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

(13 size) A Project Based Learning Report Submitted in partial fulfilment 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

SEP - 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

1) Vision-Based Mobile Indoor Assistive Navigation Aid — Li et al., 2018 (ACS/PMC)

Li and colleagues developed a holistic mobile system for indoor navigation that fuses RGB(D) sensing, semantic mapping, and speech guidance to help blind users travel indoors. The project creates semantic maps from building models, detects dynamic obstacles, and updates routes in real time. Their prototype emphasizes reliable path planning: when the system detects a moving obstacle or an unexpected blockage, it recalculates and gives spoken directions. Field tests showed promise for everyday indoor use, especially in structured environments like office buildings and campuses. This work is strong on mapping and planning, and it demonstrates how speech can be used for route-level guidance. However, it depends on good indoor maps and RGB-D sensors that are not always available on simple smartphones. [1] (B. Li, Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People, 2018)

II. Navigation aid by visual-to-auditory sensory substitution (Pilot Study) — Neugebauer et al., 2020 (PLOS ONE)

This study explored an augmented-reality smartphone app that converts 3D visual cues into audio signals (spatialized sounds) — a sensory substitution approach. Rather than narrating everything, the system translates height, distance, and obstacle shapes into auditory patterns that users can learn to interpret. The pilot showed that blind users could improve navigation and recognition tasks after short training. The contribution is important because it highlights alternatives to verbose narration: spatial audio (and minimal spoken prompts) can be more intuitive for some tasks. But sensory substitution requires user learning and can be less straightforward for reading textual information (signs) compared to direct OCR→speech. (A. Neugebauer, 2020)

III. From Vision to Voice: A Multi-Modal Assistive Framework — Bhat et al., 2025 (IEEE Access)

Bhat et al. present a practical framework that links computer vision (object detection + OCR), translation, and text-to-speech to support physically impaired users. The system pipeline extracts scene text, recognizes objects, optionally translates text across languages, and produces speech output with adjustable verbosity. Their experiments show reliable reading of scene text and useful object announcements; the framework emphasizes modularity so components can be swapped or upgraded. This paper is valuable because it is closest to your proposed stack: camera → (detection, OCR) → NLG → TTS, and it discusses latency, multi-language TTS, and edge-friendly components. The main limitations they report are sensitivity to low-quality images and trade-offs between verbosity and user overload. (S. Bhat, 2025)

IV. SeeSay: Assistive Device Using Retrieval-Augmented Generation (RAG) — Yu, 2024 (arXiv)

SeeSay is a modern approach that combines continuous visual logging with retrieval-augmented generation. It doesn’t just describe the current frame but can retrieve prior visual context (e.g., “You left your keys on the kitchen table yesterday”) and answer spoken queries using an LLM augmented with memory. SeeSay demonstrates two advantages: richer, context-aware answers and natural conversational queries. For visually impaired users, this means not only momentary descriptions but contextual reminders and question answering about their environment. However, SeeSay relies on large models and memory storage, raising privacy and latency concerns in on-device setups. (Yu, 2024)

V. Evaluating Multimodal Language Models as Visual Assistants — Karamolegkou et al., 2025 (ACL / arXiv)

This recent evaluation systematically tests multimodal LLMs (MLLMs) on tasks relevant to blind users: scene description, instruction following, text reading, and specialized tasks (like Optical Braille Recognition). The paper finds that while MLLMs are promising, they suffer from hallucinations, cultural/contextual errors, and weaknesses on small textual elements (fine OCR-like tasks). User studies reveal trust and safety concerns when users rely solely on MLLMs for visual interpretation. The paper argues for careful grounding, task-specific pipelines, and fallback verification for critical outputs (A. Karamolegkou, 2025).

Cross-paper synthesis & practical implications (analysis)

Across these works, we see three consistent themes:

1. **Modularity wins.** Systems that separate detection/OCR from higher-level language modules are easier to debug, optimize for edge devices, and safer for critical outputs (e.g., reading numeric text). Papers by Li and Bhat show modular pipelines work well in practice.
2. **Speech must be concise and context-aware.** Verbose narration overwhelms users. Sensory-substitution and spatial audio papers (Neugebauer) suggest alternatives, but for reading text and naming nearby objects, short, crisp phrases are preferred.
3. **LLMs add value but need checks.** RAG and MLLMs (SeeSay, ACL paper) allow contextual answers and conversational queries, but hallucination and latency are real problems. Best practice is to use deterministic vision modules for low-level perception and reserve LLM outputs for higher-level summaries or user-initiated queries.

Table — Key limitations across papers

Paper (year)	**Main strengths**	**Main limitations**
Li et al., Vision-Based Indoor (2018)	Robust semantic mapping; good path planning	Needs RGB-D / maps; heavy for simple smartphones.
Neugebauer et al., Visual→Auditory (2020)	Intuitive spatial audio; low narrative load	Requires user training; weak for textual reading.
Bhat et al., From Vision to Voice (2025)	End-to-end OCR→TTS; modular	Sensitive to poor image quality; verbosity control needed.
Yu, SeeSay (2024)	Contextual memory + RAG; conversational queries	Privacy, latency, and heavy compute; not fully edge-capable.
Karamolegkou et al., MLLM Eval (2025)	Systematic analysis of MLLMs; user study insights	Hallucinations; poor fine-text handling; trust/safety concerns.

References

1. B. Li, R. T. (2018). Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People. *Sensors*, 1–22.

2. A. Neugebauer, A. K. (2020). Navigation aid for blind persons by visual-to-auditory sensory substitution: A pilot study. *PLOS ONE*, 1–15.

3. S. Bhat, P. B. (2025). From Vision to Voice: A Multi-Modal Assistive Framework for the Physically Impaired. *IEEE Access*, 24311–24320.

4. Yu, M. (2024). *SeeSay: An Assistive Device for the Visually Impaired Using Retrieval Augmented Generation.* Ithaca, NY, USA: Cornell University Library.

5. A. Karamolegkou, K. B. (2025). *Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users.* Vienna, Austria: Association for Computational Linguistics.