International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 10 October-2026 | Impact Factor: 3.5



Computer Vision and Voice-Based AI Technologies for Assistive Systems

Author 1
Prof. Vijaya Jadhav
Department of Computer Applications, ASM IBMR College, Pune, Maharashtra, India
vijayaj1979@gmail.com

Author 2
Shivam Kate
Department of Computer Applications, ASM IBMR College, Pune, Maharashtra, India
shivamkate1604@gmail.com

Author 3
Mohit Wankhede
Department of Computer Applications, ASM IBMR College, Pune, Maharashtra, India
Wankhede4755@gmail.com

Author 4
Manthan Khonde
Department of Computer Applications, ASM IBMR College, Pune, Maharashtra, India
manthankonde@gmail.com

How to Cite this Article:

Kate, S., Wankhede, M. & Khonde, M. (2026). Computer Vision and Voice-Based AI Technologies for Assistive Systems. International Journal of Creative and Open Research in Engineering and Management, 02(10), 1-9. https://doi.org/10.55041/ijcope.v2i10.117

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i10.117

Abstract: Assistive technologies increasingly apply artificial intelligence to support access to physical environments, visual information, communication, and human-computer interaction. This review synthesizes eight selected studies covering indoor navigation, interactive object search, Indian Sign Language recognition, gesture-to-text interaction, real-time hand tracking, vision-language navigation, multilingual assistance, and wake-word detection. The studies are organized into two broad themes: computer vision-based assistance and voice-based AI. A narrative comparison examines the tasks addressed, technical approaches, input and output modalities, evaluation designs, reported findings, and limitations. The literature illustrates a movement from infrastructure-assisted guidance and constrained recognition toward open-vocabulary object search and vision-language interaction, while also showing the continuing importance of efficient perception and speech-activation components. The reviewed studies use heterogeneous evidence, including small user studies, image-classification experiments, device-level measurements, surveys, and prototype demonstrations; their results therefore should not be ranked through one common accuracy score. Recurring needs include broader real-world validation, generalization across users and environments, sustained contextual interaction, accessible user feedback, efficient deployment, and more consistent evaluation reporting. This paper is a synthesis of existing published work and does not propose a new system or report project-specific experiments.

Keywords: Assistive Technology; Computer Vision; Voice-Based AI; Accessibility; Human-Computer Interaction; Multimodal AI

I. INTRODUCTION

A. Background and Context

Assistive technology encompasses tools and services that help people access information, communicate, navigate environments, and participate in daily activities. Artificial intelligence has broadened this area by enabling visual recognition, language processing, speech interaction, and adaptable interfaces. Yet these capabilities respond to different needs. A person who is blind may require spatial information about a corridor or a target object, while a person who is deaf or hard of hearing may benefit from sign-language recognition or text-based communication. Voice interfaces may support hands-free interaction, but they are not suitable in every environment or for every user. For this reason, accessibility should be understood as a collection of context-dependent tasks rather than one universal technical problem.

Computer vision can convert images or video frames into representations of objects, gestures, landmarks, or scenes. In assistive applications, this may support hand tracking, sign-language recognition, object search, and navigation. The reviewed literature shows several distinct approaches. Ivanov describes indoor guidance supported by NFC and passive RFID infrastructure [1]. ObjectFinder combines open-vocabulary object detection with a multimodal language model to help users search for objects through flexible requests [2]. Rokade and Jadav examine static Indian Sign Language (ISL) finger spelling [3], and Maben, Shetty, and Salian describe a web-oriented gesture-to-text prototype [7]. MediaPipe Hands provides a reusable hand-tracking component that can operate in real time on mobile devices [4]. Li et al. investigate vision-language techniques for producing navigation instructions [6].

Voice-oriented technologies address another part of accessible interaction. A multilingual assistance platform integrates speech recognition, natural language processing, text-to-speech, and computer vision for event-related guidance [5]. HEiMDaL focuses on wake-word detection and localization, an enabling stage that can identify when a spoken command begins without requiring a full speech-understanding model to run continuously [8]. These two studies operate at different levels: one addresses an application combining several services, while the other focuses on a compact audio-processing component.

The diversity of these tasks creates a challenge for comparison. Recognition accuracy from a fixed sign dataset cannot be interpreted as equivalent to route completion, object-search usability, speech activation, or multilingual service coverage. Likewise, a component benchmark does not by itself demonstrate that an end-to-end assistive application is reliable,



understandable, or useful in daily life. A useful review must therefore preserve each study's task and evaluation context while drawing cautious cross-study conclusions.

B. Scope and Objectives

This paper reviews eight selected studies that represent complementary computer-vision and voice-based AI capabilities [1]–[8]. Its aim is not to claim that these studies exhaust the assistive-technology literature, nor to propose a new architecture. Instead, it organizes the selected publications into functional themes and examines where their methods and evidence overlap or differ.

The review has four objectives: (1) to describe the principal tasks and approaches represented in the selected literature; (2) to compare their input and output modalities, evaluation designs, and reported results; (3) to identify recurring limitations that appear across more than one technical direction; and (4) to outline research priorities that follow from the reviewed evidence. The focus is on computer vision and voice-based AI as enabling technologies, with accessibility needs providing the application context.

The scope is intentionally broad enough to include both assistive applications and components that could support such applications. This distinction matters because a hand-tracking pipeline, a wake-word detector, and a navigation interface have different responsibilities and should be assessed at different levels. The paper therefore distinguishes component-level performance from user-level outcomes throughout the discussion.

II. METHODOLOGY

A. Review Design and Study Selection

The paper uses a narrative, thematic review approach. Eight studies were selected because they cover two related technological themes and represent different positions in an assistive interaction pipeline: sensing and recognition, object or gesture interpretation, navigation and guidance, multilingual interaction, and voice activation. The set includes conference proceedings, journal articles, and preprint literature. The review does not claim a systematic search of every relevant database, nor does it apply meta-analysis or a formal risk-of-bias scoring instrument.

The selected studies were included when they addressed at least one of the review's focal functions and reported a method, implementation, or evaluation that could be described at the literature level. The selected set spans infrastructure-assisted navigation [1], open-vocabulary object search [2], static ISL classification [3], real-time hand tracking [4], multilingual assistance [5], vision-language navigation [6], gesture-to-text interaction [7], and wake-word detection [8]. This purposeful selection provides technical variety, but it also limits how broadly the conclusions can be generalized.

The review treats findings according to the type of source and its evidence. For example, a user study is described as user-level evidence; an image benchmark is described as dataset-level evidence; and a component paper is described in terms of its measured technical properties. Results from a preprint are identified as claims reported by that work rather than as settled evidence. Numerical results are included only when the cited paper reports them and are not recalculated or combined across incompatible tasks.

B. Data Extraction and Comparative Approach

For each study, the synthesis records the application task, target context, principal method, input and output modality, evaluation approach, main reported findings, and visible limitations or future needs. The comparison is organized around five dimensions: task effectiveness, interaction quality, computational or infrastructure requirements, robustness and generalization, and strength of evaluation evidence. These dimensions are descriptive rather than a weighted ranking system.

The comparison avoids a single overall score because the studies do not share a dataset, target function, participant group, or evaluation protocol. For example, the completion time of an indoor route is not interchangeable with the accuracy of sign classification; wake-word detection metrics cannot establish the usefulness of a multilingual service; and performance on annotated navigation descriptions does not establish safety in unfamiliar buildings. Where available, specific metrics are presented with their original task context.

Cross-study gaps are treated as synthesis-level observations. A limitation is attributed to an individual study only when it is reported in that work or clearly follows from its documented scope, such as fixed sign sets or dependence on installed RFID tags. Wider recommendations, including larger field studies and more consistent reporting, are presented as future research needs rather than as experimental findings.

III. LITERATURE REVIEW

A. Computer Vision-based Assistive Technologies

1) Indoor Navigation

Ivanov presents an indoor navigation system for people with visual impairment that uses mobile terminals, NFC, passive RFID tags, and a Java application [1]. Information about rooms and points of interest is associated with tags placed in the environment, allowing a user's device to obtain location-related information and deliver voice-enabled guidance. This design addresses a central indoor-navigation difficulty: satellite navigation is often unreliable inside buildings, while indoor spaces contain corridors, rooms, junctions, and destinations that may not be represented in a mobile device's usual map.

The study was evaluated in a hospital environment with eight blind participants. All participants reached the target reference point. The reported mean completion time was 136 seconds, with the best and worst times of 106 and 180 seconds, respectively [1]. These outcomes provide evidence that prepared infrastructure combined with mobile guidance can support a bounded indoor-navigation task. The participant-level result is relevant because reaching a destination is a user outcome, not only a model metric.



The approach also illustrates a deployment trade-off. Tag-based infrastructure can provide explicit location anchors and may reduce uncertainty in places where positioning from visual cues alone is difficult. However, the building must be surveyed and prepared, tags must remain readable and associated with accurate information, and changes in room use or corridor layout may require maintenance. A successful hospital test therefore does not by itself demonstrate easy deployment in a large variety of buildings. Future studies would benefit from reporting installation effort, failures caused by missing or damaged tags, performance at unfamiliar sites, and user experience during route deviations.

Indoor navigation also needs more than a sequence of correct location estimates. Guidance must be understandable, timely, and aligned with a person's movement. A destination may be temporarily inaccessible, a corridor may be blocked, or a user may need to pause and request clarification. The reviewed study establishes a foundation for navigation using environmental anchors, while its limited setting suggests that robustness and route adaptation remain important questions for the broader field.

2) Interactive Object Search

ObjectFinder addresses the problem of finding requested objects for people who are blind [2]. It combines an open-vocabulary object detector with a multimodal large language model in an interactive pipeline. Compared with a system limited to a closed list of object labels, open-vocabulary search is intended to accommodate more flexible expressions of user intent. Users may specify an item in their own words, and the system can provide spatial cues such as distance or direction, as well as additional scene context when relevant to the request.

The 2026 journal version describes an evaluation involving eight participants with blindness and reports that participants valued flexible target specification, egocentric and allocentric spatial cues, and context-sensitive descriptions [2]. The paper reports 79.8% direction accuracy for route planning and 82.5% spatial accuracy for scene description in the evaluated setup [2]. These figures relate to different component tasks and should not be interpreted as a single end-to-end success rate. Their significance lies in showing that spatial language and object discovery can be evaluated separately as interacting capabilities.

Interactive search introduces challenges that are less prominent in fixed-category recognition. A requested object may be partially hidden, visually similar to nearby objects, outside the camera's field of view, or described at a level of detail that the detector does not support. Even when an object is detected, the guidance must distinguish between an item's position in the camera view and its position relative to the user or room. Language-model reasoning may improve flexible interaction, but it can also produce a description that sounds confident without adequate visual support. Such cases require careful evaluation and clear feedback about uncertainty.

ObjectFinder is important because it connects perception to an interactive, user-directed task. At the same time, its small participant sample and the reported residual errors in spatial tasks indicate that broader testing is needed [2]. Evaluation across unfamiliar object arrangements, lighting conditions, object sizes, and everyday environments would help establish how reliably the technique generalizes beyond the study setup. Research should also distinguish between successful object localization and the user's eventual ability to approach, identify, and safely retrieve the requested object.

3) Indian Sign Language Recognition

Rokade and Jadav investigate static ISL finger-spelling recognition through a computer-vision pipeline [3]. The method includes skin segmentation, conversion to a binary representation, Euclidean distance transformation, row and column projections, Fourier descriptors, and central and Hu moments. An artificial neural network (ANN) and a support vector machine (SVM) are used for classification. The work therefore represents a feature-engineering approach in which visual hand shape is transformed into a structured feature representation before the classifier predicts the sign.

The reported dataset contains 848 images at a resolution of 320 by 240 pixels for 17 signs. The study uses 548 images for training and 300 for testing [3]. Under the reported feature configuration, the authors report average accuracy of 94.37% for ANN and 92.12% for SVM. These values describe recognition within the chosen dataset and label set. They do not by themselves establish performance for continuous signing, different camera positions, signer-to-signer variation, or real-time conversation.

Static finger spelling provides a bounded problem that is useful for evaluating image classification, but sign language communication involves movement, timing, facial expression, body posture, and linguistic context in addition to hand shape. A model trained on separated still images may not adequately represent those features. Skin-based segmentation may also be sensitive to lighting and variation in skin appearance, while engineered descriptors may respond differently to orientation, occlusion, or background clutter. These constraints do not invalidate the reported experiment; they define the scope within which its result should be interpreted.

From an accessibility perspective, recognition is only one stage of a useful communication tool. The output must be displayed or spoken in a form that supports real interaction, and mistakes must be recoverable without creating a misleading impression of certainty. Future evaluations can report per-sign confusion, performance for previously unseen signers, continuous sequences, latency, and user feedback. A clear distinction between a constrained finger-spelling recognizer and a general sign-language translator is especially important when describing capabilities to users.

4) Gesture-to-text Interaction

Maben, Shetty, and Salian describe an ISL recognition prototype that uses a web interface and machine-learning tools including TensorFlow [7]. Camera-based hand gestures are presented to a trained recognition pipeline, and the predicted gesture is returned as text through the web interface. The paper describes support for standard and custom gestures and discusses the possible extension of the prototype to Android applications or wider institutional use [7]. Its emphasis is on moving from a predicted visual gesture to a readable communication output.

Gesture-to-text systems can lower some communication barriers when the set of supported gestures is understood by both parties. A text output may be useful in places where audio is not appropriate, where users cannot hear speech, or where a message needs to



remain visible. However, usability depends on how gestures are defined, whether the system can distinguish similar movements, how it responds when no known gesture is present, and whether the user can correct an incorrect prediction. A web interface also has practical requirements related to camera permission, browser compatibility, device performance, and access to a stable display.

The reported work is prototype-oriented and does not provide enough standardized quantitative evidence for direct numerical comparison with the dataset-based accuracy results reported by Rokade and Jadav [3], [7]. Consequently, the two ISL studies are complementary rather than directly rankable: one provides explicit dataset and classifier metrics for a static-sign problem, while the other emphasizes an interactive gesture-to-text workflow. Broader testing should include diverse signers, realistic lighting, sequences with transitions between gestures, and users who can assess whether the translated text conveys the intended meaning.

5) Real-time Hand Tracking

MediaPipe Hands is a general-purpose pipeline for tracking hands in real time from a single RGB camera [4]. It separates hand localization from landmark estimation. A palm detector first locates a hand region, after which a hand-landmark model predicts a set of landmarks representing hand pose in 2.5D. To avoid repeating the full detection stage on every frame, the system uses prior landmark predictions while tracking remains reliable and reactivates detection when confidence falls [4]. This design illustrates how pipeline organization can reduce redundant computation in continuous vision tasks.

The paper reports a set of experiments examining the speed and quality trade-off across model configurations and mobile-device settings [4]. The system is a component rather than a complete assistive application: it does not by itself determine which sign a person intends, explain the position of an object, or guide a user through a building. Its contribution is an efficient source of hand landmarks that downstream applications may use for gestures or hand-related interaction.

The distinction between tracking and interpretation is consequential. Landmark stability can be affected by occlusion, rapid motion, self-contact, camera angle, and hands that leave the field of view. A downstream gesture classifier may fail even when tracking is mostly accurate, and tracking metrics alone cannot measure whether the resulting interaction is useful. Assistive applications built on hand-tracking components should therefore assess both low-level continuity and the actual task: for example, whether a gesture is recognized correctly, whether feedback is timely, and whether a user can recover from missed frames.

The on-device emphasis is relevant to accessibility because real-time functions often need predictable response times and should not depend entirely on network access. However, device-level efficiency is only one deployment factor. Battery use over sustained sessions, thermal behaviour, memory, permission handling, and performance across low-cost devices may influence whether the component is practical outside a controlled demonstration. MediaPipe Hands thus represents an enabling technology whose value must be verified again in the context of the complete application [4].

6) Vision-language Navigation

Li et al. examine visual navigation assistance through a vision-language model fine-tuned for directional instruction generation [6]. Their approach fine-tunes BLIP-2 with Low Rank Adaptation (LoRA) on a manually annotated indoor-navigation dataset. They also describe an evaluation metric that refines BERT F1 to emphasize directional and sequential variables in navigation instructions [6]. The work illustrates a move from interpreting individual objects toward producing language that describes how a person might move through an environment.

The reported results indicate improved directional instruction generation after LoRA fine-tuning compared with the original BLIP-2 model [6]. Direction words and the order of movement are important because a sentence can be linguistically plausible while still giving an incorrect turn or a misleading sequence. An evaluation measure that considers these features is therefore more aligned with navigation than a generic language-similarity score alone. Nevertheless, the validity of any metric depends on its definition and its relationship to what a user must do in a real environment.

Navigation instructions are sequential and context-dependent. A single view may not show a destination or a corridor junction clearly, and movement can change the next useful instruction. A model may need to distinguish between “turn left at the next doorway” and “continue past the doorway,” where small wording differences can lead to very different movement. Visual language models must also handle uncertainty, unusual building layouts, temporary obstacles, and routes that differ from the annotated examples. A dataset-based improvement therefore should be treated as evidence for the evaluated instruction task, not as proof of safe navigation in any building.

The study points toward the benefit of joining visual features and language generation, while leaving questions about generalization across building types, lighting, camera height, and unseen routes [6]. Future evaluations should measure instruction correctness, task completion, user comprehension, route deviations, and recovery from misleading cues. This would clarify whether gains in directional text scores translate into better navigation outcomes for people who use the assistance.

B. Voice-based AI Technologies for Assistive Applications

1) Multilingual Voice Assistance

More et al. present an AI-based multilingual assistance platform in the context of visitor support during the Nashik-Trimbakeshwar Kumbh Mela [5]. The paper brings together speech recognition, natural language processing, text-to-speech, and computer vision with the aim of supporting guidance, navigation, emergency information, and multilingual communication. It discusses an offline-first orientation intended to maintain useful functions where network connectivity is limited, and it reports a preliminary survey of pilgrims, volunteers, and event staff to understand language preferences and accessibility needs [5].

The study frames assistive support as a service-level problem rather than a speech-recognition task alone. A system may need to understand a spoken request, identify what information is relevant, select an appropriate language, and present the response in an accessible form. Computer vision may contribute to situational context, while text-to-speech can make information available through audio. In a crowded public event, the overall experience is influenced by background noise, diverse accents, unclear



requests, multilingual place names, connectivity, and the timeliness of information. These factors interact and can affect the end result even if each individual component performs acceptably in isolation.

An offline-first design is important in settings where mobile connectivity may be unreliable or costly. It also introduces a practical question: which features can remain available offline, and which require current server-side information? Static directions or stored descriptions may be available locally, while changing event schedules, crowd conditions, or emergency notices may require updated data. A robust service should make such distinctions clear rather than presenting stale information as current. The reviewed paper describes a direction for combining components and provides preliminary context about user needs, but the abstracted evidence does not justify treating the platform as a fully validated solution for every language or event condition [5].

The evaluation of multilingual assistance benefits from more than a single system demonstration. Useful measures may include speech recognition by language and accent, the correctness of intent interpretation, response latency, speech intelligibility, user task completion, and performance during noise or connectivity loss. Accessibility feedback should also examine whether users can repeat, interrupt, or correct the interaction. These dimensions help distinguish technical availability from dependable service delivery. The paper contributes an application-level perspective, while its preliminary survey and event-specific context suggest the need for subsequent validation across varied public settings [5].

2) Wake-word Detection and Localization

HEiMDaL addresses the activation stage of a voice interface: detecting a wake word and locating its start within streaming audio [8]. The authors propose a low-footprint convolutional neural network and combine an alignment-based classification loss with an offset loss for predicting the beginning of the wake word. The reported evaluation compares the approach with a DNN-HMM-style baseline and describes a 73% improvement in the paper's detection metric while maintaining equivalent localization accuracy and the same memory footprint in the evaluated setting [8]. The result should be read in terms of that paper's specific detector benchmark.

Wake-word detection matters because voice-based assistance may need to listen for a short activation phrase before processing a longer utterance. Detection that is too sensitive can trigger on unrelated conversation or environmental sounds; detection that is too conservative can miss a genuine request. Localization can help identify the relevant audio segment and may support subsequent processing. These errors are not merely technical inconveniences: false activation can interrupt a user or raise privacy concerns, while missed activation can make an interface appear unresponsive.

A small audio model can reduce computational cost, but a realistic assessment should include more than memory footprint. Continuous operation also raises questions about energy consumption, background-noise robustness, speaker variation, accent variation, detection delay, and the handling of recordings or audio data. A wake-word component does not understand the command that follows and cannot guarantee a correct answer. It is therefore best understood as a specialized building block that can help make a voice interface more responsive when integrated with speech recognition, language interpretation, and accessible feedback.

HEiMDaL complements the multilingual assistance study because the papers describe different layers of voice interaction [5], [8]. The first considers a user-facing service that combines several AI functions; the second targets efficient audio activation. A future integrated evaluation could measure both component properties and user outcomes, but the available studies do not use a shared test setting. Accordingly, their reported results should remain separate rather than being merged into a single claim about voice-assistant effectiveness.

C. Cross-study Summary

Across the eight studies, assistance can be viewed as a layered process: sensing or receiving an input, interpreting it, selecting relevant context, generating guidance or a communication output, and allowing the user to confirm or correct the result. The papers address different layers of this process. MediaPipe Hands and HEiMDaL concentrate on efficient detection components [4], [8]; the sign-language papers transform visual gestures into labels or text [3], [7]; navigation and object-search studies connect visual information to spatial guidance [1], [2], [6]; and the multilingual platform combines multiple capabilities at the service level [5]. This layered interpretation helps compare what each study establishes without assuming that any one paper covers the full interaction chain.

The central comparison is not that one modality is universally superior to another. Computer vision can supply information about visible objects, gestures, and spatial scenes, but its usefulness depends on camera placement, visibility, and interpretation. Voice can support hands-free requests and spoken responses, but speech recognition and audio output depend on language, noise, hearing, privacy, and context. Text can support silent and persistent feedback, while gesture can provide expressive input for some users. Applications intended for accessibility should select modalities based on the task and user requirements and provide workable alternatives when one channel is unavailable.

TABLE I. COMPARATIVE SUMMARY OF SELECTED STUDIES

Study	Main focus and approach	Evidence / interpretation
Ivanov [1]	Indoor guidance using mobile, NFC and RFID	Eight blind participants; all reached target; mean 136 s. Depends on prepared infrastructure.
ObjectFinder [2]	Open-vocabulary object search using detector + MLLM	Eight participants; task-specific spatial metrics and feedback; broader field testing needed.
Rokade and Jadav [3]	Static ISL finger spelling; engineered features + ANN/SVM	848 images, 17 signs; reported 94.37% ANN and 92.12% SVM in the stated setup.
MediaPipe Hands [4]	Palm detection and 2.5D hand landmarks	Real-time tracking component; not a complete accessibility application.



Study	Main focus and approach	Evidence / interpretation
More et al. [5]	Multilingual service using speech, NLP, TTS and vision	Preliminary survey and application context; language-by-language validation remains important.
Li et al. [6]	BLIP-2 fine-tuning with LoRA for navigation instructions	Directional instruction generation improved in the reported dataset setting.
Maben et al. [7]	Web-oriented gesture-to-text prototype	Practical prototype described; standardized quantitative evidence is limited.
HEiMDaL [8]	Low-footprint CNN for wake-word detection/localization	Reported 73% improvement in detection metric with equivalent localization accuracy in the evaluated setting.

IV. RESULTS AND DISCUSSION

A. Technology Directions and Task Levels

The selected studies show a shift from bounded recognition tasks or prepared environments toward richer, interactive interpretation. Infrastructure-assisted indoor navigation uses explicit anchors placed in a building [1]. Static ISL recognition defines a closed set of signs and classifies individual images [3], while gesture-to-text work emphasizes a visible output through a web interface [7]. These approaches remain useful when a task is constrained and the supported classes are known, but their results cannot automatically be extended to open-ended communication or unfamiliar environments.

More recent work in the selected set explores flexible language and visual context. ObjectFinder accepts open-vocabulary requests and attempts to provide spatial context [2], while vision-language navigation generates directional instructions from visual input [6]. These applications place more responsibility on the relationship between perception, language, and the user's intended task. The increased flexibility is valuable, but it also creates new failure modes: a detector may find the wrong object, a language model may describe a scene inaccurately, or a direction may be understandable yet spatially incorrect. Evaluation must therefore test the full sequence of perception, interpretation, and action where possible.

The voice studies illustrate a similar distinction of levels. HEiMDaL optimizes an activation component [8], whereas the multilingual assistance platform combines several technologies to answer a broader public-service need [5]. A strong activation model can improve responsiveness, but it is not a measure of the accuracy of later language understanding or the usefulness of the service. Likewise, the availability of multilingual speech functions does not establish that every component works equally well for each language, accent, or noisy environment. Research reports should state clearly which layer is being evaluated.

B. User Context, Feedback, and Recovery

Across assistive applications, correct prediction is necessary but not sufficient. The user must understand the output and be able to act on it. An object-search system should distinguish a target object from similar objects and communicate where it is located in terms the user can use [2]. A navigation assistant should offer instructions at an appropriate time and provide a way to request clarification when the route is uncertain [1], [6]. A sign-language interface should preserve the meaning of a gesture and allow correction of an inaccurate text output [3], [7]. A voice interface should show that activation occurred and support recovery when speech is missed or misunderstood [5], [8].

This emphasis on recovery is especially important because assistive systems are often used in situations where users cannot easily inspect the underlying model output. An incorrect object label, missed hand landmark, or mistaken turn instruction may not be immediately obvious. Appropriate feedback can communicate uncertainty, ask a follow-up question, or allow the user to repeat the request. Future evaluations should examine not only first-attempt accuracy but also how well a system helps users detect and correct an error. Recovery time, number of clarification steps, and the user's confidence in the result may provide meaningful information alongside standard technical metrics.

User involvement should extend beyond a single final usability questionnaire. People with the relevant accessibility needs can help identify task definitions, meaningful output, acceptable interruption rates, and conditions that should be included in testing. The small participant studies in the selected literature provide useful initial evidence [1], [2], but they cannot represent every user, impairment profile, language, or environment. Larger and more diverse evaluations should account for variation in navigation experience, preferred interaction styles, technology familiarity, and the role of assistants or companions.

C. Evaluation Evidence and Comparability

The reviewed studies use multiple forms of evidence: task completion with participants, image-classification accuracy, technical metrics for tracking or wake-word processing, surveys, and prototype demonstrations. Each is useful but answers a different question. The indoor-navigation study reports whether participants reach a target and how long the task takes [1]. The ISL classification paper reports accuracy on a dataset of defined signs [3]. ObjectFinder reports spatial-task metrics and participant feedback [2]. A component paper such as MediaPipe Hands evaluates real-time tracking properties [4], while the multilingual assistance study includes preliminary service-context feedback [5].

These evaluation differences prevent direct numerical ranking. A higher percentage in one paper does not indicate a better overall assistive technology when the task and denominator differ. The 94.37% ANN figure in the static ISL study, the spatial-task metrics in ObjectFinder, and the 73% improvement described by HEiMDaL refer to different evaluation settings [2], [3], [8]. They should be preserved with those contexts and not combined into an average or league table. Comparisons are more defensible when organized by task type, deployment conditions, and type of evidence.

Reporting standards could improve interpretability without forcing all applications to use an identical metric. A paper can identify whether a result is based on users, images, audio segments, routes, or individual model predictions; state how test cases were



selected; and describe whether test users and environments overlap with training data. If latency is measured, the device, input conditions, and measurement boundary should be clear. If user studies are used, participant counts, recruitment context, task instructions, and success criteria should be described. These details allow readers to understand what a reported result supports.

Evaluation should also distinguish intended function from downstream outcome. Correct detection of a hand landmark does not guarantee correct sign recognition; correct object localization does not guarantee retrieval; and a correctly generated direction phrase does not guarantee that a route can be followed safely. End-to-end outcome measures should therefore be added when feasible, while component metrics remain valuable for diagnosing specific weaknesses. A layered evaluation design would make it possible to understand where an interaction succeeds or fails without attributing all performance to a single model.

D. Deployment, Robustness, and Responsible Use

Deployment conditions vary substantially across the selected studies. RFID-supported navigation requires installation and maintenance of environmental tags [1]. Camera-based systems depend on visible input, suitable device placement, and adequate lighting [2]–[4], [6], [7]. Voice-based systems depend on microphones, speech conditions, language coverage, and acoustic context [5], [8]. A model that works under one set of conditions may not remain reliable when deployed in a different building, with another device, or in a noisy public space. Deployment readiness therefore requires explicit attention to the conditions under which each function has been tested.

Computational efficiency is one practical factor. The real-time hand-tracking and wake-word studies both consider low-latency or compact processing [4], [8]. This matters for functions that operate continuously, where processing cost can affect battery life, thermal behaviour, and responsiveness. Yet efficiency should not be assessed in isolation from task performance. A smaller model that frequently misses user input may not provide a better experience, while a high-quality model that requires constant connectivity may be unavailable when the connection fails. Research should report the trade-offs between performance, computation, connectivity, and reliability.

Privacy and informed interaction deserve similar attention. Camera and microphone access may capture bystanders or sensitive activity, particularly in homes, hospitals, and crowded public settings. Voice activation mechanisms must balance responsiveness with false triggers and unwanted recording; visual systems should limit collection and retention to what the task requires. The selected papers do not collectively provide a single comparable privacy evaluation, so privacy should be treated as an important cross-cutting design and reporting need rather than a measured conclusion of this review.

Robustness also includes graceful behaviour when a component is uncertain or unavailable. If an object cannot be identified confidently, the system should not invent a definite location. If a wake word is missed, the user needs a way to reactivate or switch to an alternative input. If a network-dependent service becomes unavailable, the interface should explain which functions are offline rather than silently returning outdated information. These behaviours should be evaluated alongside nominal-case accuracy because accessibility use may involve unfamiliar or time-sensitive situations.

E. Research Gaps and Future Directions

Several research priorities recur across the selected literature. First, broader validation is required beyond small, fixed, or controlled evaluations. The indoor-navigation and object-search studies report participant groups of eight, which can provide early usability evidence but cannot establish broad population-level performance [1], [2]. Second, generalization should be tested across users, environments, languages, devices, and input conditions. Fixed sign inventories, manually annotated navigation data, and defined wake-word settings represent useful starting points but narrow evaluation contexts [3], [6], [8].

Third, future systems should be evaluated across complete user tasks, including recovery from errors. Recognition and localization metrics help diagnose components, but real-world assistance also depends on timing, comprehensibility, trust calibration, and successful completion. Fourth, research would benefit from more consistent documentation of datasets, train-test separation, participant characteristics, latency, resource consumption, and failure conditions. Such reporting can improve comparability while respecting the fact that sign translation, indoor navigation, object search, and wake-word detection require different primary metrics.

Fifth, multimodal interaction should be developed with a clear accessibility rationale. Combining vision, audio, language, and text can create useful alternatives, but more modalities do not automatically improve usability. A camera may be unavailable or poorly positioned; audio may be difficult in noisy settings; and visual text may not be accessible to a blind user. Designs should let users choose or switch interaction channels where appropriate and should explain the meaning of output through a channel the user can access. Integrated systems should also avoid treating a single generic response as sufficient for distinct accessibility needs.

Finally, future studies should investigate how efficient components can be integrated while preserving transparency and reliability. The selected literature demonstrates useful component-level advances in hand tracking and wake-word detection, and application-level work in multilingual support, object search, and navigation [2], [4]–[6], [8]. Integration research should measure not only whether components communicate technically, but whether the combined pipeline reduces task effort and produces fewer unresolved errors. Such evaluation should use realistic scenarios and involve users who represent the intended audience.

TABLE II. RECURRING RESEARCH GAPS AND FUTURE DIRECTIONS

Recurring gap	Evidence across selected studies	Future direction
Field validation	Some evidence comes from small participant groups, bounded tasks, or prototype demonstrations [1], [2], [7].	Test more users and varied real-world settings over longer periods.
Generalization	Fixed sign sets, annotated indoor data, selected sites and defined wake words constrain test conditions [1], [3], [6], [8].	Evaluate unseen users, layouts, objects, languages, devices and acoustic environments.



Recurring gap	Evidence across selected studies	Future direction
Task-level success	Component metrics do not always establish whether users complete an end task [2]–[4], [6], [8].	Combine component metrics with task completion, recovery and usability outcomes.
Error recovery	Incorrect recognition or guidance can be difficult to detect without suitable feedback.	Provide uncertainty cues, clarification, correction and alternative interaction channels.
Deployment trade-offs	Infrastructure, camera, compute, connectivity and audio conditions differ by application [1], [4], [5], [8].	Report installation needs, latency, resource use, connectivity and failure modes.
Reporting consistency	Studies use non-equivalent metrics and evaluation designs.	Document datasets, test conditions, participants, device details and failure cases consistently.

F. Limitations of This Review

The review is limited to eight selected studies and is not a systematic or exhaustive survey of assistive technologies. It combines papers with different publication types, study designs, target users, and technical goals. Some studies provide participant-level findings, others report component metrics or dataset accuracy, and some are primarily prototype-oriented. These differences restrict the strength of cross-study conclusions and mean that no pooled effect size or overall ranking is appropriate.

The review also depends on what the selected source papers report. Missing evaluation details cannot be reconstructed reliably, and the absence of a reported limitation should not be interpreted as evidence that a system has no limitation. The thematic gaps identified here should be understood as priorities suggested by the selected evidence, not as a claim that every assistive-technology study shares each problem. A more systematic review would require a predefined multi-database search, explicit inclusion and exclusion criteria, duplicate screening, formal quality assessment, and a larger set of studies.

V. CONCLUSION

The selected literature demonstrates a range of ways in which computer vision and voice-based AI can support accessibility. Indoor navigation uses environmental information to provide route guidance; open-vocabulary object search and vision-language methods connect visual inputs with spatial language; sign-language recognition and gesture-to-text systems translate constrained hand input into labels or text; and hand tracking provides a reusable real-time perception component. The voice-focused studies address complementary needs through multilingual service interaction and efficient wake-word detection [1]–[8].

The comparison also shows why reported results must be interpreted in their original context. Participant task completion, dataset classification accuracy, spatial-language metrics, and wake-word improvement metrics do not measure the same outcome. Across the selected studies, the clearest research priorities are broader real-world validation, stronger generalization, effective error recovery, user-centered evaluation, transparent reporting, efficient deployment, and accessible choices of interaction modality. Future work should connect component-level progress to task outcomes without overstating what a benchmark demonstrates. This paper provides a literature-level synthesis of existing work and does not introduce a new system or project-specific experiment.

DECLARATIONS

Conflict of Interest: The authors declare no competing interests relevant to this review.

Funding: No external funding was received for this review.

Ethical Approval: Not applicable to this literature-based review, which does not report new experiments involving human participants or animals.

REFERENCES

[1] R. Ivanov, “Indoor Navigation System for Visually Impaired,” in Proc. 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing (CompSysTech’10), Sofia, Bulgaria, 2010, doi: 10.1145/1839379.1839405.

[2] R. Liu, J. Zhang, A. Schön, K. Müller, J. Zheng, K. Yang, A. Guo, K. Gerling, and R. Stiefelhagen, “ObjectFinder: An Open-Vocabulary Assistive System for Interactive Object Search by People Who Are Blind,” International Journal of Human–Computer Interaction, 2026, doi: 10.1080/10447318.2026.2695932.

[3] Y. I. Rokade and P. M. Jadav, “Indian Sign Language Recognition System,” International Journal of Engineering and Technology, vol. 9, no. 3S, pp. 189–196, Jul. 2017, doi: 10.21817/ijet/2017/v9i3/170903S030.

[4] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, “MediaPipe Hands: On-device Real-time Hand Tracking,” arXiv:2006.10214, 2020.

[5] S. More, S. Shinde, M. Sharma, M. Vyavahare, and G. Dashmukhe, “AI-Based Multilingual Assistance Platform,” International Journal of Creative Research Thoughts, vol. 13, no. 11, pp. i76–i81, Nov. 2025, ISSN 2320-2882.

[6] X. Li, B. Gandhi, M. Zhan, M. Nehra, Z. Zhang, Y. Sun, M. Song, N. Zhang, and X. Wang, “Fine-Tuning Vision-Language Models for Visual Navigation Assistance,” arXiv:2509.07488, 2025.

[7] A. A. Maben, P. Shetty, and M. Salian, “Indian Sign Language Recognition,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 7, no. 4, pp. 149–152, Jul.–Aug. 2021, doi: 10.32628/CSEIT217436.

[8] A. Kundu, M. S. Razlighi, M. Cho, P. Padmanabhan, and D. Naik, “HEiMDaL: Highly Efficient Method for Detection and Localization of Wake-Words,” in Proc. ICASSP 2023, 2023, doi: 10.1109/ICASSP49357.2023.10097018.