International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 3, May – June -2025, ISSN (Online): 2395-566X

EchoSense: A Survey on Dual-Modality Assistive Systems for the Blind and Deaf

Dr. Nandini C, Jayant Rajendra Habbu, Midhun Manoj, Moulya R Hegde, Prerana A J

Dayananda Sagar Academy of Technology and Management

Abstract - In recent years, assistive technologies for visually and hearing-impaired individuals have predominantly focused on single-modality solutions, limiting their effectiveness for users with dual sensory impairments. This paper proposes EchoSense, a novel mobile-based assistive system integrating real-time visual scene interpretation and personalized sound detection enhanced with emotion recognition. Through a comprehensive survey of recent literature encompassing scene captioning, sound classification, emotion detection, and explainable AI, key limitations such as lack of multi-modality, minimal personalization, and dependence on cloud-based processing are identified. EchoSense addresses these gaps by offering an on-device, privacy-preserving, and offline-capable platform that delivers adaptive multimodal feedback tailored to blind, deaf, and deaf-blind users. The paper details the system's conceptual framework, design considerations, and methodology, laying the groundwork for future implementation and evaluation.

Keywords - Assistive Technology, Sound Recognition, Scene Captioning, Emotion Detection, Dual-Modality

I. INTRODUCTION

Accessibility remains a critical challenge in technology, particularly for individuals with sensory impairments such as blindness and deafness. People who are blind or have low vision face difficulties with spatial awareness and interpreting visual contexts, while deaf or hard-of-hearing individuals struggle with auditory cues and emotional tone perception. Existing assistive tools often focus on a single sensory domain, such as object detection for the blind or sound alerts for the deaf and typically operate independently without personalization. Moreover, many rely on cloud-based processing, resulting in latency issues and raising privacy concerns. To address these limitations, this paper proposes EchoSense, a dual-mode mobile application that integrates object and scene recognition for blind users with environmental sound detection and emotion analysis for deaf users.

By unifying these functionalities within a customizable and context-aware system, EchoSense aims to enhance user independence and communication. The system prioritizes offline functionality through on-device machine learning to improve privacy and reduce latency. Additionally, it adapts feedback formats— including audio, visual, and haptic modes—based on the user's sensory profile, accommodating users with single or dual impairments. This paper presents the motivation, related research, and proposed architecture for EchoSense, laying the foundation for future development and deployment.

II. LITERATURE SURVEY

Investigating Use Cases of AI-Powered Scene Description
Ricardo Gonzalez, Cynthia Bennett, Shiri Azenkot, Jazmin Collins, ACM, 2024.
This paper explores how blind and low-vision users utilize AI-powered scene description tools to identify objects, settings, and avoid obstacles. The system studied relies on a static image-based interface powered by the Azure cloud API, where users capture pictures and receive descriptive captions. A user study with 16 participants showed that while helpful for object identification and surroundings, the tool's limited interactivity and cloud dependence hindered user satisfaction.
The primary gap lies in the lack of real-time scene analysis and cloud dependency, which causes latency and privacy issues. Moreover, the system does not consider emotional or situational nuances that are important for blind users to fully understand complex environments. These limitations highlight the need for offline, on-device systems capable of richer context- aware feedback [1].

A Personalizable Mobile Sound Detector App Design
Danielle Bragg, Nicholas Huynh, Richard E. Ladner, ACM,2016.
This study introduces a customizable mobile app for deaf and hard-of-hearing users, enabling them to record important environmental sounds such as alarms or door knocks and receive alerts via vibration and visual cues. The app uses

MFCC features and Gaussian Mixture Models (GMM) for personalized sound classification, fostering user independence.

However, the app lacks emotional sound analysis and does not support multi-modal accessibility for blind or dual-impaired users. Additionally, it faces challenges in noisy environments, leading to false or missed alerts. Although the personalization feature is strong, the absence of integrated feedback modalities and deeper contextual understanding limits its usability [2].

Accuracy and Usability of Smartphone-Based Distance Estimation Approaches Giles Hamilton-Fletcher et al., IEEE, 2024.
This paper evaluates five mobile-based short-range distance estimation methods on the iPhone 13 Pro, including CoreML, LiDAR, IR, and ARKit cameras. ARKit and LiDAR showed high precision, while CoreML

offered energy-efficient deployment. The study validates the feasibility of real-time spatial awareness on commercial devices.
Despite its technical depth, the study is limited to raw distance estimation and does not explore feedback mechanisms or user-specific contextualization. The lack of auditory or haptic feedback, emotion recognition, or scene captioning restricts its assistive applications. There remains an opportunity to integrate these spatial techniques with other modalities to enhance assistive technology [3].

Explainable AI in Deep Learning: Advancements, Applications and Challenges Md. Tanzib Hosain et al., ScienceDirect, 2024. This paper reviews key interpretability techniques for deep learning, such as saliency maps, attention mechanisms, and explainability frameworks like LIME and SHAP, which increase user trust through transparent predictions. Though the focus is on healthcare and industrial applications, these techniques lay the groundwork for applying explainable AI to assistive technologies.

The paper lacks direct application to users with disabilities but highlights the importance of transparency in decision-making. This is critical for blind or deaf users to trust the outputs they receive, calling for research combining explainable outputs— like verbal summaries or visual explanations—with real-time sensory analysis [4].

Understanding Emerging Obfuscation Technologies in Visual Description Services ACM, 2022.
This research explores privacy-preserving methods in AI-generated visual captions for blind users, discussing filters that omit or generalize sensitive content to protect privacy and enhance trust. The study combines technical proposals with user evaluations, emphasizing privacy-aware captioning's impact on user confidence.
However, it lacks mobile deployment or real-time integration and does not address auditory cues,

focusing only on visual modality. While the privacy aspect is valuable, future work needs to integrate these techniques with dynamic, customizable feedback across multiple sensory channels [5].

III. GAP MITIGATION

The first paper [1] identifies key gaps such as reliance on static images, cloud dependency causing latency and privacy concerns, and the absence of emotional or situational context in scene descriptions. EchoSense addresses these issues by implementing a real-time, on-device scene captioning system with lightweight models optimized for mobile deployment (e.g., YOLOv8, BLIP-2 with TensorFlow Lite). This approach reduces latency, preserves user privacy, and enhances feedback by including contextual environmental cues and emotional awareness, thereby improving the dynamic user experience.

The second study [2] offers a personalized sound detector app but lacks emotional context and is unsuitable for blind or dual-impaired users due to its reliance on visual and haptic feedback alone. It also struggles in noisy environments. EchoSense overcomes these limitations by integrating speech emotion recognition through CNN models trained on prosodic features and providing multimodal feedback: audio for blind users, vibration for deaf users, and haptic Morse code-like signals for deaf- blind users. This also includes improved robustness against noise, enhancing detection accuracy and accessibility.

The third paper [3] focuses exclusively on distance estimation without transforming the data into user- friendly feedback or integrating it with other assistive functions. EchoSense incorporates this distance estimation to augment object detection and scene narration, using proximity data to prioritize alerts and deliver personalized audio or haptic signals, thereby integrating spatial awareness into a comprehensive assistive experience.

The fourth paper [4] is theoretical and lacks practical application in assistive contexts or user-facing transparency tools. EchoSense aims to embed explainability within its feedback by converting AI decisions into natural, understandable language (e.g., "a person nearby is smiling") or intuitive haptic patterns, which enhances user trust and interpretability.

Finally, the fifth paper [5] addresses privacy in visual description services but is limited by non-real-time processing and absence of multimodal feedback. EchoSense prioritizes privacy through fully local processing and enables user-configurable filters to omit sensitive information from captions, integrating privacy considerations with real-time dual-modality feedback to balance informativeness and discretion.

Criteria / Feature	Literature Insights	EchoSense Implementation
Scene Interpretation	Static image-based [1]	Real-time object C scene recognition (YOLOv8 + BLIP-2)
Processing Mode	Cloud-based, dependent on connectivity [1][5]	Fully on-device (TensorFlow Lite)
Emotional Context Recognition	Largely absent [1][2]	Emotion detection via CNN (RAVDESS-based prosodic analysis)
Personalization	Sound customization only [2]	Multi-modal configuration for blind, deaf, and deaf-blind users
Distance Estimation	Raw data, no interpretation [3]	Integrated into user-centric narration or vibration alerts
Explainability	Lacking or theoretical [4]	Natural language or haptic feedback for AI usage
Privacy s Data Ownership	Sensitive info processed via cloud [1][5]	Full local storage, privacy-preserving filters
Feedback Modes	Visual or audio only [1][2]	Adaptive: audio, visual, haptic or combination

Table 1: Comparative Analysis of Literature vs. Proposed System

IV. PROPOSED IMPLEMENTATION

The proposed EchoSense system offers a dual-mode interface, enabling users to choose between assistive modes based on their sensory requirements: Blind Mode or Deaf Mode. Based on the selection, the
system dynamically adapts its input processing and feedback mechanisms to suit the user's needs.

Blind Mode: When the user selects blind assistance, the system activates the device's camera and processes the live feed using on-device AI models. YOLOv8 is used for object detection, and BLIP-2 generates contextual scene captions. These captions are then converted into speech using a text-to-speech engine, providing auditory descriptions of the environment to assist the user in navigation and object awareness.

Deaf Mode: If deaf assistance is selected, the system uses the microphone to capture ambient sounds. YamNet is employed for sound classification (e.g., doorbells, alarms), while Whisper is used for real-time speech transcription. Alerts for critical sounds trigger vibration feedback, whereas speech and environmental cues are shown as on-screen text. Vibration is reserved for urgent events only, enhancing responsiveness without overwhelming the user.

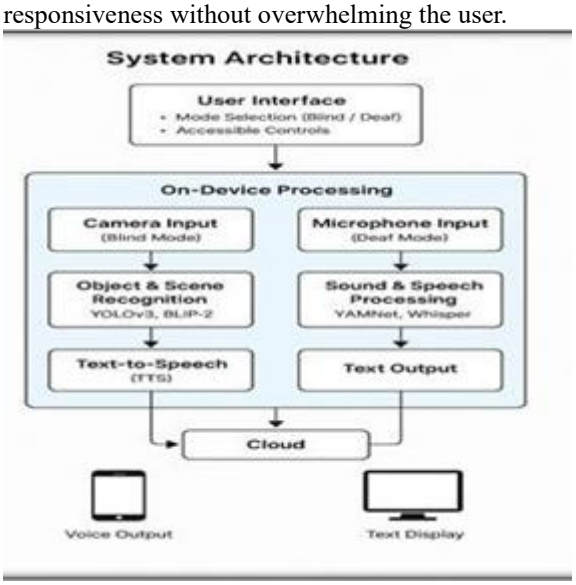

Figure1- Proposed System Architecture

In both modes, the interface ensures accessibility by offering visual, auditory, or haptic outputs, depending on the user's profile. All AI processing is executed on- device to maintain privacy, ensure low-latency performance, and support operation in offline or low- connectivity environments.

V. CONCLUSION

EchoSense proposes a novel, offline-capable, dual- mode assistive mobile application aimed at improving spatial, auditory, and emotional awareness for users with visual and/or auditory impairments. By synthesizing findings from prior works in object detection, scene captioning, sound classification, and emotion recognition, the proposed system integrates multiple sensory processing pathways into a single, user-adaptive platform. The application dynamically adjusts its interface and response mechanisms based on the selected user profile—blind, deaf, or dual-impaired—thereby delivering personalized multimodal feedback via speech, text, or vibration.

The architectural design emphasizes on-device machine learning using lightweight, mobile- optimized models such as YOLOv8, BLIP-2, Whisper, and YamNet, thereby ensuring minimal latency, offline functionality, and enhanced privacy. The system also includes customizable feedback settings and a local storage mechanism, enabling context-aware operation without dependence on cloud- based services. This approach directly addresses limitations observed in existing assistive technologies, including cloud reliance, single- modality focus, and lack of personalization. Future work will involve prototyping the system, conducting usability studies with real users, and iterating on the model performance and interface features based on user feedback. Through this approach, EchoSense aspires to serve as a unified, privacy-conscious, and practical solution for enhancing independence and situational understanding among individuals with sensory disabilities.

REFERENCES

1. R. Gonzalez, C. Bennett, S. Azenkot, and J. Collins, “Investigating Use Cases of AI-Powered Scene Description,” Proc. ACM, 2024.
2. G. Hamilton-Fletcher, M. Liu, D. Sheng, C. Feng, T.
3. E. Hudson, J.-R. Rizzo, and K. C. Chan, “Accuracy and Usability of Smartphone-Based Distance Estimation Approaches for Visual Assistive Technology Development,” IEEE, 2024.
4. D. Bragg, N. Huynh, and R. E. Ladner, “A Personalizable Mobile Sound Detector App Design,” Proc. ACM, 2016.
5. M. T. Hosain, J. R. Jim, M. F. Mridha, and M. M. Kabir, “Explainable AI Approaches in Deep Learning: Advancements, Applications and Challenges,” ScienceDirect, 2024.
6. R. Gipiškis, C.-W. Tsai, and O. Kurasova, “Explainable AI (XAI) in Image Segmentation in Medicine and Industry,” ScienceDirect, 2024.
7. Rahaf Alharbi, Robin N. Brewer, Sarita Schoenebeck “Understanding Emerging Obfuscation Technologies in Visual Description Services for Blind and Low Vision People,” Proc. ACM, 2022.
8. M. Safwan, “Vision Beyond Sight: The Role of Computer,” Proc. ACM, 2016.
9. T. A. Qureshi, M. Rajbhar, Y. Pisat, and V. Bhosale, “AI Based App for Blind People,” International Research Journal of Engineering and Technology (IRJET), vol. 8, no. 6, pp. [page numbers], 2021.