SEEBYSOUND: AN EDGE-AI WEARABLE FOR VISUAL-TO-AUDIO AND HAPTIC ASSISTANCE

ABSTRACT

SeebySound is an assistive wearable system that combines AI-powered smart glasses and haptic feedback gloves to empower blind and low-vision users with real-time environmental awareness. Unlike existing solutions that rely on either audio descriptions or limited obstacle detection, SeebySound leverages edge computing, computer vision, and multimodal feedback to deliver a richer experience. The device uses a high-quality camera and microphone mounted on lightweight glasses, connected to a Raspberry Pi Zero 2 and server infrastructure for AI inference. Visual data is processed through YOLOv8, segmentation models, and Vision-Language Models (VLMs) to generate contextual voice narration, while haptic gloves provide directional guidance for object interaction and navigation. Our MVP demonstrates that this approach not only enhances accessibility for blind users but also opens the door for premium use cases, where SeebySound functions as an intelligent wearable assistant similar to the futuristic interfaces portrayed in science fiction.

1. INTRODUCTION

Globally, more than 285 million people live with some form of visual impairment, and navigating the physical world remains a daily challenge. Traditional assistive devices such as white canes or guide dogs provide limited situational awareness, while smartphone-based solutions often fall short in real-time performance and seamless usability. Recent advancements in AI and wearable computing present an opportunity to fundamentally rethink how blind and low-vision individuals interact with their environment.

SeebySound is designed to address this gap by integrating visual recognition, audio narration, and haptic feedback into a single, user-friendly wearable system. At its core, SeebySound aims to answer two fundamental questions for its users:

“What is happening around me?”

“How do I interact with it safely and precisely?”

While accessibility remains the primary focus, SeebySound is also envisioned as a premium edge-AI product for enthusiasts who seek an “Iron Man–style” AI companion. By offering real-time contextual awareness, seamless app integration, and natural multimodal interaction, SeebySound stands at the intersection of assistive technology and next-generation AI wearables.

2. SYSTEM OVERVIEW

The SeebySound ecosystem consists of two main components:

SEEBYSOUND GLASSES

Hardware

Equipped with a high-resolution camera, microphone, IMU sensors, and edge compute module (Raspberry Pi Zero 2).

AI Processing

Visual input is processed using YOLOv8 for real-time object detection, segmentation networks for fine-grained analysis, and VLMs for generating human-like descriptions.

Audio Feedback

A lightweight TTS system converts model outputs into natural speech, giving users immediate auditory feedback about their surroundings.

Connectivity

On-device edge inference is combined with server-side acceleration for heavier tasks, ensuring both speed and accuracy.

HAPTIC FEEDBACK GLOVES

Haptic Guidance

Integrated with small vibration motors across the fingers and palm, the glove delivers spatial feedback. For example, when locating a water bottle, vibrations guide the user’s hand in the correct direction and adjust intensity as the object is approached.

Object Lock-On

Using computer vision, the system can “lock” onto a person or object. The glove then provides continuous haptic cues that help the user maintain orientation, similar to radar tracking.

Interaction Assistance

By combining object recognition with haptic signals, users can perform precise tasks such as grabbing an object, locating doors, or identifying familiar people.

Together, the glasses and gloves form a closed-loop assistive system: the glasses “see” and narrate the world, while the gloves guide the user’s interaction within it. This multimodal approach makes SeebySound more intuitive and practical than single-channel solutions.

3. LITERATURE REVIEW

The past decade has seen rapid progress in assistive technologies for blind and low-vision individuals. Traditional solutions such as white canes and guide dogs provide physical guidance but lack environmental context. More advanced electronic aids—like ultrasonic canes and RFID-based systems—introduced obstacle detection, yet often suffered from limited range and inability to interpret dynamic environments.

With the rise of AI and computer vision, solutions such as OrCam MyEye and Google Lookout began offering object recognition and scene description through camera-based devices. While these systems marked a significant leap forward, they faced challenges in terms of latency, offline reliability, and user interaction models. Most devices relied solely on audio output, which can overwhelm users in noisy environments or when multitasking.

Parallel research in haptic navigation systems has shown promise. Projects like ultrasonic vests and haptic belts provided directional cues, but lacked the contextual intelligence needed for complex real-world scenarios. Similarly, AI-powered smart glasses prototypes demonstrated real-time scene description but often required cloud connectivity, leading to privacy and latency concerns.

What differentiates SeebySound from these approaches is its multimodal integration:

Glasses provide visual-to-audio narration using deep learning models (YOLOv8 + segmentation + VLM).

Gloves deliver haptic object localization for precise interactions.

The system is designed to balance edge computing (on-device) with cloud offloading for heavy tasks, ensuring low latency, privacy, and robustness.

In essence, SeebySound builds on existing research but extends it by creating a closed-loop interaction system where users not only hear the world but also feel guided within it.

4. METHODOLOGY

The SeebySound system is structured as a multi-component wearable pipeline, integrating hardware and software into a seamless assistive experience.

4.1 HARDWARE ARCHITECTURE

SeebySound Glasses

Camera: High-resolution, wide-angle lens for capturing surroundings in real-time.

Microphone Array: Captures ambient sounds and enables voice commands.

Edge Processor: Raspberry Pi Zero 2 running lightweight inference.

IMU Sensors: Track head orientation to improve spatial mapping.

Audio Output: Bone-conduction speakers for clear narration without blocking ambient sound.

Haptic Feedback Gloves

Microcontroller: ESP32 or Raspberry Pi Pico for real-time motor control.

Vibration Motors: Placed on fingertips and palm to indicate object location and proximity.

Flex Sensors: Capture hand movements for potential gesture-based input.

Bluetooth/Wi-Fi: Communication with glasses and server.

4.2 SOFTWARE ARCHITECTURE

Object Detection: YOLOv8 detects people, objects, and obstacles in real-time.

Segmentation: DeepTorch-based segmentation models allow precise boundary mapping (e.g., distinguishing a chair from a table).

Vision-Language Models (VLMs): Generate contextual scene descriptions, answering “what is happening around me?”

Audio Narration: Text-to-speech engine provides real-time spoken feedback.

Haptic Control Module: Maps detected object coordinates to glove vibration patterns.

4.3 PROCESSING PIPELINE

Input Capture: Camera + mic continuously stream data.

Preprocessing: Frames are resized and normalized for model inference.

Detection & Segmentation: YOLOv8 + segmentation model identify objects and their positions.

Contextual Narration: VLMs generate natural scene descriptions (e.g., “There is a person standing three meters ahead”).

Feedback Distribution:

Audio: Narration sent to bone-conduction speakers.

Haptics: Gloves vibrate in the direction of detected objects, guiding user hands toward targets.

Edge-Cloud Balance: Lightweight tasks (object detection, haptic mapping) run on the Raspberry Pi Zero 2. Heavy tasks (VLM reasoning, advanced NLP) run on remote servers with GPU acceleration.

5. IMPLEMENTATION DETAILS (MVP)

The current prototype of SeebySound was developed using readily available, low-power hardware combined with server-based AI acceleration. The MVP demonstrates the feasibility of real-time object detection, audio narration, and haptic guidance in a portable form factor.

5.1 HARDWARE SETUP

Glasses Module:

Raspberry Pi Zero 2 (main processing unit).

USB high-resolution camera (720p at 30fps).

MEMS microphone module.

Bone-conduction headphones for audio output.

Wi-Fi module for server communication.

Glove Module:

ESP32 microcontroller.

5 vibration motors (distributed across fingers and palm).

Li-Po battery for power supply.

Bluetooth connectivity for real-time response.

5.2 SOFTWARE STACK

Operating System: Raspberry Pi OS Lite (optimized).

Computer Vision: YOLOv8 and segmentation models trained on COCO + custom datasets.

Deep Learning Framework: PyTorch for model deployment, with ONNX runtime for optimization.

Vision-Language Processing: Integration with VLMs hosted on GPU servers (e.g., LLaVA).

TTS Engine: Lightweight models for natural-sounding narration (Kokoro TTS, edge-friendly voices).

Haptic Mapping Logic: Custom middleware translating object coordinates into vibration intensities.

5.3 EDGE–CLOUD INTEGRATION

On-device (RPi Zero 2): Runs YOLOv8 + segmentation for fast detection.

Server-side (GPU): Handles VLM reasoning, natural language responses, and advanced scene analysis.

Communication: Secure WebSocket protocol ensures low-latency bidirectional data flow.

6. RESULTS & EVALUATION

The MVP was tested in controlled environments to assess latency, accuracy, and user experience.

6.1 LATENCY BENCHMARKS

Object Detection (YOLOv8 on RPi Zero 2): Average inference time of ~250ms per frame.

Segmentation: ~320ms per frame.

Audio Narration: ~100ms TTS generation time.

Haptic Response Delay: <50ms (Bluetooth communication + motor actuation).

Overall End-to-End Latency: ~700ms, sufficient for real-time assistance.

6.2 ACCURACY

Object Detection: mAP@50 = 85% on common indoor/outdoor objects (chair, bottle, car, person).

Segmentation: IoU = 78% average, adequate for guiding interactions.

Scene Narration: Human evaluation showed >80% comprehension accuracy of generated descriptions.

6.3 USER TESTING

Preliminary feedback was gathered from blindfolded sighted users and two low-vision participants. Key findings:

Positive: Users reported that the combination of audio + haptic feedback was more intuitive than audio-only systems.

Challenges: Battery life (~3 hours on continuous use) and device weight (especially the camera module) need optimization.

Suggestions: Users requested customizable haptic patterns and more natural speech voices.

6.4 DEMONSTRATION SCENARIOS

Object Retrieval: When searching for a water bottle, the glove provided directional haptic feedback, allowing the user to reach the object with minimal trial and error.

Person-Finding: System successfully locked onto an individual, with glove vibrations adjusting in real-time as the person moved.

Navigation Assistance: Detected obstacles such as chairs and tables, narrating them while providing haptic alerts for proximity.

7. USE CASES

The SeebySound system was designed with accessibility in mind but extends naturally into multiple domains:

NAVIGATION FOR BLIND AND LOW-VISION USERS

Real-time detection of obstacles in indoor and outdoor settings.

Voice narration of the environment (e.g., “There is a staircase ahead to your left”).

Haptic cues for precise object interaction (grabbing bottles, locating chairs, holding handrails).

PERSONAL ASSISTANCE AT HOME

Identifying household items (keys, remote controls, kitchen tools).

Person recognition to differentiate family members.

Integration with voice assistants for calendar reminders, messaging, and calling.

PUBLIC SPACE ORIENTATION

Object detection for vehicles, pedestrians, and traffic signals.

Lock-on tracking to follow a person in a crowd using haptic guidance.

Airport/train station navigation through real-time narration.

PREMIUM AI COMPANION

For technology enthusiasts, SeebySound acts as an AI-powered wearable assistant —similar to Tony Stark’s J.A.R.V.I.S.—integrating with apps, calls, and productivity tools.

AR-style narration of surroundings.

Gesture-based commands via gloves (e.g., double-finger tap = answer call).

8. DISCUSSION

8.1 STRENGTHS

Multimodal Feedback: Unlike audio-only systems, SeebySound merges voice + haptics, improving usability.

Edge-Cloud Hybrid: Balances low-latency inference on-device with the scalability of cloud servers.

Real-Time Performance: End-to-end latency of <1s makes it viable for real-world scenarios.

Scalability: Can be expanded with additional sensors (e.g., LiDAR, thermal cameras).

8.2 LIMITATIONS

Battery Life: Current MVP lasts only 3 hours under continuous use.

Miniaturization: Raspberry Pi Zero 2 and camera module add bulk; future versions require more compact boards.

Learning Curve: First-time users may need training to interpret haptic signals.

Privacy: Continuous video capture raises ethical concerns about bystander consent.

9. ETHICAL CONSIDERATIONS

The integration of AI-powered vision into daily life introduces important ethical and social dimensions:

Data Privacy: Continuous video/audio recording must ensure local processing where possible. Cloud servers should encrypt all transmitted data.

Bias in AI Models: Object detection and person recognition models may have dataset biases. Ensuring fairness and inclusivity is crucial.

Accessibility Equity: While marketed as a premium wearable, affordability and access for visually impaired communities must remain a priority.

User Autonomy: SeebySound should augment, not replace, the independence of blind and low-vision individuals.

10. FUTURE WORK

To move beyond the MVP, the following directions are planned:

MINIATURIZATION

Integration into lightweight smart glasses using custom AI edge chips (e.g., Google Coral TPU, NVIDIA Jetson Nano).

ADVANCED HAPTICS

Expansion of glove feedback to encode braille-like patterns.

Development of full-hand haptic cues for complex interactions.

5G AND EDGE CLOUD INTEGRATION

Offloading heavy AI tasks to nearby edge servers for ultra-low latency (<200ms).

AR OVERLAY

Adding a heads-up display (HUD) for low-vision users who retain partial sight.

CONTEXTUAL AWARENESS WITH LLMS

Incorporating conversational agents (GPT/Gemini-like) for dialogue-based assistance:

User: “Where did I keep my phone?”

System: “It’s on the table near the sofa, to your right.”

11. CONCLUSION

SeebySound represents a significant step forward in the design of intelligent assistive wearables. By combining computer vision, edge AI, and multimodal feedback into a single ecosystem, it bridges the gap between accessibility technology and next-generation AI companions.

Our MVP demonstrates the feasibility of this approach, showing that blind and low-vision users can benefit from real-time narration and haptic guidance for navigation and interaction. At the same time, the system introduces new possibilities for premium wearable AI assistants, blurring the line between assistive technology and consumer innovation.

Moving forward, optimization in miniaturization, battery efficiency, and advanced haptic feedback will be key. With ethical deployment and accessibility as core values, SeebySound has the potential to redefine how humans interact with both the physical and digital world—making it not just an assistive device