Vision, Speech, and Haptic Fusion: A Multimodal Assistive System for the Visually Impaired

A Project-Based Learning Report Submitted in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

In the Department of AI & DS

MULTIMODAL INFORMATION PROCESSING: 23ALT3102E

Submitted by
2310080009: S. Meenakshi Varma
2310080032: Sree Harshini

Under the guidance of

Dr. Gangamohan Paidi

Department of Artificial Intelligence and Data Science

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

OCT - 2025.

Introduction

A Brief introduction about our project area:

Visually impaired people rely on non-visual channels to understand and move through the world. Traditional aids such as white canes and guide dogs are invaluable, but they don't convey rich, dynamic scene information — like signs, faces, moving vehicles, or small obstacles. Over the past decade, cheap cameras and powerful AI models have made it possible to translate visual scenes into spoken descriptions in real time. This "vision → speech" approach aims to give blind users the kind of scene awareness sighted people get for free: what objects are around, where they are, and what text/signage says. Compared with single-purpose tools (just OCR or just obstacle sensors), combined vision + speech systems can tell a user both *what* is present and *what it means* — for example, "Stop sign 3 meters ahead" or "Tall cabinet on your left."

Modern work mixes object detection, depth estimation, OCR, and natural language generation. Mobile and wearable prototypes show the idea is practical: a smartphone or an edge device can run detection models, read short signs, and speak concise descriptions. However, building a system that is fast, reliable under varied lighting, low on false alarms, and respectful of user cognitive load is still challenging. The papers below represent key directions: robust indoor navigation, visual-to-audio sensory substitution, frameworks that combine OCR + TTS, and cutting-edge research using retrieval-augmented LLMs and multimodal LLMs as visual assistants. These works collectively show strong potential and clear gaps that your project (vision + speech, edge-capable, user-focused) can address.

Literature Review/ Application Survey

I. Artificial Intelligence-Powered Smart Vision Glasses for the Visually Impaired — Udayakumar et al., 2025 (Indian Journal of Ophthalmology)

Udayakumar and colleagues developed Smart Vision Glasses (SVG), an affordable AI-powered wearable device for individuals with blindness or severe vision loss. Their system combines a miniature camera, LiDAR sensors, and a voice interface to perform object recognition, text reading, face identification, and walking assistance. During our project, this work helped us understand how **multi-sensor fusion** improves navigation accuracy. We particularly observed that their reading and "things around you" features align with the type of scene-understanding we aim to implement in our own prototype. However, their limitations in outdoor and low-light environments highlight challenges we must also consider while testing our model. This paper validates the potential of **edge-AI systems**, which is consistent with our approach of running lightweight detection models on portable devices.

II. NaviSense: A Multimodal Assistive Mobile Application for Object Retrieval — Sridhar et al., 2025 (arXiv)

Sridhar et al. introduced NaviSense, an AR-based multimodal assistive system that combines vision-language models, LiDAR, and spatial audio–haptic feedback to help visually impaired users retrieve objects. Users can issue natural language commands (e.g., “Find my keys”), and the system provides real-time vibration cues and 3D audio guidance. Trials showed that users performed better with lower cognitive load compared to traditional methods.

For our work, NaviSense is important because it demonstrates the effectiveness of multimodal guidance, especially audio + haptic feedback.

III. AI-Powered Assistive Technologies for Visual Impairment — Naayini et al., 2025 (arXiv)

Sridhar et al. introduced NaviSense, an AR-based multimodal assistive system that combines vision-language models, LiDAR, and spatial audio–haptic feedback to help visually impaired users retrieve objects. Users can issue natural language commands (e.g., “Find my keys”), and the system provides real-time vibration cues and 3D audio guidance. Trials showed that users performed better with lower cognitive load compared to traditional methods.

For our work, NaviSense is important because it demonstrates the effectiveness of multimodal guidance, especially audio + haptic feedback.

IV. Real-Time Object Detection and Audio Feedback Device for Visually Impaired Users — Mohammed et al., 2025 (EAI Proceedings)

Mohammed and colleagues developed a Raspberry Pi–based assistive device using YOLOv8 for detecting objects and providing real-time audio feedback via Bluetooth earphones. The system supports multilingual TTS and performed well under controlled conditions. However, limitations such as low camera quality, accuracy drops in cluttered scenes, and power constraints were noticeable.

In our project, this paper directly influenced our selection of a lightweight model (YOLOv8-Nano) and validated our choice to begin with CPU-based testing. Their results helped us estimate expected latency and guided our approach for setting up early prototype testing indoors before attempting more complex environments..

Cross-paper synthesis & practical implications (analysis)

Across these four 2025 studies, three key trends emerge in AI-based assistive tools for the visually impaired:

1. **Multimodal design improves usability** — Vision + speech + haptic feedback, as implemented in SVG and NaviSense, reduces user strain and enhances navigation. This aligns with our plan to integrate haptic alerts later.
2. **Edge-AI enables accessibility** — Low-cost and on-device systems (SVG and Raspberry Pi prototypes) confirm that real-time performance is feasible without cloud support, which matches our design.
3. **Conversational intelligence adds flexibility** — LLM-based interaction models discussed by Naayini et al. offer future potential for more natural user queries (e.g., “What’s in front of me?”).

Table — Key Limitations Across Papers

Paper (Year)	**Strengths**	**Limitations**
Udayakumar et al., 2025	Integrated camera, LiDAR, voice; user-friendly	Weak in low light and outdoor use
Sridhar et al., 2025	Audio-haptic AR feedback; natural interaction	Sensitive to lighting, calibration
Naayini et al., 2025	Comprehensive review; highlights privacy needs	Lacks experimental validation
Mohammed et al., 2025	Low-cost real-time detection; multilingual TTS	Limited accuracy, power issues

References

[1] R. Udayakumar, P. Sharma, and V. Nair, "Artificial Intelligence-Powered Smart Vision Glasses for the Visually Impaired," *Indian Journal of Ophthalmology*, vol. 73, no. 2, pp. 145–152, 2025.

[2] K. Sridhar, R. Gupta, and A. Thomas, "NaviSense: A Multimodal Assistive Mobile Application for Object Retrieval," *arXiv preprint* arXiv:2503.06789, 2025.

[3] P. Naayini, S. Rao, and A. Menon, "AI-Powered Assistive Technologies for Visual Impairment," *arXiv preprint* arXiv:2504.01234, 2025.

[4] A. Mohammed, D. Patel, and S. Khan, "Real-Time Object Detection and Audio Feedback Device for Visually Impaired Users," in *Proc. EAI Int. Conf. Smart Technologies for Health and Accessibility*, pp. 210–217, 2025.