

Making Objects Speak: Spatial Audio Guidance for Object Grasping by Blind and Visually Impaired Users*

Chenxin Qin, Yukiko Iwasaki, Chenyang Li, and Hiroyasu Iwata, *Member, IEEE*

Abstract—This paper presents an assistive system that enables blind and visually impaired (BVI) users to localize and reach objects using spatialized audio cues, rendered as if the objects themselves emit sound. By integrating voice command recognition, RGB-D-based 3D localization, and head-tracked spatial audio via Apple AirPods Pro, the system transforms object positions into egocentric, directional prompts aligned with the user’s head orientation. We evaluated the system through tabletop localization-to-contact tasks with blindfolded sighted participants, comparing a spatial-audio (SA) condition against a speech-only (SO) baseline. While success rates were comparable between conditions, spatial audio significantly reduced task completion time and subjective workload and received substantially higher usability ratings. These findings suggest that spatialized object-originating sound can enhance task efficiency and user experience in near-field, non-visual interaction scenarios.

I. INTRODUCTION

According to current global forecasts, the number of people who are blind is projected to increase from approximately 43 million in 2020 to about 61 million by 2050 [1]. This stark projection underscores the need for assistive technologies that let blind and visually impaired (BVI) individuals interact with everyday objects without relying on sight. Although traditional aids such as white canes and guide dogs enhance mobility, they offer no centimeter-level spatial information for locating and grasping specific items.

A comprehensive survey of assistive-robot projects for BVI users confirms that today’s systems still lean heavily on speech prompts, vibrotactile belts or other sensory-substitution channels; these modalities rarely provide real-time, egocentric, and hands-free guidance, and often overload users cognitively or physically [2].

Spatial audio has emerged as a promising alternative because it encodes “where” information directly in sound. Studies in dynamic settings show that head-tracked spatial cues can boost localization accuracy and reduce head-movement load for blind soccer players [3]. At the tabletop, making each game card “speak” in the Auditory Uta-Karuta system enables blind and sighted players to pinpoint targets instantly and enjoy the game on equal terms [4].

Yet many spatial-audio prototypes are still confined to meter-scale navigation. Asakura et al. [5] demonstrate a bone-

conduction headset that pairs voice prompts with a 1.2 kHz warning tone to improve outdoor pathfinding, but it offers no centimeter-grade cues for fine manipulation. Paré et al. [6] convert hallway obstacles into horizontally spatialized “water-droplet” sounds that early- and late-blind users master after just 30 minutes of training, yet the feedback remains at the obstacle level.

Closer to object interaction, the wearable StereoPilot headset fuses an RGB-D camera with head-tracked spatial audio and halves grasp time while cutting localization error by 40 percent; however, it presumes that users naturally orient their head toward the workspace and does not integrate semantic voice input [7]. Other approaches revert to tactile or mixed-modality feedback: a semantic-SLAM rig delivers directional cues via a vibrating wrist array—occupying the hands and lacking head correction [8]—and map applications that blend speech with vibration report cognitive overload in dense point-of-interest scenarios [9].

Although RGB-D cameras and YOLO detectors can now pinpoint objects with centimeter accuracy, most robotic pipelines internalize this information rather than mapping it onto the user’s egocentric frame. In our own earlier work, the HOVER-BVI system employed a large-language-model planner to interpret free-form voice commands and autonomously steer a robotic arm for hand-over; yet this full automation also deprived users of the opportunity to explore and verify an object’s direction themselves, leaving them passive and uncertain during the transfer [10].

Recently, several efforts have explored the use of large language models (LLMs) to generate real-time semantic descriptions for blind users. For instance, Song et al. combine smartphone LiDAR with GPT-based prompts to describe unfamiliar environments [11], while Kumaresan et al. propose a pipeline that translates images into audio using LLM-generated captions [12]. However, such models often lack spatial precision. A recent benchmark study found that vision-language models tend to hallucinate or produce vague spatial cues when asked to support blind navigation [13], limiting their standalone utility in fine-grained object localization.

These observations reveal a clear gap: an assistive solution that fuses semantic voice input, real-time visual perception, and head-tracked spatial audio, enabling objects to “call out” from their actual 3D locations so that BVI users can

*Resrach supported by the Waseda University Green Computing Systems Research Organization

Chenxin Qin is with the Graduate School of Creative Sciences and Engineering, Waseda university, Tokyo, Japan, (e-mail: qinchenxin@fuji.waseda.jp).

Yukiko Iwasaki is the faculty with the Graduate School of Creative Sciences and Engineering, Waseda university, Tokyo, Japan, (e-mail: kamiwaza@ruri.waseda.jp).

Chenyang Li is with the Graduate School of Creative Sciences and Engineering, Waseda university, Tokyo, Japan, (e-mail: lichenyang@ruri.waseda.jp).

Hiroyasu Iwata is the faculty with the Graduate School of Creative Sciences and Engineering, Waseda university, Tokyo, Japan, (e-mail: jubi@waseda.jp).

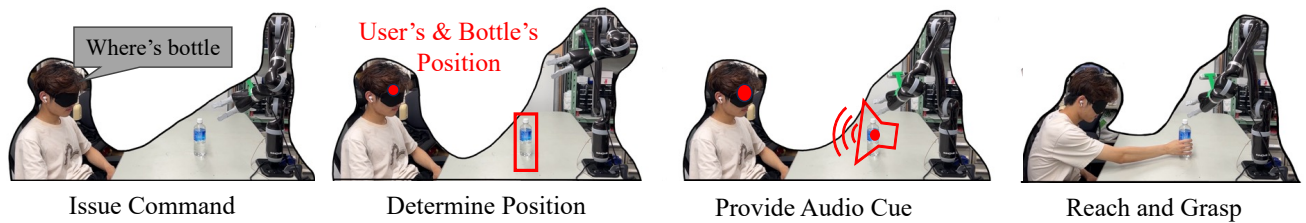


Fig. 1. Task flow: voice command, user and object localization, and spatial audio guidance for grasping.

confidently and independently reach them. By integrating an Intel RealSense D435i, a Kinova Jaco2 robotic arm, and Apple AirPods Pro, our system transforms spoken commands into head-aligned spatial audio cues, rendering the target as if it were “speaking” from its true position.

To validate the effectiveness of our proposed spatial-audio design, we conducted a user study addressing the following questions:

- (1) Can our head-tracked spatial audio guidance outperform conventional speech-only descriptions in terms of task efficiency?
- (2) Does it improve users’ subjective experience, including perceived workload and usability?

II. METHOD

A. System Overview

The system enables BVI users to locate tabletop objects using spatial audio cues. It consists of three main layers: the user, the mobile app, and the ROS backend. All interactions are voice-driven and hands-free.

Fig. 1 illustrates the task flow through real-world photos. The user issues a voice command, such as “Where’s the bottle?” The system first localizes the user and then scans the workspace to detect and locate the requested object. Finally, a directional sound is rendered to guide the user toward the target for grasping.

This figure offers a high-level overview of the closed-loop interaction, without requiring technical details. The following sections describe each component—hardware setup, mobile app, backend system, and direction estimation—in detail.

B. Hardware Setup

Kinova Jaco2 robotic arm: A 7-DOF Jaco2 was mounted on the table side to carry the RGB-D camera for workspace scanning. In this study the arm was used only for perception and object localization; no robotic grasping or hand-over was executed.

Intel RealSense D435i RGB-D camera: The camera was rigidly attached to the end-effector and provided synchronized RGB and depth for 3D localization of the user and objects.

Apple AirPods Pro: Earphones were used as the spatial audio output device with built-in head tracking to update the perceived source direction in real time. For the experiment, Active Noise Cancellation (ANC) was enabled to mask incidental cues (e.g., object placement sounds); Transparency

Mode is discussed only as a deployment consideration and was not used during data collection.

C. Mobile App: Command Input and Audio Feedback

The mobile app fulfills two sequential roles in the interaction loop. First, it captures user voice commands and head orientation data, forwarding them to the backend server via WebSocket. Then, upon receiving spatial feedback (azimuth, elevation, distance) from the backend, it renders a directional audio cue using Apple’s spatial audio framework. These two stages—input acquisition and feedback rendering—are tightly coupled via the backend perception system (Fig.2).

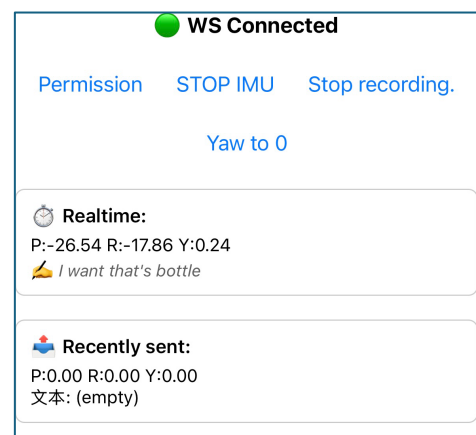


Figure 2. App screenshot

The user interface is implemented as a mobile application using React Native, a JavaScript-based framework for building native iOS and Android apps from a shared codebase. The app is optimized for BVI users and enables hands-free interaction through natural language input and audio-based guidance.

Upon launch, the app enters a passive listening state, continuously awaiting user voice input. Spoken commands (e.g., “Where is the bottle?”) are transcribed using the react-native-voice module. Simultaneously, the app activates Apple AirPods Pro’s inertial sensors through the react-native-headphone-motion library and captures head orientation data in the form of pitch, roll, and yaw.

The voice command and averaged head orientation are serialized into a structured JSON payload and transmitted to the backend system via a WebSocket connection. A heartbeat and reconnection mechanism are implemented to ensure stable communication.

After receiving spatial guidance parameters from the backend—typically azimuth, elevation, and distance relative to the user’s head—the app renders a directional audio cue to guide object localization. The spatial audio is generated using Apple’s AVFoundation framework, specifically the AVAudioEnvironmentNode, which simulates a virtual sound source in 3D space. The perceived direction is continuously corrected using real-time head-tracking to ensure alignment with the user’s orientation.

For the cue signal, we adopt a “ping-pong bouncing” sound. Prior studies have shown that such impulsive, non-verbal sounds yield higher localization accuracy than synthetic speech or white noise, particularly for BVI users during spatial navigation tasks [6].

All core interactions are fully voice driven. The graphical interface is included only for development and debugging. This mobile component serves as a lightweight, intuitive control and feedback interface, tailored for non-visual usage in dynamic environments.

D. ROS Backend & Direction Estimation

The backend is implemented on ROS Noetic (Ubuntu 20.04, Python 3.8). It performs all core perception, localization, and feedback-parameter generation required for real-time interaction. After receiving the user’s voice command and head orientation from the mobile app via WebSocket, it launches a closed-loop process to compute the spatial relationship between the user and the requested object.

Perception pipeline: An Intel RealSense D435i RGB-D camera mounted on the robotic arm first detects the user. A face-detection module identifies the user and extracts the head position in the camera coordinate frame. After localizing the user, the robotic arm executes a planar sweep over the tabletop workspace to search for the target. Object detection is carried out with a YOLOv8-based model, which returns real-time bounding boxes for predefined classes. For each detection, the system reads the depth from the RGB-D sensor and estimates the object’s 3D coordinates. The detected class is then matched to the keyword parsed from the user’s voice command to select the intended item.

Direction estimation: Let p_{obj}^c and p_{head}^c denote the 3D positions of the target object and the user’s head in the camera frame (meters). A fixed extrinsic transform $T_{W \leftarrow c} \in SE(3)$ maps camera coordinates to the world/table frame, and $T_{H \leftarrow W}$ maps world coordinates to a head-centric frame whose x axis points to the user’s right, y axis upward, and z axis forward. We compute the displacement in the head frame as

$$\Delta p^H = T_{H \leftarrow W} T_{W \leftarrow c} (p_{obj}^c - p_{head}^c) = [\Delta x \ \Delta y \ \Delta z]^T$$

The azimuth and elevation are then defined by:

$$\theta_{azimuth} = \arctan2(\Delta y, \Delta x)$$

$$\theta_{elevation} = \arctan2(\Delta y, \sqrt{\Delta x^2 + \Delta z^2}),$$

with ranges azimuth $\in (-\pi, \pi]$ (positive = right of the head) and elevation $\in (-0.5\pi, 0.5\pi]$ (positive = above the horizontal plane). The distance is $d = \|\Delta p^H\|_2$ (meters). We

wrap azimuth to $(-\pi, \pi]$, clamp elevation to $(-0.5\pi, 0.5\pi]$, and guard against $d \approx 0$ to avoid numerical instability. To suppress jitter, head pose is exponentially smoothed and time-aligned to camera timestamps by linear interpolation before each cue update. The triple (azimuth, elevation, d) is sent to the client via WebSocket.

All coordinate frames are right-handed. The world frame is anchored on the table plane estimated by RANSAC, and the head-centric frame origin corresponds to the user’s head position. The term “fixed extrinsic transform” refers to a rigid transform that was calibrated once per session through a hand-eye procedure. The 500 ms parameter serves as a temporal stability threshold for object detection rather than indicating any system latency.

Target disambiguation. If multiple detections match the requested class, we select the instance minimizing azimuth and, secondarily, d ; if no instance remains stable within the persistence window, the backend returns a “not found” status and the client prompts for re-query.

Coordinate note. This calibration was performed once per session and stored in the ROS tf tree.

E. End-to-End Task Pipeline

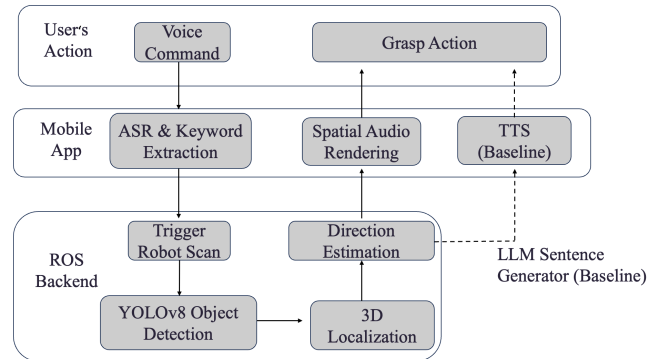


Figure 3. System Architecture and Module Interaction

Fig. 3 summarizes the overall system architecture and end-to-end task flow. It integrates the previously described modules—voice command processing, visual perception, 3D localization, and directional audio feedback—into a closed-loop pipeline. Solid lines represent the spatial audio (SA) condition, while dashed lines indicate the speech-only (SO) baseline. This diagram highlights the modular interaction across system layers and completes the implementation overview.

III. EXPERIMENTS

A. Experimental Setup

Workspace: A rectangular table (1.20 m \times 0.80 m, height [0.73 m]) served as the interaction surface. A 7-DOF Kinova Jaco2 was mounted on one side solely to carry the RGB-D camera and scan the workspace; no robotic manipulation was performed in this study.

Objects: Plastic bottles filled with water were selected as target items and placed randomly on the table in each trial.

Their positions were changed between trials to prevent users from relying on memory and to simulate unpredictable object locations.

Participant condition: Sighted participants were blindfolded to simulate non-visual interaction. Apple AirPods Pro (2nd gen) were worn throughout; ANC was enabled to mask incidental cues (e.g., object placement sounds). Participants sat centered along the table’s long side facing the workspace.

Environment: The study took place in a shared laboratory (not acoustically isolated). Background noise was no interference with command recognition or cue playback was observed. All modules ran on a locally connected machine.

B. Baseline Condition

To create a speech-only control condition, a lightweight Python node on the ROS backend invokes the OpenAI GPT-4 API. Once the target object is localized, the node computes its straight-line distance from the user (centimeters), horizontal bearing (positive = right, negative = left), and vertical offset (positive = above eye level, negative = below). These three values are inserted into a fixed prompt, and GPT-4 returns a concise English sentence—for example:

“The bottle is about 35 centimeters to your right and 12 degrees below eye level.”

The sentence is encapsulated in a JSON packet and transmitted over the existing WebSocket channel to the mobile client. The client employs react-native-tts, which wraps iOS AVSpeechSynthesizer, to render the text as monaural speech and play it once through the AirPods. Because this baseline provides guidance solely through a spoken sentence—without spatial audio rendering or head-tracking—it represents a conventional speech-based assistive approach against which the proposed spatial-audio method is evaluated.

C. Evaluation Procedure

Each session began with four practice trials (two speech-only, two spatial-audio) that were not analyzed. The formal experiment then comprised two twelve-trial blocks, one SO and one SA, block order was counter-balanced across participants. At the start of every trial, the experimenter—while the blindfolded participant wore the previously specified ANC mode AirPods Pro 2—placed the bottle at a uniformly random point on the tabletop. The system calculated the head-to-target vector and issued the corresponding audio cue.

A trial was scored as successful if the participant’s first reach contacted the bottle; it was scored as a failure if the reach missed or if no contact occurred within twenty seconds. For each trial, we recorded success/failure, and response time, measured from the moment the system completed scanning and localization to the participant’s first physical contact with the target.

This measure therefore represents the user’s reaching duration after visual sensing and computation were completed, excluding internal system delays.

After each twelve-trial block, the participant completed the SUS and RAW NASA-TLX questionnaires.

Ethical Considerations—This research meets the criteria for exemption from formal ethics committee review, as it does not involve identifiable human subject data. All participants were members of the university laboratory and voluntarily took part in the study with informed consent. Importantly, no personal or biometric data were collected, and all experimental results were anonymized at the source. The study design involves no invasive procedures, and no physical contact occurred between the robotic system and participants at any stage. The equipment used, including the Apple AirPods Pro and Kinova Jaco2 robotic arm, complies with relevant safety certifications (e.g., CE marking). The experimental protocol was conducted in accordance with standard institutional guidelines for minimal-risk human interaction studies.

IV. RESULTS

A. Objective Task Performance

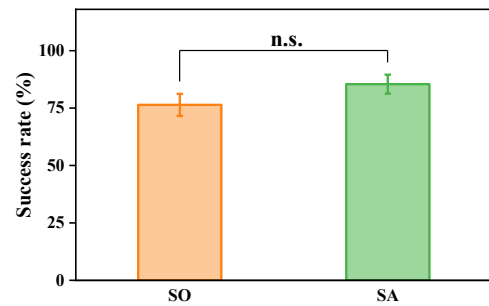


Fig 4. Success rate under SO and SA conditions. Significance assessed with the Wilcoxon signed-rank test (n.s.= not significant).

Success Rate: Twelve participants completed 12 trials per condition (SO, SA). Per-participant success rates were compared with a Wilcoxon signed-rank test. SA tended to be higher for 8/12 participants (4/12 lower), but the difference was not significant ($Z = -1.219, p = 0.223, N = 12; r = 0.35$) This indicates that task accuracy was comparable between conditions (Fig.4).

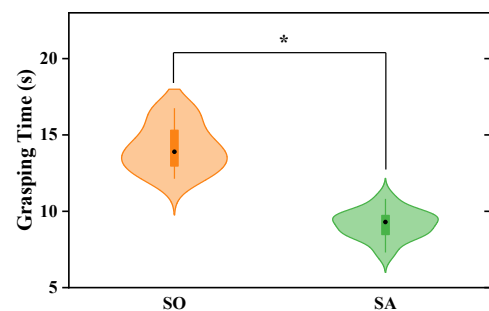


Fig 5. Localization and reaching under SO and SA conditions. Error bars represent ± 1 standard deviation. * $p < 0.05$.

Localization and reaching time: To evaluate task efficiency, we compared the localization and reaching times between the SO and SA conditions. The time was measured from scan completion to object contact. As shown in Fig. 5, the mean localization and reaching time was significantly shorter in the SA condition ($M=9.15$ s) than in the SO condition ($M=14.18$ s), $t(11)=9.454$, $p<0.05$. These results suggest that spatial-audio guidance effectively reduced the time required for object acquisition.

B. Subjective Task Performance

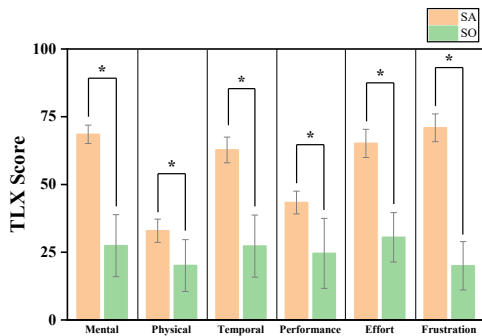


Fig 6. NASA-TLX workload scores under SO and SA conditions

As shown in Fig.6, the SA condition resulted in significantly lower workload ratings across all six NASA-TLX subscales—Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration—compared to the SO condition ($p<0.05$ for all). This consistent reduction indicates that spatial-audio guidance was perceived as less mentally and physically demanding and induced less temporal pressure, effort, and frustration during the task.

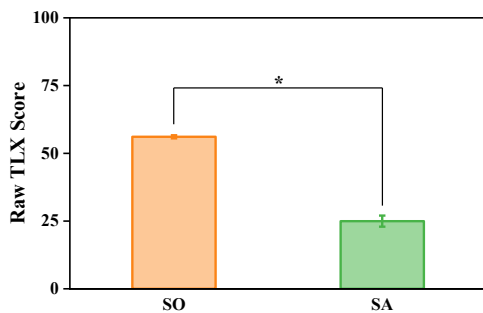


Fig 7. Raw NASA-TLX workload scores under SO and SA conditions.

As shown in Fig.7, the overall task load measured by RAW NASA-TLX was also significantly lower under the SA condition ($M=24.98$) than under the SO condition ($M=56.09$), $p<0.05$. This confirms that spatial-audio guidance substantially reduced the participants' overall perceived workload.

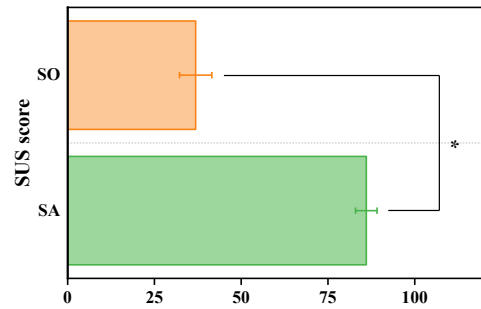


Fig 8. SUS scores under SO and SA conditions.

System Usability Scores. As shown in Fig.8, the SA condition yielded substantially higher SUS scores than the SO condition. The mean difference was 49.17 points, $p<0.05$, and this difference was statistically significant.

V. DISCUSSION AND CONCLUSION

A. Summary of Key Findings

This study investigated a spatial audio-based prompting approach that simulates target objects as if they emit sound from their actual location. By rendering directional audio cues aligned with the user's head orientation, the system aimed to provide intuitive, egocentric guidance for object localization and reaching. To evaluate its effectiveness, we compared this method with a baseline SO condition, in which object locations were described verbally using synthesized speech. Results showed that while both conditions achieved comparable success rates, the SA condition significantly reduced the time required to complete the task. In addition, participants rated the SA system as substantially more usable, as reflected by significantly higher SUS scores. These findings suggest that simulating object-originated spatial sound can enhance user efficiency and experience, even when the overall task accuracy remains similar.

B. Interpretation and Implications

The advantage of spatial audio lies in its egocentric and continuous directional feedback. By aligning virtual sound sources with the user's head orientation, it enables intuitive localization without requiring verbal-spatial translation. In contrast, the speech-only condition provides discrete information such as "the bottle is at 45 degrees to your right, 0.8 meters away," which requires cognitive mapping from linguistic descriptions to motor actions. Notably, sound-based localization is a perceptual skill deeply rooted in biological instincts. By leveraging this natural mechanism, spatial audio offers a more embodied and efficient form of non-visual guidance.

The similar success rates observed across both conditions suggest that each method is functionally reliable within the tabletop task setting. This study serves as an intermediate validation step toward broader spatial applications. While both methods suffice for near-field reaching tasks, verbal descriptions may become less effective in large-scale

navigation scenarios—where spatial complexity increases, and linguistic cues grow harder to process. Spatial audio holds potential as a scalable alternative in such contexts, especially for guiding blind users during real-world movement.

C. Limitations

This study has several limitations that constrain the generalizability of its findings. First, one participant with unilateral hearing loss showed a markedly lower success rate under the spatial-audio condition (2/12) compared to the speech-only condition (9/12). The participant's data were excluded from the main analysis due to this extreme deviation. This case highlights the system's dependence on binaural auditory perception, suggesting reduced effectiveness for users with asymmetric or impaired hearing.

Second, the participant pool consisted exclusively of sighted male students. While blindfolding simulates visual occlusion, it does not fully capture the perceptual strategies or compensatory behaviors of individuals with long-term visual impairments. Moreover, the limited sample size (N=12) restricts statistical power and may limit the detection of subtle interaction effects.

Third, the present work did not directly measure the system's internal latency—from voice command reception to cue onset—which would help to distinguish system delay from user response time.

Lastly, the experimental setting was confined to tabletop object localization. The effectiveness of spatial audio in larger-scale or dynamic tasks, such as indoor navigation or mobile grasping, remains unexplored.

D. Future Work

Future research should extend the system to support multi-object spatial cueing. In practical scenarios, users often need to distinguish and locate multiple items simultaneously. Enabling concurrent auditory streams—each spatialized to a different target—would require solving challenges related to signal separability, prioritization, and cognitive load management.

In addition, the current system could be integrated into larger-scale assistive navigation frameworks. Combining object-level spatial guidance with broader mobility support would benefit BVI users in real-world environments. For example, spatial audio could be embedded in wearable navigation robots or smart canes, allowing users to localize both their next step and a desired object in the same auditory space. This dual-level guidance may enhance task continuity, reduce verbal dependency, and offer a more seamless and autonomous user experience.

Future versions may incorporate monaural or rhythmic encoding schemes to improve accessibility for users with asymmetric hearing.

health: Vision beyond 2020," *Lancet Glob. Health*, vol. 9, no. 4, pp. e489–e551, Apr. 2021.

- [2] M. Kyrarini, M. Zand, and K. Kodur, "Assistive robots for persons with visual impairments: Current research and open challenges," in *Proc. 16th Int. Conf. PErvasive Technol. Related to Assistive Environ.* (PETRA '23), Corfu, Greece, Jul. 2023, Art. no. 7, pp. 1–7.
- [3] A. Tsuji, S. Aihara, S. Tanaka, L. Guinot, and H. Iwata, "Enhancing sound-source localization in blind soccer: Analyzing head-movement strategies and localization abilities for static and dynamic auditory cues via a virtual spatial-acoustic system," in *Proc. 2024 IEEE Int. Conf. Syst., Man, Cybern.* (SMC), Maui, HI, USA, Oct. 2024, pp. 3763–3768.
- [4] H. Miyakawa, N. Kuratomo, H. E. B. Salih, and K. Zempo, "Auditory Uta-Karuta: Development and evaluation of an accessible card game system using audible cards for the visually impaired," *Electronics*, vol. 10, no. 6, Art. no. 750, 2021.
- [5] T. Asakura, "Bone conduction auditory navigation device for blind people," *Applied Sciences*, vol. 11, no. 8, Art. no. 3356, 2021.
- [6] S. Paré, M. Bleau, I. Djerourou, V. Malotau, R. Kupers, and M. Ptito, "Spatial navigation with horizontally spatialized sounds in early and late blind individuals," *PLoS ONE*, vol. 16, no. 2, Art. no. e0247448, 2021.
- [7] X. Hu, A. Song, Z. Wei, and H. Zeng, "StereoPilot: A wearable target-location system for blind and visually impaired using spatial audio rendering," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1621–1630, 2022.
- [8] F. Fei, S. Xian, R. Yang, C. Wu, and X. Lu, "A wearable visually impaired assistive system based on semantic vision SLAM for grasping operation," *Sensors*, vol. 24, no. 11, Art. no. 3593, 2024.
- [9] G. M. T. Paratore and B. Leporini, "Exploiting the haptic and audio channels to improve orientation and mobility apps for the visually impaired," *Universal Access in the Information Society*, vol. 23, no. 2, pp. 859–869, 2024.
- [10] N. Bonda, C. Qin, Y. Iwasaki, H. Iwata, "HOVER-BVI: Handover Vision-Language Embodied Robot System for BVI Users," *2025 9th International Conference on Robotics and Automation Sciences (ICRAS)*, pp. 1–5, 2025.
- [11] J. E. Kim, G. Sahas, and M. Bessho, "Toward assisting blind individuals in exploring unfamiliar indoor environments using multimodal LLM and smartphone LiDAR," in *Proc. IEEE Int. Conf. Consumer Electron.* (ICCE), Las Vegas, NV, USA, Jan. 2025, pp. 1–6, doi: 10.1109/ICCE63647.2025.10930138.
- [12] T. Kumaresan, J. B. G. Thangam, G. A. Rajkumar, M. Eswaran, R. G. Kumar, and T. Rajkumar, "Image to audio using LLM for blind people" in *Proc. 1st Int. Conf. Data, Comput. Commun.* (ICDCC), Bhopal, Madhya Pradesh, India, Nov. 2024, pp. 610–614.
- [13] P. Rahmazadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen, "Vision-language models are blind," in *Proc. Asian Conf. Comput. Vis.* (ACCV), Hanoi, Vietnam, Dec. 2024, pp. 18–34.

REFERENCES

- [1] M. J. Burton, J. Ramke, A. P. Marques, R. R. A. Bourne, N. Congdon, I. Jones, *et al.*, "The Lancet Global Health Commission on global eye