

An AI-Based Shopping Assistant System to Support the Visually Impaired

Larissa R. de S. Shibata^{1†}, Ankit A. Ravankar¹, Jose Victorio Salazar Luces¹, and Yasuhisa Hirata¹

¹Department of Robotics, Tohoku University, Sendai, Japan
(E-mail: de.souza.shibata.larissa.rika.r4@dc.tohoku.ac.jp, {ankit.j.salazar, hirata}@srd.mech.tohoku.ac.jp)

Abstract: Shopping plays a significant role in shaping consumer identity and social integration. However, for individuals with visual impairments, navigating in supermarkets and identifying products can be an overwhelming and challenging experience. This paper presents an AI-based shopping assistant prototype designed to enhance the autonomy and inclusivity of visually impaired individuals in supermarket environments. The system integrates multiple technologies, including computer vision, speech recognition, text-to-speech synthesis, and indoor navigation, into a single, user-friendly platform. Using cameras for ArUco marker detection and real-time environmental scanning, the system helps users navigate the store, identify product locations, provide real-time auditory guidance, and gain context about their surroundings. The assistant interacts with the user through voice commands and multimodal feedback, promoting a more dynamic and engaging shopping experience. The system was evaluated through experiments, which demonstrated its ability to guide users effectively and improve their shopping experience. This paper contributes to the development of inclusive AI-driven assistive technologies aimed at enhancing accessibility and user independence for the shopping experience.

Keywords: Assistive Robotics Technology, Visual impairment, Semantic Query Processing, Smart shopping cart, Indoor Navigation, Human-Machine Interaction.

1. INTRODUCTION

Shopping is a vital aspect of everyday life that extends beyond the acquisition of goods. It plays a role in personal identity, social participation, and the sense of autonomy. However, for individuals with visual impairments, shopping in physical environments introduces considerable barriers. Navigating stores, locating specific products, and making informed purchasing decisions are tasks heavily reliant on visual input, often requiring assistance from others. According to the World Health Organization (WHO), over 2.2 billion people globally are affected by visual impairments, highlighting the broad social importance of inclusive and accessible shopping solutions [1].

In recent years, several assistive technologies have emerged to address these challenges[2]. Mobile applications such as TapTapSee, Navilens, Seeing AI, and Vision utilize device cameras and voiceover functionalities to identify objects and texts[3]. For navigation, GPS-based solutions including BlindSquare, VoiceMap, and SWAN offer route guidance, primarily in outdoor environments [4]. Additionally, crowd-sourced platforms like “Be My Eyes”[5] connect visually impaired individuals to volunteers or AI services for ad-hoc assistance. While these solutions have made significant contributions, they face notable limitations. Specifically, mobile apps often require manual control and familiarity with multiple interfaces, GPS-based solutions are unsuitable for precise indoor navigation, and remote human-based solutions may not offer real-time or contextually rich guidance. More recently, robotics and AI-based systems have been explored to provide embodied assistance to elderly or visually impaired individuals [6-10]. Robotic guides and wearable navigation aids have been proposed to offer real-time environment feedback using sensors such as LiDAR, depth cameras, and computer vision techniques[11-14]. For example, autonomous robotic assistants have been trialed

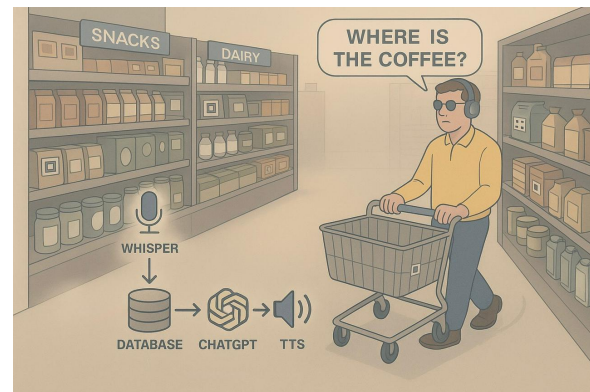


Fig. 1: Concept image of our AI-based Shopping Assistant System for Visually Impaired. Our proposed system guides user to different sections in a store and provide real-time multi-modal information about the environment and products on the shelves.

for guiding users in structured indoor spaces like museums and airports. Recent advancements in sensor technology, including multi-sensor fusion and autonomous robot mobility have brought significant progress in realizing smart and autonomous navigation for diverse applications [15-18]. While promising, these systems are often specialized, infrastructure-dependent, or focused solely on navigation rather than comprehensive in-store support. Furthermore, indoor localization technologies, including RFID tags, BLE beacons, and vision-based approaches such as ArUco markers have been investigated for assistive applications, yet integrating them seamlessly with user interaction models remains an open challenge[19].

Furthermore, most existing systems approach shopping as a rigid and goal-oriented process, supporting only predefined queries or fixed routes. This perspective overlooks the dynamic and opportunistic nature of in-store shopping, where consumers often make spontaneous decisions influenced by promotions, store layouts,

† Larissa R. de S. Shibata is the presenter of this paper.

and cultural preferences[20]. Such subtle and context-dependent behaviors remain largely inaccessible to visually impaired individuals using current technologies [21-24].

To address these gaps, this work proposes a unified AI-based shopping assistant system embedded within a smart shopping cart. Unlike prior solutions that rely on disparate devices and fragmented functionalities, the proposed system integrates computer vision, automatic speech recognition (ASR), natural language interaction, and indoor localization using ArUco markers into a single platform. Through voice commands and multimodal feedback, users can explore store environments, identify product locations and details, and receive context-aware auditory guidance—all without switching devices or interfaces. This design aims to promote not only accessibility but also autonomy and spontaneity during shopping experiences. The main contributions of this work are summarized as follows:

- A unified shopping assistance platform that integrates environment description, product information retrieval, and navigation into a single embodied system tailored for visually impaired users.
- An AI-driven interaction model combining Whisper ASR, ChatGPT for dialogue and contextual reasoning, and ArUco marker-based localization for precise indoor navigation.
- User-centered experimental evaluation under normal and simulated visual impairment conditions, demonstrating system effectiveness and analyzing usability and workload through NASA-TLX metrics.

The remainder of this paper is organized as follows. Section 2 introduces the proposed system architecture and implementation details. Section 3 presents the experimental design, results, and analysis. Finally, Section 4 concludes with discussions on system performance, limitations, and future development directions.

2. PROPOSED SYSTEM

This section introduces the AI-based assistive shopping system developed to enhance the autonomy of visually impaired individuals in supermarket environments. The system is designed to address common challenges such as locating store sections, retrieving product information, and navigating aisles. To provide these functionalities in an intuitive and accessible manner, the system integrates computer vision, speech recognition, natural language processing, and indoor localization into a unified smart shopping cart platform. The proposed system supports three core user functions:

- **Environment Understanding and Awareness:** Delivering real-time auditory descriptions of store sections and surroundings.
- **Product Information Retrieval:** Enabling users to query product details, including location and price, through voice commands.
- **Indoor Navigation Assistance:** Guiding users to desired products or sections with step-by-step voice instructions.

To achieve these capabilities, the system architecture consists of five functional modules: *Listening*, *Capture*,

Processing, *Navigation*, and *Speaking*. The relationship among these modules is illustrated in Figure 2.

2.1. System Operation Workflow

The system remains in an idle state within the *Listening Module* until activated by the user via one of three physical buttons on the cart interface:

- **Environment Description Button:** Initiates the *Image Capture Module*, which activates multiple onboard cameras to capture images from various angles and forwards them to the Processing module.
- **Voice Command Button:** Triggers the *Audio Capture Module*, allowing the user to dictate questions or requests. The audio input is transcribed into text using the Whisper speech recognition model. If necessary, this also activates the Image Capture and Processing modules sequentially.
- **Section Description Button:** Provides real-time auditory feedback by utilizing ArUco marker recognition data. The recognized marker is interpreted to determine the user's current section, and the information is announced via the Speaking module.

While in operation, the system's cameras continuously monitor the environment for ArUco markers, independent of button presses. Detection of these markers plays a critical role in indoor localization, allowing the system to determine the cart's precise position and orientation relative to store sections or shelves.

When user input is processed in the *Processing Module*, the AI interprets the command and selects the appropriate action pathway. This may involve:

- Performing a semantic search in the local product database using text embeddings to retrieve relevant product information such as price or shelf location.
- Sending images for analysis via a vision-language model integrated with ChatGPT to describe the environment or answer user-specific queries.
- Accessing indoor navigation functions to calculate and initiate guidance to a target destination.

If navigation is required, the system activates the *Navigation Module*. This module computes the current position using detected ArUco markers and plans a path to the target location. The path is translated into a sequence of human-understandable audio instructions, which are then relayed to the user through the *Speaking Module*.

The Speaking Module utilizes a text-to-speech (TTS) engine to convert AI-generated textual responses or navigation instructions into natural-sounding speech. The audio output is transmitted through the user's headphones to ensure clear communication without external disturbances. Upon completion of each response cycle, the system returns to the *Listening Module* to await the next user interaction.

2.2. ArUco Marker Encoding and Database Structure

ArUco markers are fundamental to the system's indoor positioning capability. Each marker encodes a unique identifier (ID) that corresponds to either a specific shelf or a designated location within the supermarket layout. These marker IDs are mapped to human-readable labels, such as "Dairy," "Snacks," or "Cleaning Products," via

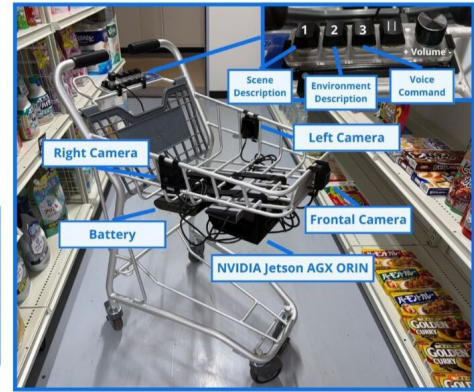
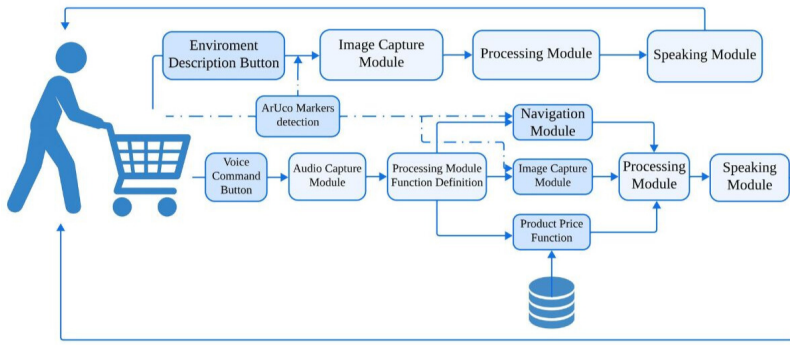


Fig. 2: Architecture of the proposed system, including the image capture, audio capture, processing, navigation, and speaking modules. The developed smart cart with sensor integration is also presented. Inset image shows the button module developed for the user.

a local reference table stored in the database. The product database is structured to store comprehensive information, including:

- Product name, brand, and variety.
- Shelf ID linked to the corresponding ArUco marker.
- Vertical position on the shelf (from top to bottom).
- Horizontal position on the shelf (from left to right).
- Product price.

This structured organization enables accurate retrieval and delivery of product location and pricing information in response to user queries. By leveraging semantic search capabilities through text embeddings, the system can handle variations in user input and still return precise or close-matching product results.

3. SYSTEM ARCHITECTURE

The proposed system adopts a modular and scalable architecture, where each functional block—Listening, Capture, Processing, Navigation, and Speaking—is implemented as an independent thread. These threads communicate and synchronize using shared queues and signaling mechanisms to ensure proper coordination and responsiveness. For example, when a user presses the voice command button, the Audio Capture thread captures and transcribes the audio, then passes the resulting text to the Processing thread through a shared communication buffer. Depending on the interpretation of the input, the Processing thread may trigger additional actions, such as image processing, product lookup, or navigation planning.

This multithreaded design enables parallelism while preserving modularity, making it possible to update or replace individual components without affecting the overall workflow. Each module operates autonomously and only activates when relevant input is received, ensuring efficient resource usage and low-latency responses.

Moreover, this architecture promotes the reusability of each module in other applications. For instance, the voice interaction and environment description threads can be integrated into service robots or public information kiosks for visually impaired users. Similarly, the navigation planning logic based on visual markers and route generation can be repurposed for indoor guidance sys-

tems in hospitals, libraries, or other indoor spaces. The system’s thread-based modularity lays the groundwork for developing a more general methodology for integrating multi-modal AI components into a wide range of assistive and service-oriented technologies.

3.1. Image Capture Module

The Image Capture Module is responsible for acquiring visual data from the supermarket environment and detecting ArUco markers for localization. This module plays a two important role: recognizing the user’s location and surroundings, and providing visual input for environment description when requested.

To achieve wide coverage and robust perception, the system employs three ELP-USBG1200P01 cameras, each equipped with a 120-degree distortion-free lens. These cameras are strategically mounted on the left, front, and right sides of the shopping cart to capture a comprehensive view of the environment. We used the NVIDIA Jetson AGX Orin platform for collecting and processing the sensor data, doing all the computations, and running the AI modules. The system was running on Ubuntu Linux 22.04.

As soon as the system is powered on, it continuously begins to recognize ArUco markers. The cameras detect 5x5-sized ArUco markers placed in front of each shelf. If the “*Section Description*” button is pressed, the name of the section associated with the marker is read aloud to the user. Since supermarkets have many shelves, to avoid overlapping commands, the system applies two criteria for reading a marker: the first is the marker’s size, which varies depending on the distance from the cart when the image is captured, larger markers (closer) have higher relevance. The second factor is the time since the marker was last read; a marker is only re-read after a certain interval or after a different marker has been recognized. After being recognized, the last marker ID from the left and right cameras is stored in two separate variables to help determine the user’s location and orientation. However, if no new marker is recognized after a few seconds, the variable is cleared to avoid false localization.

In contrast to continuous marker detection, photographic image capture for environmental analysis is performed on-demand. This occurs only when the user

presses either the *Environment Description Button* or the *Voice Command Button*. In such cases, the Image Capture Module simultaneously captures three images — one from each camera — representing the left, front, and right views.

These captured images are then transmitted to the Processing Module, where they are analyzed to generate environment descriptions or to assist in answering user queries via the integrated AI system.

3.2. Audio Capture Module

The Audio Capture Module is responsible for acquiring and processing user voice commands, which form a core part of the system’s interactive capabilities. This module enables users to query product information, request environment descriptions, or initiate navigation through simple verbal instructions.

When the user presses the *Voice Command Button*, the Audio Capture Module is activated. At this point, the system begins recording audio through an onboard microphone. To efficiently manage the recording process and avoid unnecessary data capture, the module employs a *Voice Activity Detection (VAD)* mechanism. Specifically, the WebRTC Voice Activity Detector (*webrtcvad*) is used to continuously analyze short frames of audio in real time. The VAD determines when speech is present and automatically stops the recording once speech ceases, signaling that the user’s utterance has concluded.

The recorded audio is then transcribed into text using the Whisper-1 automatic speech recognition (ASR) model provided by OpenAI. This model is equivalent to the Large-v2 version of Whisper and contains approximately 1.55 billion parameters, offering robust transcription capabilities for diverse speech patterns and environments.

Once transcription is complete, the resulting text is forwarded to the *Processing Module* for further interpretation. Depending on the nature of the user’s request, the Processing Module may perform semantic queries in the local database, initiate navigation instructions, or, if needed, activate the Image Capture Module to support context-aware responses.

By integrating VAD and ASR technologies, the Audio Capture Module provides an intuitive and streamlined mechanism for users to interact naturally with the system without requiring visual input or complex operations.

3.3. Question and Data Processing Module

The Question and Data Processing Module serves as the central decision-making component of the system, interpreting user commands and coordinating subsequent actions. Whether initiated through voice commands or environment description requests, this module determines the appropriate response path based on the nature of the user’s query.

Upon activation, the module processes inputs received from either the Audio Capture Module (in the form of transcribed text) or the Image Capture Module (as visual data). For voice-triggered queries, the transcribed command is first analyzed to classify the user’s intent into one of the following categories:

- **Product Information Retrieval:** Queries related to the location or price of products.
- **Navigation Requests:** Commands that require guiding the user to a particular product or store section.
- **General Environment or Contextual Queries:** Requests for broader information regarding the surroundings or requiring scene analysis.

If the command relates to product information, the system performs a *semantic search* within the local product database. To achieve this, the transcribed user query is converted into a high-dimensional vector representation using the *text-embedding-ada-002* model from OpenAI. This vector is then compared with precomputed vectors for each product entry in the database using cosine similarity, as defined in Equation 1:

$$\text{similarity}(A, B) = \frac{A \cdot B}{|A||B|} \quad (1)$$

where A and B represent the embedding vectors of the user query and the database entry, respectively, and $\|\cdot\|$ denotes the Euclidean norm. If the computed similarity exceeds a threshold of 90%, the system considers this a precise match and directly announces the product’s location (including shelf ID, vertical and horizontal shelf position) and price to the user. If the similarity falls below this threshold, the system retrieves and presents the top three most similar candidates, offering the user a choice to select the correct product.

When a user request pertains to navigation, the Processing Module engages the Navigation Module. It provides the current shopping cart location (based on ArUco marker detection) and the desired destination (obtained from the product database). The Navigation Module then generates a step-by-step route to guide the user to the target location.

If the user’s request falls outside the scope of product queries or navigation—for example, general questions about the surroundings or descriptions of the environment—the Processing Module forwards both the transcribed user input and any captured images to the integrated AI model for open-ended reasoning and natural language generation.

In this system, *ChatGPT 4o*, a recent and advanced version of the GPT (Generative Pre-trained Transformer) family, is utilized for this purpose. ChatGPT 4o is optimized for real-time dialogue, multimodal understanding, and context-aware reasoning tasks. Unlike base GPT models, ChatGPT 4o maintains conversational coherence, processes multimodal inputs (including images), and supports function calling mechanisms, making it especially suitable for dynamic and natural communication in assistive scenarios such as this[25].

Once ChatGPT 4o generates a contextually appropriate response, the system converts this text output into audio through the Speaking Module and delivers it to the user. Following the completion of any processing path, the system returns to the *Listening Module* to await the next user command.

3.4. Navigation Module

The navigation module provides optimized guidance to the user based on the location identified through con-

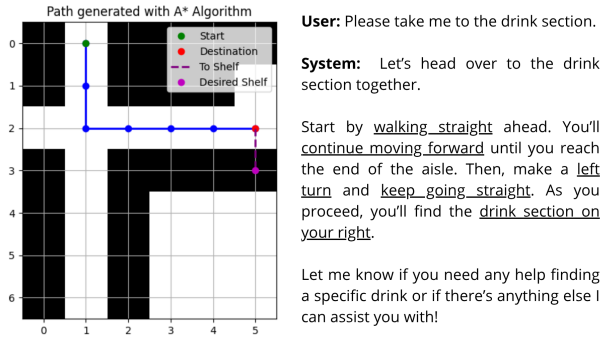


Fig. 3: Example of a route generated by the A* algorithm and the corresponding dialogue between the user and the system. The visualized path represents the optimal trajectory from the user's current location to the desired shelf, while the dialogue illustrates how the system translates this route into natural language instructions for a visually impaired user.

tinuous recognition of ArUco markers placed throughout the environment. Detecting these markers with the system's cameras makes it possible to determine both the user's current position on the supermarket map and their initial movement orientation, depending on whether the left or right camera detected the marker.

For route planning, the A* algorithm is used, known for its efficiency in finding optimal paths in graphs. It calculates the total cost $f(n)$ of each node based on the sum of the actual cost $g(n)$ from the starting point and a heuristic $h(n)$ that estimates the cost to the destination, as shown in Eq. (2):

$$f(n) = g(n) + h(n) \quad (2)$$

In this system, the algorithm's inputs are the initial position (obtained from the recognized ArUco marker), the target position (e.g., desired section or product), and the initial orientation (based on which camera identified the marker).

The module's output is an optimized path, represented as a sequence of movements on the supermarket map. To make navigation accessible and understandable for visually impaired individuals, this sequence of movements is converted into natural language instructions. This conversion is performed using a GPT-based model, which interprets the sequence of directions and generates clear, humanized audio commands such as "turn right," "go straight for two aisles," or "turn left at the next section." An example of an interaction with the system can be seen in Figure 3.

3.5. Speaking Module

Finally, the speaking module is responsible for converting the responses generated by the AI into natural-sounding audio output, enabling auditory interaction with the user. This module is essential for providing feedback to visually impaired users in a seamless and intelligible manner.

For this task, the system employs the *OpenAI TTS* (Text-to-Speech) tool with the Onyx voice. After processing the request, the textual response is broken into smaller segments, especially in cases of longer messages,

to reduce the perceived latency for the user. Each segment is individually converted into audio and stored in a playback queue, ensuring that the information is delivered in a smooth and efficient manner.

In addition to text-to-speech conversion, the system incorporates non-verbal auditory feedback at strategic moments during the interaction. A brief sound is played to indicate the start of voice recording in the audio capture module, informing the user that the system is actively listening. Another sound is emitted while the AI processes the response, signaling that the system is working on an answer. These audio cues enhance the interactivity and transparency of the system's behavior, especially for users with visual impairments and who rely entirely on auditory perception.

All synthesized speech and sound cues are delivered through a headphone output, allowing the user to receive private and uninterrupted feedback even in noisy store environments. Once the message or instruction is fully delivered, the system transitions back to the *Listening Module*, ready to accept the next input.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Two experiments were conducted to evaluate the usability and effectiveness of the proposed system under normal and impaired vision conditions. The experiments were designed to assess the system's ability to support in-store navigation and product detection tasks in a realistic shopping scenario. Each experiment consisted of two phases: one performed under normal vision, and another using a cataract-simulation device (see Fig. 4) to mimic moderate to severe visual impairment.

Both experiments were conducted with four participants, two of whom had received prior training with the system. To evaluate cognitive and physical workload, participants completed the NASA Task Load Index (NASA-TLX) questionnaire after each phase. The NASA-TLX assesses subjective workload across six dimensions: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. Each dimension is rated from 0 (low) to 100 (high), and the weighted sum provides an overall workload score. Higher scores indicate increased perceived workload or emotional strain[26].

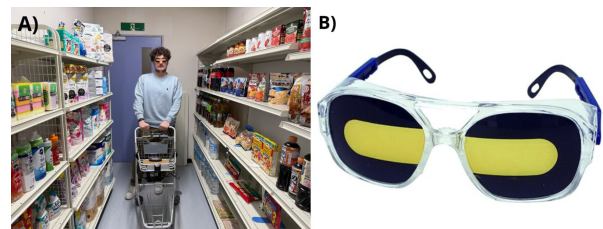


Fig. 4: (A) Experimental setup with the AI shopping assistant cart during tests. (B) Cataract simulation goggles used to reproduce the visual impairment condition during the evaluation.

4.1. Navigation Task Results

In this experiment, the participants were instructed to locate three store sections within a simulated supermarket. The requested sections were not visually associated with nearby products, ensuring that participants relied on system guidance. Different section targets were assigned across conditions to prevent memory bias. After each phase, participants completed the weighted NASA-TLX questionnaire. Results for the navigation task are summarized in Table 1.

Table 1: Navigation Task: Summary of results.

Condition	Normal Vision	Impaired Vision
Success Rate	91.6%	100%
Avg. Time (s)	59.6	57.4
Avg. Queries	4.3	3.75
NASA-TLX score	35.9	40.4

Although participants with normal vision could theoretically rely on visual cues, all information was intentionally hidden or unlabeled to force dependence on the AI system. Interestingly, one participant failed to reach a target section due to incorrect system feedback. This incident, combined with initial unfamiliarity, likely explains why the average success rate and NASA-TLX score appear worse under normal vision than impaired vision. This highlights how prior expectations and the mismatch between visual input and auditory feedback can increase cognitive load.

Among trained participants, the average number of queries was lower (3.0) compared to untrained participants (5.0), and the time to reach each section was shorter (73 s vs. 137 s), demonstrating a learning effect. However, the Frustration score was paradoxically higher among trained users (277.5 vs. 45). This may reflect raised expectations and increased sensitivity to system limitations.

In this experiment, Section 1 took the longest time to be located (105 seconds on average) and required the highest number of queries to the system (2.25). NASA-TLX scores revealed that Frustration and Performance were the most influential dimensions, with average weighted scores of 161.3 and 116.3, respectively.

Under simulated visual impairment, all participants completed the task successfully. Fewer queries were needed, and the average time was slightly lower. These results likely reflect both the absence of conflicting visual cues and increased familiarity with the system by the second phase.

For impaired vision, the most influential NASA-TLX dimensions were Performance (150.0) and Effort (135.0). These results suggest that visual impairment led participants to perceive the task as more cognitively demanding, requiring greater effort to achieve satisfactory performance.

Voice system latency (measured from the voice button press to audio output) averaged 19.3 seconds under normal vision and 20.9 seconds under impaired vision. This includes time for audio recording, transcription, processing, and text-to-speech synthesis. A total of six system errors or unexpected behaviors were reported, which likely

contributed to increased task duration or failure in some cases.

4.2. Product Detection Task Results

In the second experiment, participants were asked to find two products on store shelves using voice queries. The product names varied across phases and were generic (e.g., "instant noodles"), allowing the participant to choose among system-suggested options. The task was performed with and without the vision impairment simulation, and participants completed a NASA-TLX questionnaire after each phase. Results for this experiment are summarized in Table 2.

Table 2: Product Detection Task: Summary of results.

Condition	Normal Vision	Impaired Vision
Success Rate	87.5%	100%
Avg. Time (s)	143.8	136.1
Avg. Queries	9.25	7.25
NASA-TLX score	50.7	44.8

In the normal vision condition, one participant failed to locate a product. On average, more queries were needed (9.25), especially for the first item (6.25 queries), which took significantly longer to find (218 seconds vs. 69.5 seconds for the second). This suggests a learning curve during interaction.

Despite the availability of visual input, participants found the system difficult to use at first. The discrepancy between what they saw and what the system communicated may have caused frustration, reflected in high NASA-TLX Frustration (243.8) and Mental Demand (183.8) scores.

With simulated vision impairment, performance improved—likely due to increased system familiarity and focused reliance on auditory input. NASA-TLX scores showed highest weights on Frustration (172.5) and Effort (157.5). Voice latency averaged 18.0 seconds (normal vision) and 17.1 seconds (impaired).

Voice latency was slightly lower than in the navigation task, averaging 18.0 seconds (normal) and 17.1 seconds (impaired). Seven system errors occurred, often related to Whisper transcription issues.

4.3. NASA-TLX Analysis

A summary of weighted NASA-TLX scores per dimension is presented in Figure 5 and Figure 6 for each experiment.

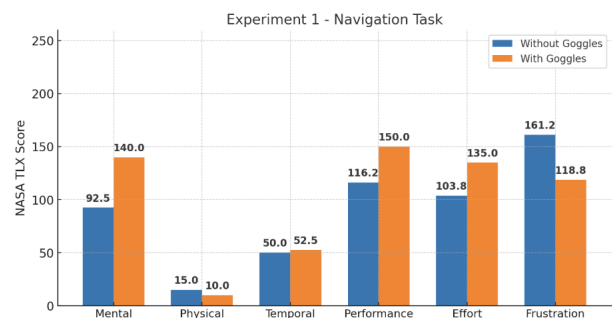


Fig. 5: Nasa TLX scores per dimension - Navigation Experiment.

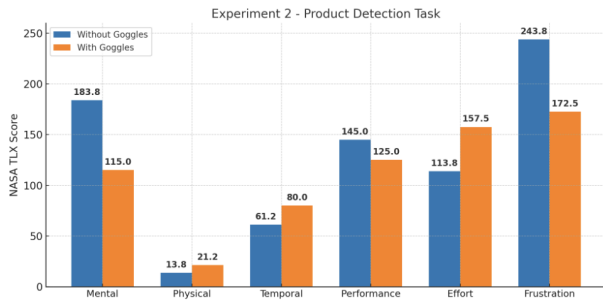


Fig. 6: Nasa TLX scores per dimension - Product Recognition Experiment.

Frustration and Perceived Performance were the most influential dimensions, especially during early interactions or when user expectations were not met. Over time, participant familiarity with the system led to reduced workload scores and increased efficiency.

5. CONCLUSION AND FUTURE WORK

This work presented an AI-based shopping assistant designed to support blind and visually impaired individuals by promoting greater autonomy during grocery shopping. Implemented as a smart shopping cart, the system integrates vision, speech, and language technologies to guide users, identify products, and interact conversationally.

Experimental evaluations conducted under both normal and simulated vision impairment conditions confirmed the system's ability to assist users in navigating the shopping environment and locating products, with improved performance through training. However, delays in response generation were observed due to processing time. To address this, faster local modules like faster-whisper and piper will be adopted. Improvements in navigation instruction clarity and user localization are also planned.

Future efforts include testing with visually impaired participants in real supermarkets to validate usability and expand functionality in real-world conditions.

6. ACKNOWLEDGMENTS

This work was partially supported by JST [COINEXT][Grant Number JPMJPF2201] and JSPS Kakenhi [Grant Number JP24K07399].

REFERENCES

- [1] W. H. Organization *et al.*, "World report on vision," in *World report on vision*, 2020.
- [2] A. Ogura and A. Ravankar, "Development of a shopping assist system," *The Proceedings of JSME annual Conference on Robotics and Mechatronics (Robomec)*, vol. 2023, pp. 1A2-H26, 2023.
- [3] O. Serena Sukhija and M. Bradley, "Innovative mobile technology targets low vision," *Optometry Times*, vol. 11, no. 2, pp. 18-20, 2019.
- [4] S. M. Baker, "Consumer normalcy: Understanding the value of shopping through narratives of consumers with visual impairments," *Journal of retailing*, vol. 82, no. 1, pp. 37-50, 2006.
- [5] "Be my eyes." <https://www.bemyeyes.com/>. [Accessed 07-05-2025].

- [6] A. Ravankar, Y. Kobayashi, A. Ravankar, and T. Emaru, "A connected component labeling algorithm for sparse lidar data segmentation," in *2015 6th International Conference on Automation, Robotics and Applications (ICARA)*, pp. 437-442, IEEE, 2015.
- [7] M. C. Domingo, "An overview of the internet of things for people with disabilities," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 584-596, 2012.
- [8] A. A. Ravankar, S. A. Tafrishi, J. V. S. Luces, F. Seto, and Y. Hirata, "Care: Cooperation of ai robot enablers to create a vibrant society," *IEEE Robotics & Automation Magazine*, vol. 30, no. 1, pp. 8-23, 2022.
- [9] M. A. Khan, P. Paul, M. Rashid, M. Hossain, and M. A. R. Ahad, "An ai-based visual aid with integrated reading assistant for the completely blind," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 6, pp. 507-517, 2020.
- [10] A. Ravankar, A. A. Ravankar, A. Rawankar, Y. Hoshino, and Y. Kobayashi, "Itc: Infused tangential curves for smooth 2d and 3d navigation of mobile robots," *Sensors*, vol. 19, no. 20, p. 4384, 2019.
- [11] M. Bamdad, D. Scaramuzza, and A. Darvishy, "Slam for visually impaired people: A survey," *IEEE Access*, 2024.
- [12] G. Lacey and K. M. Dawson-Howe, "The application of robotics to a mobility aid for the elderly blind," *Robotics and Autonomous Systems*, vol. 23, no. 4, pp. 245-252, 1998.
- [13] J. Salazar, K. Okabe, and Y. Hirata, "Path-following guidance using phantom sensation based vibrotactile cues around the wrist," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2485-2492, 2018.
- [14] Z. Liao, J. V. S. Luces, and Y. Hirata, "Human navigation using phantom tactile sensation based vibrotactile feedback," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5732-5739, 2020.
- [15] M. B. Alatise and G. P. Hancke, "A review on challenges of autonomous mobile robot and sensor fusion methods," *IEEE access*, vol. 8, pp. 39830-39846, 2020.
- [16] A. Ravankar, A. A. Ravankar, Y. Hoshino, M. Watanabe, and Y. Kobayashi, "Safe mobile robot navigation in human-centered environments using a heat map-based path planner," *Artificial Life and Robotics*, vol. 25, no. 2, pp. 264-272, 2020.
- [17] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52-57, 2002.
- [18] H. Beomsoo, A. A. Ravankar, and T. Emaru, "Mobile robot navigation based on deep reinforcement learning with 2d-lidar sensor using stochastic approach," in *2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR)*, pp. 417-422, IEEE, 2021.
- [19] F. Zafari, A. Gkeliias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568-2599, 2019.
- [20] E. Kostyra, S. Żakowska-Biemans, K. Śniegocka, and A. Piotrowska, "Food shopping, sensory determinants of food choice and meal preparation by visually impaired people. obstacles and expectations in daily food experiences," *Appetite*, vol. 113, pp. 14-22, 2017.
- [21] C. W. Yuan, B. V. Hanrahan, S. Lee, M. B. Rosson, and J. M. Carroll, "Constructing a holistic view of shopping with people with visual impairment: a participatory design approach," *Universal Access in the Information Society*, vol. 18, pp. 127-140, 2019.
- [22] S. Tullio-Pow, H. Yu, and M. Strickfaden, "Do you see what i see? the shopping experiences of people with visual impairment," *Interdisciplinary Journal of Signage and Wayfinding*, vol. 5, no. 1, pp. 42-61, 2021.
- [23] H. Yu, S. Tullio-Pow, and A. Akhtar, "Retail design and the visually impaired: A needs assessment," *Journal of Retailing and Consumer Services*, vol. 24, pp. 121-129, 2015.
- [24] M. Balconi, L. Angioletti, and C. Acconito, "choose with your eyes closed" instore shopping experience and spatial maps in visually impaired and sighted persons," *Universal Access in the Information Society*, pp. 1-13, 2024.
- [25] "OpenAI — ChatGPT." <https://openai.com/>. [Accessed 03-05-2025].
- [26] S. G. Hart, "Nasa-task load index (nasa-tlx): 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904-908, Sage publications Sage CA: Los Angeles, CA, 2006.