World Navigation Hat – Visual to Audio and World Mapping

Jason D'Souza*
University of San Agustin
jdsouza@usa.edu.ph

Ethel Herna Pabito
University of San Agustin

ChenLin Wang
University of San Agustin

Vince Ginno Daywan
University of San Agustine

Abstract

Visually impaired individuals face significant challenges in spatial navigation. Traditional aids (e.g. white cane, guide dogs) provide limited environmental information, and electronic travel aids can overwhelm users with raw data. We propose the *World Navigation Hat*, a wearable system that uses cameras, inertial sensors, and microphones to build a real-time 3D representation of the environment and render it as audio. A depth sensor (e.g. stereo camera or RGB-D unit) captures spatial structure, and an onboard processor (Raspberry Pi) converts this into a continuous spectrogram-like "soundscape." In this audio display, obstacle distance is encoded by loudness and pitch, and spatial layout by stereo panning, all calibrated by human hearing models (e.g. equal-loudness curves) to optimize clarity. Psychoacoustic adjustments (equalization, perceptual tuning) are applied so that only critical cues are conveyed, minimizing cognitive load. The Hat's modular Linux-based OS enables additional features (hand-gesture input, voice-assistant integration, IoT connectivity) and easy reconfiguration. We will develop and iterate the Hat through expert feedback and low-fidelity prototyping, followed by qualitative user trials (with visually impaired volunteers) to evaluate usability, learnability, and navigation performance. Insights from expert review and user interviews (analyzed via thematic coding) will guide design refinements.

Table of Contents

1. Introduction

Despite policy and infrastructure efforts, the daily mobility of visually impaired people remains restricted. For example, guide dogs are effective but entail high cost and long training times. Simple tools like the white cane give local touch feedback, but they require active probing and can miss overhead or distant obstacles. Electronic aids (e.g. sonar canes, GPS apps) can supplement these methods, yet they often provide sparse or non-spatial cues. Recent advances in wearable computing offer new possibilities: on-board cameras and sensors can continuously scan the surroundings, and processors can analyze scenes in real time. In particular, sensory substitution – conveying visual information via sound – has shown promise for non-invasive assistance

Visual-to-auditory sensory substitution devices (SSDs) acquire visual data (from a camera or depth sensor) and render it through the auditory channel. Classic examples like The vOICe transform images to soundscapes, enabling blind users to “hear” shapes and spatial layouts. Neuroscientific studies find that blind users of SSDs can develop visual-like perceptions from these sounds, leveraging cross-modal brain plasticity. However, mapping a complex 3D world into audio is inherently challenging: human vision conveys far more detail than hearing can naturally represent. Designers therefore emphasize transmitting only critical information (to avoid overload) and choosing audio parameters that remain distinct in perception

The World Navigation Hat builds on these ideas. It uses a depth camera and inertial sensors to create a live spatial map. This map is sonified: for example, closer obstacles trigger higher intensity and pitch, while lateral position is mapped to stereo pan. Crucially, we adjust these mappings using psychoacoustic models. For instance, we apply equal-loudness weighting to ensure all frequencies are perceived at comparable loudness, preventing some cues from dominating due to human ear sensitivity. We also tune the audio so that mapped dimensions (e.g. pitch vs. volume) are perceptually independent. By grounding the design in human hearing characteristics, our goal is an intuitive soundscape requiring minimal training. The system is implemented on a Raspberry Pi with a Linux-based, modular OS. This enables flexible extension:

we can add hand-gesture input or integration with cloud services without redesigning the core. For example, voice-controlled assistants (e.g. Alexa, Google Assistant) have proven highly useful for blind users, and our Hat's platform can interface with these. Likewise, connecting to IoT services (as in systems like VISISENSE) could allow offloading heavy vision processing to the cloud or sharing data with smart infrastructure.

In summary, this project aims to create a head-mounted navigation aid that (a) translates depth and spatial layout into an auditory display using human-centered psychoacoustic mapping, and (b) provides a modular, augmented platform for control and connectivity. The following sections detail these two philosophical objectives, survey related work, and outline the planned evaluation methodology.

a. Sensory Substitution

The first core principle is sensory substitution: conveying visual-spatial information through sound. In our Hat, live video/depth data is processed into audio in real time. Research in neuroscience indicates this can be effective: when blind users wear an auditory SSD for months, they often report detailed "visual" sensations triggered by sound. This is explained by the brain's multimodal organization: in the absence of vision, occipital (visual) cortex can be repurposed to interpret auditory or tactile input for spatial cognition. Importantly, this plasticity is not limited to early blindness – even adults can learn new cross-modal mappings given sufficient training. Thus, we design the Hat's audio output to be as intuitive as possible to speed learning.

Our audio encoding strategy is inspired by proven SSD designs. For example, "Stoll et al.'s MeloSee" device mapped a 2D depth image into a "melody" in which object distance controlled loudness, pitch encoded vertical position, and stereo panning indicated left-right position. Similarly, our system uses a spectrogram-like scheme: each image column (at a given angle) becomes a time-slice of audio, where higher rows (closer objects) are rendered with higher frequencies or greater volume. Unlike naive sonification, we incorporate human psychoacoustics. For instance, audio output is calibrated by equal-loudness contours so that sounds at different pitches are perceived equally loud when they should be. We also ensure orthogonality: changing obstacle distance alters loudness without inadvertently shifting perceived pitch, and vice versa. These precautions follow recommendations from auditory display research, which shows that arbitrary mappings often confuse users unless perceptual interactions are managed.

To avoid cognitive overload, we filter and prioritize inputs. Only key obstacles or features within a certain field of view are sonified, consistent with expert SSD design rules. This "focal" sonification approach prevents the auditory channel from being flooded with information. In practice, the hat's camera depth map might be downsampled or segmented so that only the nearest objects produce sounds. The user should hear a stable, continuous audio scene that highlights immediate hazards while suppressing distant or irrelevant details. By focusing on salient spatial cues and tuning the audio to human hearing, we aim to make the navigation sounds learnable and actionable with minimal training.

b. Argumented Modular Operating System

The second principle is a flexible software platform: an augmented modular operating system. We implement this on a Raspberry Pi single-board computer. The Pi is chosen for its accessibility, low cost, and community support. Its Linux OS (e.g. Raspbian) allows rapid development of modular components (camera drivers, sensor fusion, audio synthesis, etc.) that can be updated independently. This modularity means the Hat can be extended: for example, a computer vision module (e.g. object recognition via TensorFlow) could be added as a separate process without altering the core navigation logic.

The modular OS also enables advanced interaction modes. We plan to incorporate hand-gesture controls: for instance, a palm sensor or Leap Motion unit can be attached to the hat, allowing the user to issue simple gestures (swipe, tap) to change modes or ask for information. Such alternatives improve usability: voice or gesture commands are intuitive and hands-free. They have proven value; for example, a Raspberry Pi–based wheelchair used voice and hand gestures for control, making it accessible to users with limited mobility. Likewise, we will interface the Hat with voice assistants. Contemporary smart speakers (Amazon Alexa, Google Home) are extremely useful to blind users, as they provide spoken information on demand. By connecting our device to these platforms (via internet), a user could query the environment ("What's in front of me?") or request external data (directions, weather) through the hat's audio output.

Networking (Wi-Fi, Bluetooth) opens further IoT possibilities. For example, we might pair the hat with smartphones or cloud services for heavy computation: depth data could be streamed to a remote server for scene understanding. This concept is supported by systems like VISISENSE, an IoT-based assistive navigation aid that achieved >99% object detection accuracy and low latency by offloading vision tasks to cloud AI. In future iterations, the Hat could similarly leverage external processing or share maps with other smart city infrastructure. The underlying Raspberry Pi OS handles networking and allows these modules to plug into the system easily.

In summary, our augmented OS goal is to make the World Navigation Hat a "platform" rather than a closed device. Sensors and audio synthesis form the core, while optional modules (gesture recognition, voice assistant integration, web connectivity) can be enabled. This approach aims to increase the device's longevity and adaptability; as new sensors or algorithms emerge, they can be incorporated into the Hat's stack with minimal redesign.

2. Related Work

Researchers have long explored wearable navigation aids. Early electronic travel aids (ETAs) typically used sonar or infrared to detect obstacles, providing simple auditory or tactile

feedback. More recent work uses cameras and computer vision. For instance, Stoll et al. (2015) demonstrated the MeloSee system: blindfolded users navigated hallways using a Kinect depth camera whose output was converted to sound. Participants learned to approach targets more quickly over sessions, showing that real-time depth sonification can guide movement. Ghaderi et al. (2015) created a "retina-inspired" wearable using a dynamic vision sensor (DVS) for low-latency detection of obstacles. Chang et al. (2021) built a head-mounted navigational aid combining RGB-D sensing and semantic SLAM. Their device understands the environment (e.g. recognizing doors, people) and gives voice feedback, demonstrating wearable SLAM feasibility. Like Chang's design, our Hat uses an RGB-D camera, but we focus on continuous audio display rather than discrete voice prompts.

Color-based sonification has also been studied. Singh et al.'s Colorophone (2021) is a head-mounted system that converts the color distribution in the camera's view into a spatialized soundscape. It uses a dedicated color-opponent space and stereo audio so that different hues and regions produce distinct sounds. Although it targets color recognition, Colorophone illustrates how continuous visual information can be rendered as an intuitive ambient audio stream. Our design is analogous, but emphasizes depth (distance) as the primary cue.

The design of auditory navigation aids must consider perceptual factors. Ziemer and Schultheis (2018) proposed a psychoacoustic navigation display and emphasized that physical sound parameters often interact in perception. For example, simply varying pitch and volume in orthogonal ways can confuse users, since louder sounds can also seem "brighter" in pitch. They argue (α) each audio dimension should be interpretable on its own, (β) parameters should be orthogonal in perception, and (γ) multiple cues must form a coherent single stream. We follow these guidelines by carefully choosing mappings (e.g. loudness for distance, filtering to isolate pitch changes) to meet such demands. Other studies highlight SSD design issues: only a few wearable devices provide full spatial audio feedback, and many overload the user. By limiting information to salient obstacles and using perceptual equalization, our Hat builds on this literature to create a more manageable audio display.

Finally, the Hat's platform builds on trends in assistive hardware. Microcomputers like the Raspberry Pi are increasingly used in prototypes because they support multimedia and I/O for low cost. Voice and gesture control have been integrated in devices from smart canes to wheelchairs with positive results. In the IoT realm, the VISISENSE study showed the power of linking wearable sensors to cloud AI for navigation. Our Hat adopts this practice by enabling network connectivity and standardized interfaces, aiming to create a next-generation assistive system that synthesizes the best of prior work.

3. Methodology

We will follow a user-centered, iterative design and evaluation process, integrating expert feedback and user testing in each cycle. The methodology comprises the following steps:

1. **Concept Refinement with Experts**: We will consult domain experts (orientation & mobility instructors, audio engineers, cognitive psychologists) through interviews and design workshops. Their input will refine requirements (e.g. which environmental features are most critical, safe operating constraints) and assess early prototype ideas.
2. **Prototyping**: Based on feedback, we will build successive prototypes of the Hat. Early prototypes may use off-the-shelf components (Raspberry Pi, USB depth camera, headphones) to demonstrate core functionality. Later versions will integrate custom PCBs or enclosures. Each prototype iteration will implement the depth-to-sound algorithm and the modular OS framework (supporting gestures and connectivity). Code will be developed in a modular way so components (audio engine, sensor drivers, gesture modules) can be swapped or tuned independently.
3. **Expert Evaluation**: After each prototype, we will conduct heuristic evaluations and cognitive walkthroughs with experts. For example, an occupational therapist or visually-impaired consultant will wear the device in a controlled setting and provide feedback on comfort, intuitiveness of controls, and perceived audio clarity. We will also perform technical tests (e.g. measuring audio latency, sensor range) to ensure specifications are met. Insights will directly inform refinements (e.g. adjusting audio filters, reordering sonification strategy).
4. **User Testing**: Once the system is functionally stable, we will recruit visually impaired participants for qualitative user studies. In lab-controlled scenarios (mock-up corridors or obstacle courses), users will perform navigation tasks wearing the Hat (blindfolded sighted participants may be included in early tests). We will observe performance (time to navigate, collisions) and conduct semi-structured interviews or think-aloud protocols to capture users' experiences. Questionnaires (such as the NASA-TLX for workload or SUS for usability) can be adapted for each trial. Each session will be video-recorded (with consent) and audio-recorded, and any user comments and behaviors noted.
5. **Data Analysis**: Qualitative data (interview transcripts, observational notes) will be analyzed using thematic analysis. We will code responses to identify common themes (e.g. "pitch perception difficulty", "occlusion of environmental sounds", "gesture control confusion"). This analysis will highlight usability issues and user preferences. Quantitative measures (navigation time, errors) will also be compared across prototype versions to assess improvements. The findings from analysis will guide the next design iteration, closing the loop.

Throughout development, design choices (e.g. audio mapping parameters, sensor placement) will be documented and justified. By involving both experts and end-users iteratively, we ensure the final system is not only technically sound but also practically usable by the visually impaired community.