Semantic and Spatial Sound-Object Recognition for Assistive Navigation

Daniel Gea
Faculty of Engineering, University of Porto
dngeap@gmail.com

Gilberto Bernardes
INESC TEC, Faculty of Engineering,
University of Porto
gba@fe.up.pt

ABSTRACT

Building on theories of human sound perception and spatial cognition, this paper introduces a sonification method that facilitates navigation by auditory cues. These cues help users recognize objects and key urban architectural elements, encoding their semantic and spatial properties using non-speech audio signals. The study reviews advances in object detection and sonification methodologies, proposing a novel approach that maps semantic properties (i.e., material, width, interaction level) to timbre, pitch, and gain modulation and spatial properties (i.e., distance, position, elevation) to gain, panning, and melodic sequences. We adopt a three-phase methodology to validate our method. First, we selected sounds to represent the object's materials based on the acoustic properties of crowdsourced annotated samples. Second, we conducted an online perceptual experiment to evaluate intuitive mappings between sounds and object semantic attributes. Finally, in-person navigation experiments were conducted in virtual reality to assess semantic and spatial recognition. The results demonstrate a notable perceptual differentiation between materials, with a global accuracy of $.69 \pm .13$ and a mean navigation accuracy of $.73 \pm .16$, highlighting the method's effectiveness. Furthermore, the results suggest a need for improved associations between sounds and objects and reveal demographic factors that are influential in the perception of sounds.

1. INTRODUCTION

In a world where sight is often prioritized, blind and visually impaired (BVI) individuals face significant challenges in semantic and spatial navigation. When a sighted person encounters an obstacle, they intuitively navigate around it. For BVI individuals using a sound-based system, a simple beep lacks the detail needed for intuitive navigation. Using the same beep for all obstacles creates ambiguity, as different objects require different responses. Although vision allows us to form a mental map of our surroundings, a binary sound system limits spatial awareness, making navigation in unfamiliar spaces challenging for blind individuals [1].

In this context, assistive technology for BVI individuals calls for two levels of recognition: semantic and spatial. Semantic navigation consists of recognizing the attributes and functions of objects, allowing us to distinguish between *obstacles* to be circumvented and *delimiters*, i.e., fixed boundaries that define the limits of accessible areas. Spatial navigation involves understanding physical layouts, i.e., identifying the distances, positions, and elevations of objects.

In recent decades, obstacle detection systems have increasingly relied on computer vision, utilizing cameras and algorithms to identify obstacles for users [2]. While ultrasonic and laser detection are traditionally more accurate, camera-based methods are becoming popular due to their accessibility on devices like smartphones and wearable glasses. These systems detect obstacles and identify features like color and texture, enhancing usability. However, computer vision systems face limitations in real-world conditions, such as low lighting, adverse weather, and visual obstructions, along with current technological constraints.

This study proposes a sonification method, i.e., the process of converting data into non-speech audio signals [3], to translate semantic and spatial object-related information into non-speech audio cues, enhancing the BVI individuals' understanding of their surroundings. In the context of this study, sonification translates spatial and semantic information about objects into sound cues that can assist users in navigating. Semantic information helps distinguish elements like pavements, traffic lights, and doors, while spatial information conveys objects' left or right position, distance, and elevation.

The novelty of the proposed method lies in sonification mappings that convey semantic (object function, width, and interaction) and spatial (distance, position, elevation) properties designed explicitly for BVI navigation. The proposed mappings are tested in perceptual and navigation experiments to evaluate their effectiveness.

The remainder of this paper is organized as follows. Section 2 reviews audio-driven semantic and spatial navigation strategies. Section 3 details the proposed sonification method, namely the mappings between objects and their timbral qualities and the interactions between users and objects with sound transformations. Sections 4 and 5 present the evaluation and results of the sonification method. Finally, Section 6 concludes the paper by highlighting its main contributions and directions for future work.

DOI: https://doi.org/10.5281/zenodo.15032513

2. RELATED WORK

Auditory display methods attempt to map data to sound for intuitive understanding, addressing challenges such as cognitive load and auditory fatigue. These approaches prioritize user-friendly designs that balance clarity, aesthetics, and functionality, especially for visually impaired users [4].

Auditory displays include sonification, auditory icons, and musification methods, which have been instrumental in conveying data to users, especially those with visual impairments [5]. These auditory approaches lie on a symbolic-analogic continuum, where symbolic sounds like earcons serve as basic notifications. In contrast, more analogic displays such as auditory icons reflect event-specific details, enhancing user interaction with relevant context [6–8].

Sonification techniques in outdoor navigation experiments enable users to interpret information despite background noise, reducing cognitive load and overcoming language barriers [9]. This technique employs auditory attributes such as pitch, timbre, amplitude, and spatiality to establish intuitive connections between data characteristics and sound, fostering greater understanding and situational awareness.

Auditory icons and earcons serve as distinct event indicators. While auditory icons leverage intuitive associations—like glass breaking to signal an error—earcons require a learned association between abstract tones and data elements [5]. Morphocons presents another innovative auditory display technique. Unlike earcons, which associate fixed sounds with specific meanings, morphocons create a dynamic sonic grammar by modifying parameters such as rhythm and frequency. This approach allows for adaptable sounds that convey abstract information, making them accessible to both blind and sighted users [10].

For Presti et al. [11], musification extends beyond basic sonification by incorporating tonality, modal scales, and higher-level musical features (i.e., polyphony, tonal modulation). Other techniques for sonification include integrating both speech and non-speech cues, these systems can convey layered information through metaphors of shared human experiences, reducing cognitive load and enhancing accessibility in complex environments [12].

In sound-assisted navigation, auditory dimension mappings make spatial data interpretable. For example, properties like distance are mapped through repetitive increasing or decreasing sound frequencies—continuous values—while others, such as position and object width, are mapped to more straightforward binary categories—discrete values—to avoid information overload [13].

Research emphasizes the importance of auditory aesthetics when designing auditory icons and earcons for users, noting preferences for lower frequencies and reduced loudness to avoid user discomfort, particularly with high frequencies that many find unpleasant [14]. In sound-guided navigation experiments [13], the authors demonstrated that users' emotional responses to sounds significantly impact task performance and engagement, highlighting the importance of sonification designs that balance clarity, confidence, and pleasantness [15].

Despite significant advances in auditory display and sonification technologies, several limitations remain. Most existing methods rely on symbolic or abstract earcons, thus requiring users to learn associations between sounds and events, potentially increasing cognitive load and limiting usability. Furthermore, while auditory icons can more intuitively represent events, they can be confused with environmental sounds, background noise, or a lack of universally intuitive mappings. Current sonification approaches also struggle to balance the amount of information conveyed without overwhelming users, often requiring careful calibration of auditory dimensions such as pitch, volume, and spatialization to avoid user fatigue or sensory overload. High-frequency sounds and specific auditory mappings can be unpleasant or counterproductive for users, limiting the overall user experience and accessibility, particularly for individuals with sensory sensitivities.

3. SONIFICATION METHOD

Based on principles of spatial cognition and sound perception, this section presents a sonification method to assist people from the BVI in navigating outdoor and indoor environments. It aims to transform critical spatial and semantic information about obstacles into non-speech sounds, enabling users to perceive their surroundings through sounds.

In detail, the method tackles four primary challenges faced by BVI individuals. First, it recognizes the position and distance of objects in the surroundings at any height. Second, distinguish between obstacles (i.e., trash bins, chairs, doors, poles, etc.) and objects delimiting areas of navigation (i.e., walls, curbs, stairs, etc.). Third, understanding spatial layouts by recognizing specific objects to navigate unfamiliar environments, such as an outdoor street with a conventional road and a small square or an indoor office workplace. Fourth, preserving auditory surrounding awareness while enhancing spatial and semantic understanding of the surroundings.

These challenges are addressed by the sonification of *semantic* and *spatial* properties of surrounding objects. Semantic properties characterize the nature and function of objects within the environment and the user. The term *obstacles* refers to objects that can be navigated around, such as poles or benches. The term *delimiters* describes objects that define boundaries, such as walls or curbs, indicating the end of a path. *Width* indicates an object's width or size. And *interaction* refers to an object's affordance level for BVI users, emphasizing those objects that users may find relevant and wish to interact with (i.e., doors, chairs, traffic lights, etc.). The mappings of semantic attributes to sound are detailed in Section 3.1.

Spatial properties encompass attributes related to the physical space and positioning of objects. These include 1) the *distance* to an object from the user location, 2) the *position* to the object's lateral placement relative to the user's path, and 3) the object's *elevation* to the ground level (i.e., the elevation of steps, curbs, or stairs). The mappings of spatial attributes to sound are detailed in Section 3.2.

The selection of objects for indoor and outdoor environments is shown in Table 1 and was informed by frequent appearances in extensive reviews of photographs of the streets and offices. In addition, some references were drawn from previous studies with similar experimental setups [1].

List of Objects for Experiments		
Benches	Bin trashes	Bollards
Bushes	Chairs	Columns
Constructions	Crosswalk	Curbs
Doors	Elevators	Fences
Fountains	Mailboxes	Potholes
Ramps	Roads	Single-steps
Stairs	Statues	Street lights
Tables	Traffic light	Trees
Vases	Walls	Windows

Table 1. List of objects adopted in our experimental sonification method.

3.1 Semantic properties

The sonification of semantic properties of objects included the following four primary elements: obstacles, delimiters, width, and interaction. Obstacles identification is represented by the timbre and amplitude envelope of their sonic material characteristics—such as brick, glass, metal, and wood, the most common construction materials in cities and interior spaces. Delimiters are represented by a simple sinewave within the 100-140 Hz range to contrast with the complex sounds of obstacles. These uncircumventable boundaries, like walls or construction barriers, define indoor and outdoor layouts.

The width of the obstacle is mapped to the pitch of its sound. We adopted discrete pitch values instead of a continuous scale to simplify the implementation of the method. We defined three categories of width: narrow, medium, and wide. The medium width serves as the baseline, with pitches adjusted one octave higher for narrow obstacles (multiplying the baseline frequency by 2) and one octave lower for wide obstacles (dividing the baseline frequency by 2).

Interaction is indicated by gain modulation and categorized into three levels: high, low, and none. Objects with high interaction (i.e., doors, traffic lights, crosswalks) are beneficial, while those with low interaction (i.e., benches, trash cans, mailboxes) have moderate utility. Objects without interaction (i.e., walls, trees, flowerpots) offer no practical affordance during navigation. Interaction is conveyed through discrete modulation of sound gain, implemented using additive synthesis with sinewave oscillators. The sound undergoes a modulation rate of 5 Hz for high-interaction objects and, for low-interaction objects, a modulation rate of 1 Hz. This modulation is applied through audio processing software, where the gain is modulated by altering the amplitude of the base sound.

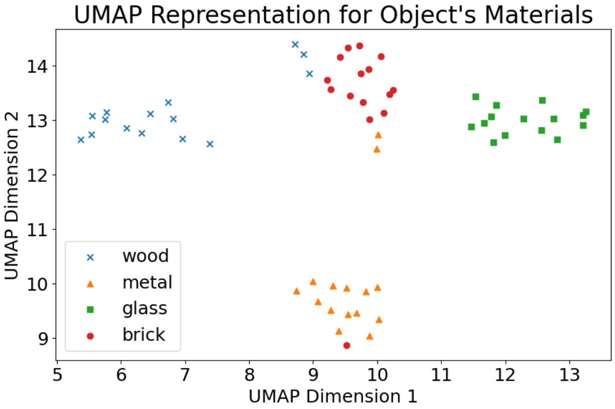

Figure 1. UMAP visualization of the content-based audio description space of obstacles' materials for the semantic labels wood, metal, glass, and brick.

We sonify objects in terms of timbre and amplitude envelope following a methodology proposed in [16] to identify acoustic attributes that best differentiate semantic labels of non-auditory cues, aiming to adopt the most intuitive sounds for a given material. Departing from semantic queries of the four most common object's materials, wood, metal, glass, and brick, a collection of 15 sound samples per semantic label retrieved from the crow-sourced platform Freesound. [1] A mathematical space of acoustic content-based descriptors, such as spectral brightness, roughness, and MFCCs, covers temporal, spectral, rhythmic, and tonal descriptions. From the resulting (multidimensional) space of about 50 features, a two-dimensional uniform manifold approximation and projection (UMAP) is created, as shown in Figure 1. After removing the outliers and attentively listening to the sounds at the center of each material's clusters, we selected representative sounds.

3.2 Spatial properties

The work of Presti et al. [11] inspires the spatial mappings adopted, who establish correlations between object properties and specific sound dimensions, organized to be meaningfully interpreted by the auditory sense [18].

The sonification of spatial properties of objects features the following three elements: distance, position, and elevation. Distance from the object is mapped to the gain, with closer objects having higher gain values and further objects having lower gain values. The distance range considered is from 4 meters (-20 dB) to 40 cm (0 dB). The gain adjustment follows a logarithmic relationship, using continuous values between these two points, with the sound intensity gradually increasing as the object approaches the user.

The position provides an indicator of the laterality of the object from the user and is mapped to the sound panning. When an object is directly ahead, the sound is equally panned to the left and right channels. As the object moves to the left or right, the panning adjusts accordingly, making the sound more pronounced in the corresponding channel.

[1] `https://freesound.org/` (accessed on November 13, 2024) is a repository of sounds where users upload labeled sounds by their semantic attributes. The platform allows for textual queries and outputs a list of audio samples ranked by the proximity to the query. Furthermore, each sample is annotated with features based on acoustical content computed by the Essentia library [17].

The detection range is based on the average shoulder width of a person (approximately 40 cm) plus an additional 20 cm, resulting in a detection field of about 60 cm. If no object is detected, no sound is played. The panning effect is achieved by linear interpolation between the left and right channels.

Elevation indicates changes in the ground level (i.e., steps, stairs, or ramps) through melodic tone sequences. Two tones are played consecutively for a single step or a ramp: the first tone represents the lower octave, and the second represents the higher octave. When the elevation increases—stepping from the road to the pavement—the sequence ascends, and vice versa when descending. For stairs, a sequence of four tones progresses by one octave each. All tones last one-quarter note (at ≈70 BPM), except for the last tone, which is held slightly longer (whole-note) to help users identify the start and end of the sequence and other modified properties.

One of the main intentions for these properties is to mimic the physical behavior of any sound source, which has been the study of psychoacoustics and simulated in multiple virtual environments to convey more realistic sound scenes. [2]

4. EVALUATION

The evaluation of our method is twofold. 1) A perceptual listening experiment assessing the intuitive association between the *semantic* properties of objects to the sound mapping designed and detailed in Section 3.1. 2) An in-person navigation experiment to assess the joint *semantic* and *spatial* navigation abilities in both blind and sighted participants within a 3D virtual reality (VR) environment. Additionally, 3) the data collection process for both experiments and 4) the methods used for data analysis are described.

4.1 Perceptual Listening Test

A listening test was conducted to validate the mappings developed in the sonification method by examining how well participants can *intuitively* interpret the semantic information conveyed, namely the *obstacles'* material identification, as well as their *width* and degrees of *interaction*.

4.1.1 Test Structure

The listening test was implemented using LimeSurvey [3] and comprised three groups of questions, each with four questions, totaling 12 questions per participant. The test was preceded by a demographic section that collected information on age, nationality, vision type, and musical knowledge. Each group evaluated a specific relationship between sound and semantic properties, with responses recorded in a confusion matrix format [19]. The question groups are as follows:

- *Obstacles material*-timbre: each question presented one audio clip with five possible answer options: A) Brick, B) Glass, C) Metal, D) Wood, or E) None of the above.
- *Width*-pitch: each of the four questions was a question structured like a matrix response and included three audio clips (same material, varying pitch), with three answer options for each sound: A) Narrow, B) Medium, or C) Wide.
- *Interaction*-modulation: each of the four questions was a question structured like a matrix response and included three audio clips (same material, varying modulation levels), with three response options for each sound: A) No interaction, B) Medium interaction, or C) High interaction.

Participants were instructed to rely solely on judgment, with no prior knowledge of the mappings, as there were no correct answers. Questions within each group were randomized to minimize order bias. Only semantic properties were tested, as spatial properties are more useful for locating objects than recognizing them. Additionally, adding more questions could increase complexity and cause participants to discontinue the test.

4.1.2 Participants

The listening test was primarily distributed through mailing lists within the University of Porto, online academic communities, via the Association for Visually Impaired Individuals in Portugal [4] , and international research communities focused on music and sound [5] . In total, 50 participants completed the entire survey, resulting in a total of 600 trials conducted. Participants ranged in age, with a slightly higher concentration in the 45–54 age group. Fourteen reported being "blind or low vision," while the remaining 36 reported being "sighted". Regarding musical knowledge, 26 participants claimed to have either "Strong or basic knowledge", while 29 stated having "No experience." Most participants were from "Spain" (23) and "Portugal" (19).

4.2 In-person Navigation in Virtual Reality

This experiment evaluated spatial navigation in blind and sighted participants within a 3D VR environment, providing users with both semantic and spatial object properties. Participants were tasked with navigating two simulated scenarios. Both semantic and spatial properties were assessed, thus adopting the totality of the proposed sonification method detailed in Section 3.

4.2.1 Experimental Set Up

We have adopted Unity [6] to create a VR simulation of the two indoor and outdoor scenarios under evaluation, where common tasks can be tested, such as crossing a street or walking through an office. The virtual environment was

[2] The repository containing the tested sounds is available on GitHub (https://github.com/dngea/Semantic-and-Spatial-Sound-Object-Recognition-for-Assistive-Navigation).

[3] `https://www.limesurvey.org/` (Last accessed April 23, 2024).

[4] ACAPO `https://www.acapo.pt/` (Last accessed November 24, 2024).

[5] NIME `https://www.nime.org/`, ISMIR `https://ismir.net/` (Last accessed April 20, 2024).

[6] Unity Technologies, `https://unity.com` (Last accessed November 24, 2024).

chosen to avoid the complexity of integrating sonification with AI-based object recognition, focusing solely on sonifying objects rather than developing detection algorithms.

The experimental framework builds on prior research, which utilized audio cues in a virtual maze [20], and investigations into mental mapping for spatial navigation [21]. Participants used an external controller—specifically, a smartphone equipped with a gyroscope—to simulate forward movement and lateral rotation. Forward movement was activated by pressing a button, with a footstep audio cue indicating a walking speed of 1.20 m/s. As participants navigated, objects emit sounds through headphones, requiring them to adjust their paths based on auditory cues, while the absence of sound indicated a clear trajectory. Visual input was entirely omitted to simulate a blind navigation experience. An examiner supervised the experiment, monitoring participants' locations and interactions within the virtual environment.

4.2.2 *Experiment Dynamics and Tasks*

Participants received a briefing on the study's objectives and procedures, followed by the experiment, with a total duration of approximately 45 minutes. During this training session, they were introduced to the objects listed in Table 1 and their corresponding sounds. They familiarised themselves with the movement controls, identical to those used in the experiment maps. The experiment was structured as follows: it began with the outdoor map, where participants were given a raised relief map—which represented the layout of the 3D scenario built in Unity, as shown in Figure 2—to help them navigate. They were informed that the objects in this space were the same as those encountered during training. Participants were instructed to move from point A (marked by a raised dot) to location B, which remained consistent across all experiments and represented a reasonable endpoint within the map, offering a trajectory with moderate challenges and objects. Employing a think-aloud methodology, participants were required to identify the semantic and spatial properties of each sound they heard before attempting to guess the corresponding object. After approximately 5 minutes of navigation, they were asked to position themselves on the relief map or inform the examiner when they believed they had reached location B. This same methodology was applied during the indoor map navigation. Therefore, the experiment consisted of three primary tasks:

1. **Recognition of Sound Cues**: Participants identify sound cues, associating them with semantic (obstacles, delimiters, width, interaction) and spatial (distance, position, elevation) properties.

2. **Object Recognition**: Participants apply deductive reasoning in a think-aloud format to identify objects based on sound cues. This process encourages participants to articulate their thought processes, aiding in data analysis.

3. **Self-Location**: Participants identify landmarks and their positions using contextual information, the raised relief map, and their starting point. They are instructed to navigate from point A to location B, utilizing auditory cues to aid their journey.

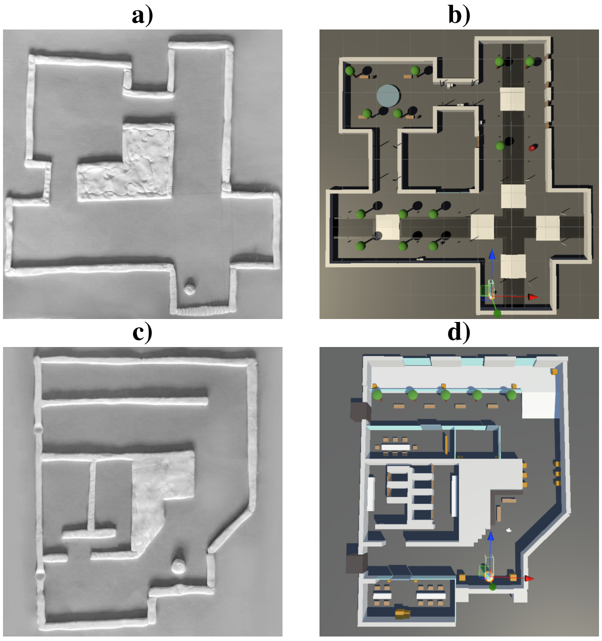

Figure 2. Comparison of raised relief and 3D map visualizations for indoor and outdoor settings. Each sub-figure is labeled with a) Raised relief outdoor, b) Outdoor 3D map, c) Raised relief indoor, and d) Indoor 3D map.

4.2.3 *Participants*

Invitations to participate in the study were disseminated through word-of-mouth and mailing lists, resulting in the recruitment of 10 participants: five BVI participants sourced from a mailing list of ACAPO, in the city of Porto, and five sighted participants recruited via the University of Porto mailing list. Table 2 summarizes the participants' demographic details. Five of the 10 participants (average 23.5 years old) were females. Breaking down the visual conditions, three participants were classified as blind, two had low vision, and five were sighted. Most participants (eight) were students, and two were employed. Regarding musical proficiency, four had no experience, four had basic knowledge, and two had a solid musical background. All participants were familiar with current digital technologies. Given the experiment's VR context, video game experience was also recorded, as it could affect navigation abilities. Five participants reported no video game experience, while the other five had some or considerable experience.

4.3 Data Collection

The perceptual listening test used a quantitative data collection process with predefined questions to ensure data validity. Administered via LimeSurvey, the survey included three sections examining the relationships between semantic properties and sounds: obstacle material-timbre, width-pitch, and interaction-modulation. Participants answered 12 questions each, resulting in 600 trials (12 questions ×

ID	Sex	Age	Vision	Music Exp.	Game Exp.
P1	F	20	Low	No	No
P2	M	21	Low	Low	Low
P3	F	22	Blind	Low	No
P4	F	24	Sighted	Low	High
P5	M	27	Sighted	No	No
P6	M	25	Blind	High	High
P7	F	21	Blind	Low	No
P8	M	23	Sighted	High	High
P9	F	24	Sighted	High	No
P10	M	27	Sighted	High	Low

Table 2. Participants' demographic information

50 participants). Responses were recorded in a confusion matrix format.

In the in-person navigation experiment, 10 participants navigated two maps with 27 objects and 42 unique sound combinations. Data collection included both quantitative and qualitative components. Quantitative data focused on object recognition, self-positioning, and the identification of sound properties, documented in a confusion matrix. A total of 270 trials (27 sound combinations × 10 participants) were conducted.

Qualitative data was gathered through think-aloud methodology, with participants providing feedback via open-ended questions and completing the System Usability Scale (SUS) questionnaire at the session's end.

4.4 Data Analysis

In both experiments, we report the results in terms of accuracy, computed as the ratio of correctly categorized stimuli to the total number of tests for a given category. In detail, for the perceptual listening test, accuracy measures the number of correct associations between the sound properties. For the in-person VR experiment, it measures the number of correctly identified obstacles' semantics and their spatial properties. Results are reported for the total number of participants per experiment and sub-groups of visual condition, musical knowledge, age, and nationality. Descriptive statistics, including the average and standard deviation, measuring the dispersion of the values from the mean, are reported per group. To infer statistical differences between groups, we adopted the non-parametric Mann-Whitney U test.

5. RESULTS

5.1 Recognition of Sound and Semantic Properties Mappings

The global accuracy for the semantic object recognition tasks from the online listening test was .69 ± .13. When examining the individual mappings, we obtain an accuracy of .66 ± .42 for *obstacle material*-timbre, .64 ± .48 for *width*-pitch, and .75 ± .43 for *interaction*-modulation. Significant variability in participant responses is observed while presenting relatively high average accuracy, suggesting differing levels of familiarity and understanding of the sound associations.

Metal and glass were identified as the most recognizable materials within the *obstacle material*-timbre associations, with .95 ± .06 and .72 ± .13 average accuracy, respectively. Wood and brick proved more challenging to distinguish, with an accuracy of .49 ± .15. In Figure 3 b), a confusion matrix illustrates a tendency among participants to mismatch wood and brick while consistently identifying metal and glass.

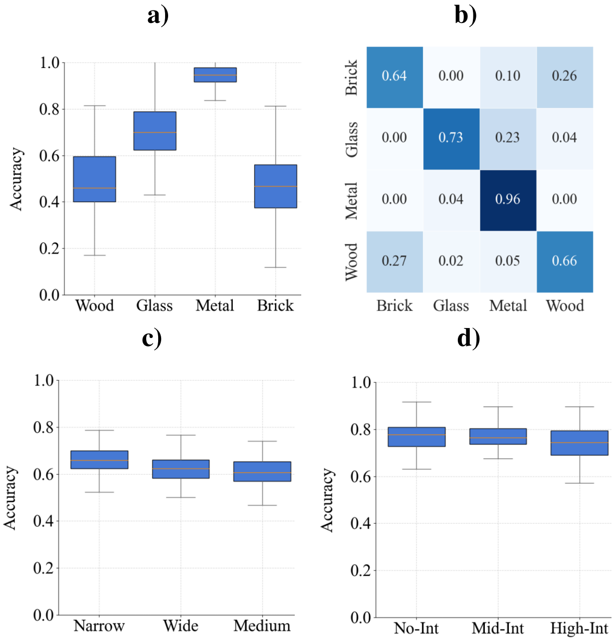

Figure 3. Accuracy of obstacles' semantic properties. a) Accuracy of obstacle materials, b) confusion matrix for obstacle material, c) accuracy for width-pitch, and d) accuracy for interaction-modulation.

Similar statistics resulted for *width*-pitch associations, with high-pitch narrow objects achieving an accuracy of .67 ± .07. Individual contributions showed that wood and brick performed notably well, scoring .64 ± .1 and .62 ± .1, respectively. However, glass lagged with a score of .53 ± .07, suggesting that the acoustic properties of materials significantly influence participants' ability to make accurate associations. This finding contrasts with the *obstacle material*-timbre results, revealing that different auditory dimensions interact in complex ways.

For the *interaction*-modulation dimension was particularly recognizable for high-interaction sounds with .75 ± .07 accuracy. These results suggest a strong link between modulation and interaction associations, indicating that participants were better equipped in this auditory category.

5.2 Navigation Performance

The in-person navigation experiment evaluated the participants' performance in recognizing sound cues, identifying objects, and self-locating within the virtual environment. Participants achieved an average accuracy of .73 ± .16 across all tasks, with sound recognition output the highest accuracy at .88 ± .14, reflecting the effectiveness of the sonification technique. Object recognition followed with a score of .72 ± .13, while the most challenging task—self-location—scored an accuracy of .60 ± .2.

In the sound cues recognition task, timbre attained an overall accuracy of .88 ± .12, with pitch and modulation scoring .9 ± .1 and .93 ± .05, respectively. Volume and panning also received high ratings of .94±.05 and .94±.1. In contrast, tone intervals scored lower at .77 ± .2, indicating high variability in participants' responses. The material recognition was improved in comparison with the perceptual experiment, with accuracy results of .86±.1 for wood, .85.11 for metal, .80±.17 for glass, .93±.12 for brick, and .94 ± .1 for sinewaves. This improvement is noteworthy, particularly following training sessions before the experiment, which reduced accuracy dispersion across materials.

The object recognition task showed participants encountered an average of ten objects per map, with an overall accuracy of .72 ± .13, correctly identifying about seven out of ten objects. Recognition times decreased significantly with repeated encounters, from 10.5 seconds on the first encounter to 6.0 seconds by the sixth. Qualitative feedback indicated increased confidence as participants became familiar with the objects, although some felt overwhelmed by multiple stimuli.

The self-location task was challenging, with participants averaging 2.4±.6 instances of disorientation before examiner intervention, improving to 1.4 ± 1.28 afterward. The average experiment duration was about 7 minutes, with more time spent on the indoor map due to its size and complexity. Self-location accuracy improved from .44 ± .20 before intervention to .62 ± .27 afterward. The heatmap analysis in Figure 4 highlights areas of prolonged engagement, emphasizing the need to gather information before making navigational decisions.

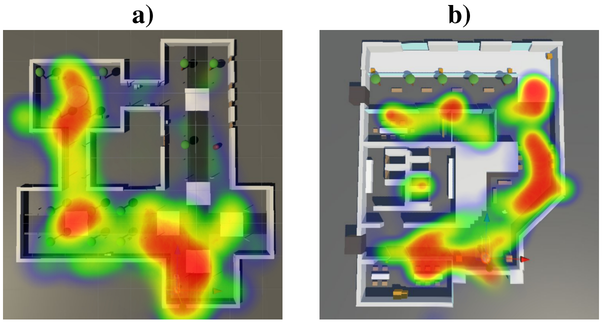

Figure 4. Comparison of a) outdoor and b) indoor hotspots of average stop time.

5.3 Comparison of Demographic Factors

Demographic comparisons provided valuable insights into the influence of musical background and vision condition on accuracy. Participants with some musical training achieved a higher overall accuracy (.79 ± .14) than those without musical knowledge (.63 ± .24). Significant statistical differences (p-value of .029) were found, suggesting that musical knowledge is critical to improving sound recognition abilities. Although participants with musical training demonstrated higher accuracy in *obstacle material*-timbre and *width*-pitch associations, these differences did not reach statistical significance, indicating the need for further exploration.

Comparative analysis revealed that sighted participants had slightly higher accuracy (.49±.34) than blind and low-vision participants (.45±.38), though no significant statistical differences were found (p = .86). This may be due to the limited representation of blind and low-vision participants, who comprised 22% of the sample. Age also played a role, with younger participants (ages 18-24) showing lower accuracy (.32 ± .70) and greater variability, while older participants had more consistent accuracy (.56±.64). This highlights the impact of age on auditory perception and sound association.

6. CONCLUSIONS AND FUTURE WORK

We proposed a sonification method to assist BVI individuals in navigating two (outdoor and indoor) environments by mapping spatial and semantic information about objects into non-speech sounds.

Our evaluation of the proposed mappings in an online listening test showed an overall accuracy of .69 ± .13 in recognizing semantic attributes of objects, indicating that participants could make meaningful associations between sounds and object properties. The analysis highlighted variability in performance across different sound dimensions, underscoring sound perception's inherent complexity and subjectivity, emphasizing the need for refinement in sound design to improve recognition rates.

An in-person navigation experiment in VR further confirmed the effectiveness of auditory cues in aiding navigation, achieving an average accuracy of .73 ± .16 across three primary tasks: recognition of sound cues, identification of objects, and self-location within a virtual environment. The strong correlation between sound cue recognition and repeated exposure improved accuracy in object recognition, with an overall accuracy of .72 ± .13, emphasizing the importance of a robust base of intuitive sounds to reduce the learning curve in real-world applications. Self-location tasks proved challenging, highlighting the added difficulty simulated 3D spaces pose for both BVI and sighted individuals. However, after the examiner intervention, participants showed notable improvement in spatial perception, overcoming the initial disorientation common in such complex environments.

Qualitative feedback from participants regarding the prototype was positive, indicating satisfaction with the clarity and logical structure of the sound mappings. This was supported by a System Usability Scale (SUS) score of 75, exceeding the industry average of 68 and indicating strong usability. However, the feedback also pointed to potential areas for improvement, particularly concerning user experience in navigating complex environments. This emphasizes the need for continued system refinement to enhance usability and further accommodate the diverse needs of participants.

Overall, the findings from this research highlight the potential of sonification techniques as valuable aids for BVI individuals, enhancing spatial awareness and navigation through auditory cues. By integrating sound perception with practical navigation tasks, we gain insight into how intuitive sounds can support learning and navigation. Future research should involve a more diverse participant pool and investigate the effects of sound training on recognition abilities, leading to better auditory interfaces for improved accessibility.

Future work will focus on refining the sonification method by adjusting sound categories and testing them in sound perceptual experiments to reduce subjectivity. Expanding these experiments across various channels will help gather more comprehensive data. Additionally, integrating sonification with computer vision can improve real-world navigation systems.

While current research emphasizes vision, our experiments suggest hearing plays a crucial role. Exploring nonverbal sound-based languages could prove essential, embracing the abstract nature of sound as a strength. Ultimately, our work aims to demonstrate that the sum of these efforts creates more effective systems for the BVI community.

7. REFERENCES

[1] E. Nuhn, K. Hamburger, and S. Timpf, "Urban sound mapping for wayfinding – a theoretical approach and an empirical study," *AGILE: Giscience Series*, vol. 4, pp. 1–13, 2023.

[2] M. Hersh, "Wearable travel aids for blind and partially sighted people: A review with a focus on design issues," *Sensors*, vol. 22, no. 14, p. 5454, 2022.

[3] G. Kramer, *Auditory Display: Sonification, Audification, And Auditory Interfaces*, 1994.

[4] D. Gomez, J. Bologna, and T. Pun, "See color: an extended sensory substitution device for the visually impaired," *Journal of Assistive Technologies*, vol. 8, no. 2, pp. 77–94, 2014.

[5] D. K. Mcgookin and S. A. Brewster, "Understanding concurrent earcons," *ACM Transactions on Applied Perception*, vol. 1, no. 2, pp. 130–155, 2004.

[6] B. N. Walker and G. Kramer, "Auditory displays, alarms, and auditory interfaces," in *International Encyclopedia of Ergonomics and Human Factors*, 2006, pp. 1021–1025.

[7] B. N. Walker and M. A. Nees, "Theory of sonification," in *The sonification handbook*, 2011, vol. 1, pp. 9–39.

[8] B. N. Walker and L. M. Mauney, "Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 2, no. 3, pp. 1–16, 2010.

[9] I. I. Bukvic, G. D. Earle, D. Sardana, and W. Joo, "Studies in spatial aural perception: establishing foundations for immersive sonification," *Unpublished work*, 2019.

[10] F. Morando, A. Lepetit, J. Espinosa, and S. Trentin, "Morphocons: A new sonification concept based on morphological earcons," *Proceedings of the 13th International Conference on Auditory Display (ICAD 2006)*, pp. 307–312, 2006.

[11] G. Presti, D. Ahmetovic, M. Ducci, C. Bernareggi, L. Ludovico, A. Baratè, F. Avanzini, and S. Mascetti, "Watchout: Obstacle sonification for people with visual impairment or blindness," New York, NY, USA, p. 402–413, 2019. [Online]. Available: https://doi.org/10.1145/3308561.3353779

[12] T. L. Smith and E. B. Moore, "Storytelling to sensemaking: A systematic framework for designing auditory description display for interactives," New York, NY, USA, p. 1–12, 2020. [Online]. Available: https://doi.org/10.1145/3313831.3376460

[13] T. Senan, B. Hengeveld, and B. Eggen, "Sounding obstacles for social distance sonification," in *Proceedings of the 17th International Audio Mostly Conference*, September 2022, pp. 187–194.

[14] K. Kurakata, T. Mizunami, and K. Matsushita, "Sensory unpleasantness of high-frequency sounds," *Acoustical Science and Technology*, vol. 34, no. 1, pp. 26–33, 2013.

[15] A. Sharif, O. Wang, and A. Muongchan, "What makes sonification user-friendly? exploring usability and user-friendliness of sonified responses," pp. 1–5, 10 2022.

[16] Z. Cao, A. Pinto, and S. Bernardes, "Bisaid: Bipolar semantic adjectives icons and earcons dataset," in *Proceedings of the Sound and Music Computing Conference*, 2024.

[17] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, and X. Serra, "Essentia: An audio analysis library for music information retrieval," in *International Society for Music Information Retrieval Conference*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:11200511

[18] G. Eshetu, *Factors affecting instructional leaders perception towards educational media utilization in classroom teaching*. Anchor Academic Publishing, 2015.

[19] K. W. Ma, H. M. Wong, and C. M. Mak, "A systematic review of human perceptual dimensions of sound: Meta-analysis of semantic differential method applications to indoor and outdoor sounds," *Building and Environment*, vol. 133, pp. 123–150, 2018.

[20] E. Gandolfi and R. Clements, "Alternative embodied cognitions at play evaluation of audio-based navigation in virtual settings via interactive sounds," *Journal For Virtual Worlds Research*, vol. 12, no. 1, 2019.

[21] A. Afonso-Jaco and B. F. G. Katz, "Spatial knowledge via auditory information for blind individuals: Spatial cognition studies and the use of audio-vr," *Sensors*, vol. 22, no. 13, p. 4794, 2022.