org/10.1016/B0-08-043076-7/01459-5
[63] Austin M Mulloy, Cindy Gevarter, Megan Hopkins, Kevin S Sutherland, and Sathiyaprakash T Ramdoss. 2014. Assistive technology for students with visual impairments and blindness. *Assistive technologies for people with diverse abilities* (2014), 113–156.
[64] Javier Nistal, Stefan Lattner, and Gaël Richard. 2020. DrumGAN: Synthesis of Drum Sounds With Timbral Feature Conditioning Using Generative Adversarial Networks. In *International Society for Music Information Retrieval Conference*.
[65] Dianne K Palladino and Bruce N Walker. 2007. Learning rates for auditory menus enhanced with spearcons versus earcons. In *Proceedings of the 13th international conference on auditory display*. 274–279.

[66] Amy T Parker, Martin Swobodzinski, Julie D Wright, Kyrsten Hansen, Becky Morton, and Elizabeth Schaller. 2021. Wayfinding tools for people with visual impairments in real-world settings: a literature review of recent studies. In *Frontiers in Education*, Vol. 6. Frontiers Media SA, 723816.

[67] Shishir G Patil, Don Kurian Dennis, Chirag Pabbaraju, Nadeem Shaheer, Harsha Vardhan Simhadri, Vivek Seshadri, Manik Varma, and Prateek Jain. 2019. Gesturepod: Enabling on-device gesture-based interaction for white cane users. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 403–415.

[68] Helen Petrie, Valerie Johnson, Thomas Strothotte, Andreas Raab, Steffi Fritz, and Rainer Michel. 1996. MoBIC: Designing a travel aid for blind and elderly people. *The Journal of Navigation* 49, 1 (1996), 45–52.

[69] Lorenzo Picinali, Amandine Afonso, Michel Denis, and Brian FG Katz. 2014. Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge. *International Journal of Human-Computer Studies* 72, 4 (2014), 393–407.

[70] Giorgio Presti, Dragan Ahmetovic, Mattia Ducci, Cristian Bernareggi, Luca Ludovico, Adriano Baratè, Federico Avanzini, and Sergio Mascetti. 2019. WatchOut: Obstacle sonification for people with visual impairment or blindness. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 402–413.

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[72] Kyle Rector, Keith Salmon, Dan Thornton, Neel Joshi, and Meredith Ringel Morris. 2017. Eyes-free art: Exploring proxemic audio interfaces for blind and low vision art engagement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

[73] Pablo Revuelta Sanz, Belén Ruiz Mezcua, José M Sánchez Pena, and Bruce N Walker. 2014. Scenes and images into sounds: a taxonomy of image sonification methods for mobility applications. *Journal of the Audio Engineering Society* 62, 3 (2014), 161–171.

[74] Stefano Scheggi, A Talarico, and Domenico Prattichizzo. 2014. A remote guidance system for blind and visually impaired people via vibrotactile haptic feedback. In *22nd Mediterranean conference on control and automation*. IEEE, 20–23.

[75] Eldon Schoop, James Smith, and Bjoern Hartmann. 2018. Hindsight: enhancing spatial awareness by sonifying detected objects in real-time 360-degree video. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[76] Andrew Sears and Vicki Hanson. 2011. Representing users in accessibility research. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2235–2238.

[77] Hongje Seong, Junhyuk Hyun, Hyunbae Chang, Suhyeon Lee, Suhan Woo, and Euntai Kim. 2019. Scene Recognition via Object-to-Scene Class Conversion: End-to-End Training. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–6. https://doi.org/10.1109/IJCNN.2019.8852040

[78] Monika Sharma and Sudha Morwal. 2015. Location Tracking using Google Geolocation API. *Int. J. of Sci. Tech. & Eng* 1, 11 (2015), 29–32.

[79] Roy Sheffer and Yossi Adi. 2023. I Hear Your True Colors: Image Guided Audio Generation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096023

[80] Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. 2015. FingerReader: a wearable device to explore printed text on the go. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2363–2372.

[81] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. 2021. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems* 34 (2021), 1415–1428.

[82] Lee Stearns, Uran Oh, Leah Findlater, and Jon E Froehlich. 2018. Touchcam: Realtime recognition of location-specific on-body gestures to support users with visual impairments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23.

[83] Thomas Strothotte, Steffi Fritz, Rainer Michel, Andreas Raab, Helen Petrie, Valerie Johnson, Lars Reichert, and Axel Schalt. 1996. Development of dialogue systems for a mobility aid for blind people: initial design and usability testing. In *Proceedings of the second annual ACM conference on Assistive technologies*. 139–144.

[84] Garreth W Tigwell, Benjamin M Gorman, and Rachel Menzies. 2020. Emoji accessibility for visually impaired people. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[85] Wallace Ugulino and Hugo Fuks. 2015. Landmark identification with wearables for supporting spatial awareness by blind persons. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 63–74.

[86] Bruce N Walker and Lisa M Mauney. 2010. Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. *ACM Transactions on Accessible Computing (TACCESS)* 2, 3 (2010), 1–16.

[87] Bruce N Walker, Amanda Nance, and Jeffrey Lindsay. 2006. Spearcons (speech-based earcons) improve navigation performance in auditory menus.. In *Proceedings of the 12th International Conference on Auditory Display*. 63–68.

[88] Sue Watkinson and Eileen Scott. 2004. Managing the care of patients who have visual impairment. *Nursing times* 100, 1 (2004), 40–42.

[89] Keith White et al. 1991. Training Program for Individuals Working with Older American Indians Who Are Blind or Visually Impaired. Training Manual. (1991).

[90] Lonce Wyse, Purnima Kamath, and Chitralekha Gupta. 2022. Sound Model Factory: An Integrated System Architecture for Generative Audio Modelling. In *Artificial Intelligence in Music, Sound, Art and Design: 11th International Conference, EvoMUSART 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20–22, 2022, Proceedings*. Springer, Springer, 308–322. https://doi.org/10.1007/978-3-031-03789-4_20

[91] Shuchang Xu, Ciyuan Yang, Wenhao Ge, Chun Yu, and Yuanchun Shi. 2020. Virtual Paving: Rendering a smooth path for people with visual impairment through vibrotactile and audio feedback. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–25.

[92] Ciyuan Yang, Shuchang Xu, Tianyu Yu, Guanhong Liu, Chun Yu, and Yuanchun Shi. 2021. LightGuide: Directing Visually Impaired People along a Path Using Light Cues. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–27.

[93] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[94] Alireza Zare, Kyla McMullen, and Christina Gardner-McCune. 2014. Design of an accessible and portable system for soccer players with visual impairments. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 1237–1242.

[95] 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*. 1–14.

[96] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A face recognition application for people with visual impairments: Understanding use beyond the lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

APPENDIX

A PROCEDURE AND QUESTIONNAIRE FOR THE FORMATIVE STUDY

We recruited participants through local community email channels. We informed them about the goal of this study through the participant information sheet before they decided to participate in the study. This study was done with the participants individually, and there were two researchers conducting the interview. Once the participant was comfortably seated, we began by introducing ourselves, describing the goal of the study again, and then we started a semi-structured interview with the following prompting open-ended questions. This interview lasted for 20 to 30 minutes per participant.

Regarding vista-spatial awareness:

(1) What is your process of understanding the scenes around you when you walk from A to B?
(2) When you are on top of a building or a hill, what is your process of experiencing or thinking about the sceneries around you?
(3) What is your process of orienting yourself when you are outside? For example, do you make a mental map of the 3D or 2D space around you to orient yourself?
(4) Do you think being aware of the scenes around you could be useful to you?

Regarding their current assistive technology usage:

(1) What kinds of assistive technologies do you currently use? (if they talk about mobile apps, then) Are there some specific apps that you find helpful?
(2) What are the modalities in which you get feedback in your assistive tools (audio, haptic?) -
(3) What kind of audio feedback - speech, earcons, spearcons?
(4) Would you find it useful to hear audio cues to become aware of objects or scenes?
(5) Do you use earphones/bone-conduction headphones?

B PROCEDURE AND QUESTIONNAIRE FOR THE WIZARD-OF-OZ STUDY

Similar to the formative study, we recruited the same participants informing them about the study prior to the study through an email. When they arrived, we described the project, conducted a short intuitiveness and learnability test with the subset of sounds, and then explained the two listening tests (central awareness and peripheral awareness tests) and their setup. They wore the bone-conduction headphones that we provided and listened to the sounds. We mentioned that both the tests are going to be a simulation of a system that sonifies the scenes around, and that we are only testing the utility of the sounds.

Before starting the central awareness test, we mention, “Imagine you are sitting on a bench on top of a hill. Now you hear the sounds through the headphones, that are simulating the scenes around you through sounds. You can turn your head around in a 180-degree view to experience the scenes spatially (four auditory icons)”.

After each of the tests with each audio type (spearcons or auditory icons), we asked them the following open-ended questions.

(1) Describe the scenes you just heard (scene names, and direction in which they occur, i.e. four directions in the front in central vista test, and right or left and sequence in peripheral vista test).
(2) So at the moment this was a simulation. But lets say in a real-world situation, there is an app or a device that converts the scenes around you into these kinds of sounds. Would these sounds enhance your experience of the scenes around you?
(3) Would you like to share any other thoughts about this experience?
(4) Would you see yourself using a friendly system that has these sounds for ambient awareness?
(5) Do you think this awareness will help you orient yourself?

(6) Do you see any use cases for such a system? When and how would you think you will be using it? What are the other kinds of scenes you would want to sonify?

C ADDITIONAL DETAILS OF THE RESULTS REPORTED IN SECTION 4

Table 4 provides a detailed performance of intuitiveness and learnability scores (out of 3.0) for sighted participants. Table 5 provides the average self-report intuitiveness ratings (out of 5.0), pleasantness, and comfort level.

Additional Notable Points:

- According to the self-report ratings of intuitiveness, the participants found the AudioLDM sounds for *door*, *stairs*, and *bush* to be more intuitive than the other two sound types. The AudioLDM-generated sound of *stairs* was somewhat more learnable than the other two sound types because the sound was distinct compared to the other sounds in the AudioLDM set of sounds and clearly represented footsteps, while the handcrafted sound for *stairs* got confused with wall, door, and chair.
- The Im2Wav sound for *bench* was not intuitive but was clearly more learnable than the other two sound types for this object. The image used to generate this sound had a bench situated in a park, so the generated sound contained some music, implying music being played in a park, which was a distinctive feature of this sound and the participants were able to learn it. On the other hand, in the learnability test, the handcrafted *bench* sound was confused with other possible bench-like wooden objects such as chair, and fence, while the AudioLDM *bench* sound, which had a rhythmic pattern, was confused with train.
- Interestingly, the two AI-generated sounds for *person/people* were significantly more intuitive and learnable than the handcrafted sound, possibly because the handcrafted sound was designed to capture the movement of a person which got incorrectly recognized as the sound of a bush, while the AI sounds had unintelligible voices of people which were easier to recognize.
- The handcrafted and AudioLDM sounds for the *kids playing* and *park* scenes were highly intuitive as they had distinct signatures of these scenes, i.e. high pitched children's voices and sound of birds and wind respectively. The corresponding Im2Wav sounds were dominated by music which got highly confused with the *mall* scene.

Table 4. Intuitiveness and Learnability scores (out of 3.0) [Mean ± Std. Dev.] from Phase 1 of the study. The sound type acronyms are - H: Handcrafted, A: AudioLDM, I: Im2Wav. (↑ means higher is better)

	Intuitiveness Scores ↑				**Learnability Scores↑**			
Sound Type→	H	H	A	I	H	H	A	I
Participants→	(PVI) [15]	(Sighted)			(PVI) [15]	(Sighted)		
Non-Sonic Objects								
bench	1.00 ±0.00	1.17 ±0.37	1.08 ±0.28	1.0 ±0.00	2.20 ±0.84	1.09 ±0.29	1.33 ±0.47	1.67 ±0.47
building	-	1.18 ±0.39	1.0 ±0.00	1.17 ±0.37	-	1.67 ±0.62	1.27 ±0.45	1.55 ±0.50
bush	2.40 ±0.89	1.45 ±0.50	1.25 ±0.43	1.0 ±0.00	2.40 ±0.80	2.08 ±0.49	1.36 ±0.48	1.27 ±0.45
chair	1.40 ±0.89	1.67 ±0.62	1.0 ±0.00	1.09 ±0.29	2.80 ±0.45	2.09 ±0.90	1.58 ±0.49	1.27 ±0.45
cycle	2.80 ±0.45	1.92 ±0.86	1.0 ±0.00	1.0 ±0.00	2.40 ±0.89	2.50 ±0.76	1.25 ±0.60	1.18 ±0.39
door	2.60 ±0.55	1.64 ±0.64	1.67 ±0.47	1.18 ±0.39	2.80 ±0.45	1.50 ±0.67	2.18 ±0.57	1.83 ±0.37
fence	2.00 ±1.00	1.58 ±0.76	1.09 ±0.29	1.27 ±0.45	2.40 ±0.89	2.00 ±0.91	1.17 ±0.37	1.09 ±0.29
stop-sign	1.00 ±0.00	1.36 ±0.48	1.0 ±0.00	1.08 ±0.28	2.75 ±0.50	2.18 ±0.94	1.09 ±0.29	1.33 ±0.47
stairs	1.40 ±0.89	1.75 ±0.60	1.73 ±0.62	1.0 ±0.00	2.60 ±0.55	1.73 ±0.62	2.0 ±0.43	1.33 ±0.47
wall	1.20 ±0.45	1.18 ±0.57	1.09 ±0.29	1.09 ±0.29	2.60 ±0.55	2.18 ±0.83	1.67 ±0.75	1.83 ±0.55
Overall	-	1.50 ±0.65	1.19 ±0.41	1.09 ±0.28	-	1.90 ±0.82	1.49 ±0.61	1.44 ±0.51
Sonic Objects								
bells	-	2.64 ±0.64	2.64 ±0.64	2.36 ±0.88	-	2.91 ±0.29	2.82 ±0.57	2.73 ±0.62
birds	-	2.92 ±0.28	2.83 ±0.37	3.0 ±0.00	-	3.0 ±0.00	2.92 ±0.28	3.0 ±0.00
bus	2.00 ±0.71	1.0 ±0.00	1.09 ±0.2	1.18 ±0.27	2.20 ±0.84	1.42 ±0.49	1.33 ±0.47	1.25 ±0.43
car	1.40 ±0.89	1.08 ±0.28	1.45 ±0.50	1.50 ±0.50	2.60 ±0.55	1.17 ±0.55	1.36 ±0.48	1.75 ±0.60
dog	2.40 ±0.89	1.92 ±0.64	1.92 ±0.28	2.00 ±0.00	2.80 ±0.45	2.45 ±0.66	2.25 ±0.43	2.17 ±0.37
ducks	-	2.91 ±0.29	2.08 ±0.76	2.42 ±0.86	-	3.00 ±0.00	2.5 ±0.65	2.73 ±0.62
traffic light	1.80 ±0.84	1.75 ±0.83	1.09 ±0.29	1.17 ±0.37	2.60 ±0.55	2.33 ±0.62	1.17 ±0.37	1.83 ±0.37
motorbike	2.40 ±0.89	2.00 ±0.82	1.27 ±0.45	1.25 ±0.43	2.60 ±0.55	2.45 ±0.78	1.91 ±0.51	1.55 ±0.66
person/people	1.40 ±0.98	1.09 ±0.29	2.00 ±0.00	2.00 ±0.00	2.75 ±0.50	1.36 ±0.64	2.45 ±0.50	2.08 ±0.28
railway	1.60 ±0.89	1.55 ±0.50	1.45 ±0.50	1.17 ±0.37	2.60 ±0.55	1.75 ±0.72	1.50 ±0.50	1.25 ±0.43
train	2.00 ±0.00	2.18 ±0.83	1.17 ±0.37	1.36 ±0.48	2.84 ±0.45	1.83 ±0.80	1.73 ±0.45	1.64 ±0.48
truck	1.00 ±0.00	1.00 ±0.00	1.08 ±0.28	1.17 ±0.37	3.00 ±0.00	1.33 ±0.62	1.45 ±0.50	1.45 ±0.50
Overall	-	1.83 ±0.87	1.68 ±0.73	1.70 ±0.77	-	2.07 ±0.88	1.95 ±0.76	1.94 ±0.74
Scenes								
beach	-	1.58 ±0.86	2.00 ±0.95	2.25 ±0.83	-	2.55 ±0.66	2.33 ±0.75	2.75 ±0.60
canteen	-	2.73 ±0.62	1.09 ±0.2	1.18 ±0.29	-	2.82 ±0.57	2.09 ±0.90	2.33 ±0.62
kids playing	-	2.64 ±0.77	2.75 ±0.43	1.67 ±0.85	-	2.92 ±0.28	2.27 ±0.86	2.64 ±0.64
mall	-	1.83 ±0.53	2.0 ±0.61	1.91 ±0.63	-	2.45 ±0.55	2.55 ±0.46	2.33 ±0.49
park	-	2.27 ±0.68	2.64 ±0.45	1.25 ±0.29	-	2.42 ±0.57	2.58 ±0.42	2.18 ±0.5
street	-	3.0 ±0.00	1.25 ±0.39	2.27 ±0.53	-	2.92 ±0.18	2.08 ±0.57	2.64 ±0.34
Overall	-	2.33 ±0.90	1.96 ±0.92	1.75 ±0.84	-	2.68 ±0.65	2.32 ±0.81	2.48 ±0.67

Table 5. Self Reported Ratings from Web Survey of the sighted participants (out of 5.0) [Mean ± Std. Dev.] from Phase 1 of the study. The sound type acronyms are - H: Handcrafted, A: AudioLDM, I: Im2Wav. (↑ means higher is better)

	Intuitiveness↑			**Pleasantness↑**			**Comfort-level↑**		
Type→	H	A	I	H	A	I	H	A	I
Non-Sonic Objects									
bench	1.5 ±0.93	1.21 ±0.41	1.08 ±0.28	2.25 ±1.15	2.33 ±1.13	3.29 ±1.08	1.92 ±1.06	2.12 ±1.12	2.88 ±1.3
building	2.33 ±1.05	1.42 ±0.97	1.25 ±0.61	3.25 ±1.07	2.08 ±1.14	1.71 ±0.69	2.25 ±1.15	1.71 ±0.91	1.46 ±0.66
bush	2.0 ±1.06	2.54 ±1.14	1.38 ±0.58	2.38 ±0.88	2.38 ±0.92	2.08 ±1.06	2.04 ±1.12	2.29 ±0.95	1.96 ±1.04
chair	3.33 ±0.87	1.5 ±0.66	1.04 ±0.2	2.83 ±1.05	2.75 ±1.11	1.71 ±0.69	2.08 ±1.06	2.17 ±0.92	1.54 ±0.78
bicycle	2.62 ±1.41	1.46 ±0.66	1.42 ±0.65	2.67 ±1.05	1.75 ±0.94	2.25 ±0.85	2.54 ±1.1	1.62 ±0.92	2.21 ±1.06
door	2.46 ±1.06	3.5 ±1.25	1.54 ±0.98	2.83 ±0.96	2.21 ±0.93	1.71 ±0.69	2.38 ±1.01	1.79 ±0.83	1.46 ±0.59
fence	2.33 ±1.34	1.58 ±0.93	1.08 ±0.41	2.46 ±0.98	2.54 ±0.98	1.79 ±0.78	2.08 ±1.02	2.42 ±1.1	1.71 ±0.81
stop sign	2.0 ±1.41	1.17 ±0.38	1.58 ±0.83	2.71 ±1.08	2.08 ±0.78	1.38 ±0.65	1.92 ±1.06	2.0 ±1.02	1.21 ±0.59
stairs	2.79 ±1.5	4.17 ±0.82	1.21 ±0.51	2.92 ±1.1	3.29 ±1.0	2.04 ±0.81	1.92 ±1.02	2.79 ±1.14	1.92 ±0.88
wall	2.46 ±1.22	2.17 ±1.31	1.46 ±0.78	2.83 ±1.31	2.88 ±0.95	2.21 ±1.1	2.29 ±1.08	2.67 ±1.09	1.83 ±0.96
Overall	2.38 ±0.47	2.07 ±0.98	1.30 ±0.19	2.71 ±0.28	2.43 ±0.43	2.02 ±0.50	2.14 ±0.20	2.16 ±0.37	1.82 ±0.45
Sonic Objects									
bells	4.96 ±0.2	4.75 ±0.68	4.5 ±0.72	3.62 ±1.06	2.67 ±0.96	2.88 ±1.15	2.67 ±1.17	2.17 ±1.13	2.0 ±1.14
birds	5.0 ±0.0	5.0 ±0.0	4.92 ±0.41	4.04 ±1.08	3.58 ±0.93	3.62 ±1.1	3.67 ±1.24	3.29 ±1.08	3.12 ±1.36
bus	1.88 ±0.95	2.04 ±0.91	1.83 ±0.87	2.29 ±1.12	2.25 ±0.85	2.25 ±0.85	1.83 ±1.01	2.17 ±1.01	2.04 ±0.69
car	1.58 ±0.88	2.83 ±0.92	3.88 ±0.74	2.54 ±0.98	2.79 ±0.98	2.67 ±1.05	2.04 ±0.95	2.88 ±0.95	2.25 ±0.99
dog	3.58 ±1.41	4.38 ±0.77	4.83 ±0.38	2.62 ±1.1	2.92 ±1.06	3.38 ±0.92	1.96 ±1.16	2.33 ±1.01	2.54 ±1.22
ducks	4.88 ±0.34	3.88 ±1.23	3.38 ±1.21	3.46 ±1.02	3.0 ±1.02	2.46 ±1.1	3.17 ±1.13	2.42 ±0.97	2.25 ±1.26
traffic light	3.29 ±1.6	1.54 ±0.93	2.12 ±0.99	2.96 ±1.16	2.71 ±0.81	2.62 ±0.82	2.21 ±1.22	2.54 ±1.18	2.42 ±1.1
motorbike	3.62 ±1.35	3.21 ±1.22	3.04 ±1.4	2.83 ±1.09	2.88 ±0.85	3.0 ±0.98	2.21 ±1.18	2.46 ±1.02	2.58 ±1.21
person/people	1.08 ±0.28	4.96 ±0.2	4.79 ±0.51	2.42 ±0.93	3.5 ±0.93	3.38 ±0.97	2.17 ±1.13	3.21 ±1.22	2.88 ±0.9
railway	2.25 ±1.15	3.17 ±1.13	2.54 ±1.28	2.33 ±1.13	2.46 ±0.83	2.38 ±0.92	1.88 ±0.99	2.21 ±0.78	2.04 ±1.0
train	3.25 ±1.42	1.79 ±1.14	2.79 ±1.38	2.17 ±1.17	2.83 ±1.09	2.54 ±0.98	1.67 ±0.92	2.58 ±1.02	2.33 ±0.96
truck	1.62 ±0.82	2.5 ±1.06	1.79 ±0.78	2.12 ±0.95	2.62 ±0.97	2.42 ±1.02	1.79 ±1.06	2.54 ±1.1	2.21 ±1.02
Overall	3.08 ±1.34	3.34 ±1.19	3.37 ±1.14	2.78 ±0.60	2.85 ±0.37	2.80 ±0.43	2.27 ±0.58	2.57 ±0.36	2.39 ±0.33
Scenes									
beach	2.83 ±1.17	3.08 ±1.41	3.25 ±1.39	3.42 ±0.88	2.79 ±1.06	2.75 ±1.22	3.46 ±0.93	2.71 ±1.3	2.58 ±1.41
canteen	4.75 ±0.44	2.54 ±1.22	1.83 ±0.96	3.42 ±0.97	2.75 ±1.22	1.96 ±0.81	3.33 ±1.13	2.46 ±1.18	1.75 ±1.07
kids playing	4.96 ±0.2	3.83 ±1.34	3.25 ±1.03	3.79 ±0.83	2.92 ±1.1	2.5 ±0.88	3.46 ±1.14	2.5 ±1.29	2.0 ±0.88
mall	3.83 ±0.96	3.46 ±1.1	2.75 ±1.19	3.42 ±0.97	2.96 ±1.12	2.54 ±0.83	3.08 ±1.25	2.5 ±1.18	2.42 ±1.02
park	4.04 ±1.08	3.42 ±1.06	1.54 ±0.78	3.83 ±0.92	3.04 ±1.2	3.42 ±0.83	3.62 ±1.13	2.79 ±1.02	3.12 ±0.99
street	4.62 ±0.58	2.75 ±1.11	3.21 ±1.22	3.29 ±1.04	2.75 ±0.79	2.54 ±0.93	2.88 ±1.19	2.58 ±0.88	1.83 ±0.7
Overall	4.17 ±0.72	3.18 ±0.44	2.64 ±0.70	3.53 ±0.20	2.87 ±0.11	2.62 ±0.43	3.30 ±0.25	2.59 ±0.12	2.28 ±0.48