WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Voice and Vision: Using AI to Empower Vision and Bridge Communication Gap Among Individuals with Sensory Disabilities

Monica S[1],Nandini B[2],Pravalika N S[3],Sneha K M[4] and Dr. Leena Giri G[5]

[1]Student,CSE Programme,Dr. Ambedkar Institute of Technology,Bengaluru,India Email: monicashankar2020@gmail.com
[2]Student,CSE Programme,Dr. Ambedkar Institute of Technology,Bengaluru,India Email: bnandu252004@gmail.com
[3]Student,CSE Programme,Dr. Ambedkar Institute of Technology,Bengaluru,India Email: pravalikans2004@gmail.com
[4]Student,CSE Programme,Dr. Ambedkar Institute of Technology,Bengaluru,India Email: sneha.km.2204@gmail.com
[5]Associate Professor,CSE Programme,Dr. Ambedkar Institute of Technology,Bengaluru,India

A B S T R A C T

This paper introduces Voice and Vision, a unified platform that uses Artificial Intelligence to address communication challenges faced by sensory disabled people. This application assists mute users by translating sign language gestures to speech. For visually impaired individuals, this application uses a camera to detect the objects around them and provides audio feedback and also provides text-to-speech functionality to read printed text aloud. For deaf users, the platform provides instantaneous speech-to-text transcription. A multi-threaded architecture helps in parallel execution of camera processing, speech recognition, text-to-speech generation and Flask based web interactions. By integrating all these features Voice and Vision application assists users with all sensory disabilities.

Keywords: Artificial Intelligence, Computer Vision, Sign language recognition, Speech to Text, Text to Speech, Object Detection, multimodal communication

1. Introduction

People with sensory disabilities such as blindness, deafness, and muteness often find it difficult to communicate with the people around them. However, there are several assistive technologies that have been developed to support such people. But these technologies concentrate on particular impairments and give solutions to a single type of disability which requires the users to manage multiple tools at once.

Rapid advancements in Artificial Intelligence (AI) and Computer Vision have created opportunities for innovation in assistive technologies. Machine learning models enable real-time analysis of gestures, speech and environmental factors; computer vision allows devices to recognize and describe surroundings. These innovations play an important role in improving independence and social inclusion among individuals with sensory disabilities.

Despite technological advancements, current assistive technologies are often fragmented. Many tools provide assistance to single disabilities, thus the users with visual impairment, deafness or muteness are required to frequently switch between the applications. This lack of integration creates barriers to independent communication.

To bridge this communication gap, we propose Voice and Vision: a unified, AI-powered assistive application that supports people with sensory disabilities. Voice and Vision aims to:

1. Provide visually impaired people with the awareness about common objects in their surroundings.
2. Support visually impaired people by reading printed or digital text aloud making the information more accessible.
3. Support deaf users by converting the spoken words into text allowing them to engage in day-to-day conversations.
4. Support mute individuals to communicate by converting sign-language gestures into text which helps other people to understand.

2. Literature Review

Artificial intelligence and deep learning have transformed assistive technology, especially for people with sensory disabilities. The problem is, most systems today only address one issue at a time they might help with hearing or vision, but rarely both. That leaves accessibility pretty fragmented.The Voice and Vision project aims to change this. The goal is to combine functionalities such as sign language recognition, speech-to-text, text-to-speech, real-time object detection, and reading printed text into one unified, intelligent system.

2.1. Sign Language Recognition

Deep learning-based sign language recognition (SLR) has become essential for connecting deaf and mute individuals with the wider world. Systems built on convolutional neural networks, like those described in [1] Deep Learning Based Indian Sign Language Recognition for People with Speech and Hearing Impairment and [2] Gesture Recognition in Indian Sign Language Using Deep Learning Approach , deliver impressive accuracy when recognizing static gestures. Still, they tend to shine only in controlled settings where the background doesn't change much.

Landmark-driven SLR approaches stand out as stronger, more reliable options. [3] Indian Sign Language Recognition Using LSTM and MediaPipe demonstrates that extracting hand and pose landmarks greatly improves generalization across different lighting conditions. LSTM modeling captures dynamic gesture transitions. These lightweight pipelines are suitable for mobile use, which is important for real-time applications.

User-centered improvements are clear in [4] Gesture Guide: Empowering Deaf and Mute Individuals with Indian Sign Language Recognition and Guidance System Using An AI Powered App Hand Signs. This system incorporates guided visual feedback to help users with gesture formation and enhance usability. Deployment-focused research such as [5] Sign Comm: A Real-Time Indian Sign Language Recognition System Using Deep Learning for Inclusive Communication highlights low latency as a key factor in natural communication scenarios.

End-to-end translation pipelines, like those in [6] Indian Sign Language Recogniser with Text and Speech Translation, confirm the feasibility of converting gestures into both text and speech output. Although recognition systems have advanced, challenges like signer independence, vocabulary expansion, and environmental variance remain.

2.2. Object Detection

Object detection is vital for helping blind or low-vision users navigate and understand their surroundings. [7] An Advanced Auditory Response for Object Detection Using Deep Learning investigates ways to connect detected objects with sound cues. The focus is on clarity and managing mental effort to aid real-time understanding. This highlights the need for sound feedback systems that provide useful information without flooding the user.

Similarly, [8] Smart Blind Assistant Using Deep Learning for the Visually Impaired Users describes a supportive system that merges camera input with deep learning models to recognize nearby objects and deliver information through speech. This study focuses on practical issues like processing speed, device portability, changing environment, and the limits of on-device processing. These findings highlight the need for efficient, low-latency visual recognition models in a support platform.

2.2.Multimodal Communication

Multimodal communication is essential for users with hearing or vision impairments. [9] Speech-to-Text Conversion and Text Summarization provides a practical way to turn raw speech into structured text. It highlights challenges like environmental noise and transcription accuracy, which are essential for real-time captioning in assistive systems.

Learning methods like [10] Joint Speech-Text Embeddings for Multitask Speech Processing suggest shared embedding spaces that improve performance across ASR, keyword spotting, and semantic retrieval tasks. These methods increase robustness, context awareness, and accuracy, offering useful insights for Voice and Vision's speech-understanding module.

Text-to-speech (TTS) capabilities show up in sign-to-speech systems like Indian Sign Language Recogniser with Text and Speech Translation. They prove that synchronous communication works well when visual gestures map to speech. Adding these features with OCR-based printed-text reading broadens accessibility for users who are visually impaired or face literacy challenges.

The majority of speech and language systems operate independently and are not integrated with visual or gestural modalities, despite encouraging results. By combining text extraction, SLR, speech recognition, and TTS into a single, multimodal communication system, Voice and Vision fills this gap.

According to the paper [11] "A Unified Communication Model for Multiple Sensory Disabilities Using Convolutional Neural Networks" there is lack of an integrated communication environment which restricts independence for users with multiple impairments. This shows that there is need for a combined, multimodal assistive platform. The gaps found in gesture recognition, object detection, and multimodal communication highlight the need for an integrated system. By combining these features into one platform, Voice and Vision assists people with various sensory disabilities.

3. Methodology

This App brings together computer vision, speech and sign-language recognition into one accessible platform. Every feature responds directly to the real, everyday challenges faced by people who are visually impaired, deaf, or speech-impaired. Every feature is designed for real time assistance, using multiple communication modes.

3.1. Features that Support Visually Impaired Users

For visually impaired users the app turns what the camera sees into spoken descriptions. Two main features are :

- Real-Time Object Detection

We use a YOLO-based detector that watches the camera feed nonstop. When it spots any object it detects it and produces text output and the Text-to-Speech engine reads it out loud. This way, users can recognize what's around them, dodge obstacles, and move independently.

- Text Extraction and Reading

This App uses Pytesseract OCR which extracts text from camera feed and the app reads the text out loud. Thus it can be used for reading out signs, menus or any printed text aloud.

3.2. Features that Support Deaf Users

For deaf users the app turns speech into text and allows for sign language recognition.

- Speech-to-Text
  Speech gets converted into text on the spot and pop up on the web interface. Deaf users just read what's being said, as it happens.
- Real-Time Sign Language Recognition

  A deep learning model scans hand gestures on live camera feed. Sign Language letters and words are recognized and is displayed as text on the screen. This allows for a two-way communication: signers have their gestures translated, while non-signers can understand without learning the language.

3.3. Features that Support Speech-Impaired Users

This App captures hand movements and shapes and then matches it to sign language gestures. Once the sign is recognized the message is displayed on the screen.

3.4. Multimodal Interaction:

The platform combines different input and output options, so each group interacts in their own way:

Table 1 - Multiple modes of interaction

USER TYPE	INPUT MODE	OUTPUT MODE
Visually Impaired	Voice Commands and Camera View	Audio Feedback
Deaf	Camera and Recorded Audio files	Text displayed on the screen
Speech Impaired	Sign Language Gestures	Text output

3.5. Real-Time Performance with Multi-Threaded Design

The system is efficient and operates in real-time through the multi-threading capabilities of the operating system. One thread is always active listening to the user's voice to identify any requests spoken by the user in real time without lag and ensuring that no words are missed. At the same time, another thread is live, running the camera feed and analyzing the feed of frames in real time to detect objects, faces, or movements instantaneously. A Text-to-Speech thread is live building the audio response, and the speech output is obtained immediately without any delay.

Flask handles web requests on its own thread, ensuring the web interface is quick and responsive, even when audio or image is being processed in the backend. This multi-threaded architecture allows the entire system to listen, see, and respond at the same time. This means it is able to observe, signs, read text in the camera's view, and respond to the user in that context all in real time. Each of these are processed without lag and in parallel with other processes

3.6. Core Components

This application integrates Flask, OpenCV, YOLO, Tesseract OCR, Speech Recognition, Text-to-Speech, and a Tensorflow based sign language model. These components operate together, forming a real-time, multimodal assistant.

- **Flask Backend & Routing**

Flask manages everything. It handles user requests, serves web pages, and manages routes for object detection, OCR, speech input, sign language recognition, and live video streaming. The server runs in multi-threaded mode to stay responsive, even with multiple tasks running at once.

- **Camera Processing Pipeline**

OpenCV continuously captures frames from the camera. Each frame is sent directly to the chosen module—YOLO for object detection, Tesseract for extracting text, or the sign language model.

- **Object Detection (YOLO)**

YOLO detector is used to process every frame, and it helps in identifying objects and drawing labeled bounding boxes immediately. For visually impaired users, the detections are converted into text descriptions which are then read out loud.

- **OCR (Text Extraction)**

Pytesseract is used to extract printed text from the camera feed. The app cleans up the recognized text, displays it on screen, and it is then read out loud using text to speech engine.

- **Speech Recognition**

The SpeechRecognition library transcribes what it hears into text, which appears for the user and also acts as voice commands for switching modes.

- **Text-to-Speech (TTS) Engine**

The TTS engine runs in its own thread. Using pyttsx3 or gTTS, it generates and caches audio responses, ensuring quick feedback and avoiding repetition. This provides users especially those with visual or speech impairments with real-time audio prompts.

- **Sign Language Recognition**

A trained sign language model interprets hand gestures directly from the camera. When it detects a sign—be it a letter or a word—it displays it on the screen and says it aloud. This gives people who can't speak an alternative way to communicate, and it functions in real time.

- **Frontend Interface**

The frontend is built with HTML, CSS, and JavaScript. The interface is kept simple, so anyone can use it easily.

- **Multi-Threading & Real-Time Performance**

Each main process which includes camera capture, speech recognition, and text-to-speech operates on its own thread. Thus listening, detecting, recognizing, and responding all occur simultaneously**.**

4. Results

The primary achievement of this project is the successful creation of the Voice and Vision system, a unified, AI-powered mobile application that is designed to overcome communication and accessibilitiy challenges for people who are physically disabled. People needing assistance relied on multiple, separate tools, which was often fragmented, costly and inefficient. This system directly solves this by consolidating four key features into efficient platform:

1. Sign Language Detection
2. Speech to Text Converter
3. Text to Speech Converter
4. Object Detection

The System's foundation uses adaptive Artificial Intelligence Machine Learning (AIML) capabilities and is built with Python libraries like Open CV and pyttsx3, ensuring it is both scalable and efficient.

Functionality

Communication Support

- This app can understand hand signs instantly and change them into spoken words and text. This is very helpful in places like hospital where staff may not know sign language, making communication easier for mute users.
- It also works in opposite way it can take someone's voice or types message and show is as sign gestures, helping both sides communication smoothly.
- The Speech to Text module allows deaf users to actively participate in conversation ,meetings or classrooms by capturing live speech and converting it into readable text.

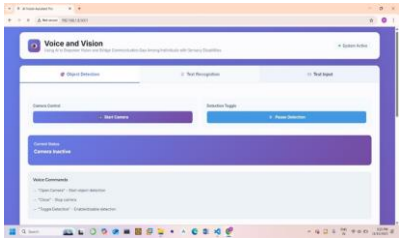
Navigation and Safety Enhancement

- The Object Detection and Identification feature uses the camera to recognize and describe nearby objects, such as furniture and traffic lights etc.
- This feature is useful during the emergencies, where system can alter the user about the danger or allow them to activate voice based emergency message instantly through specific hand signs.

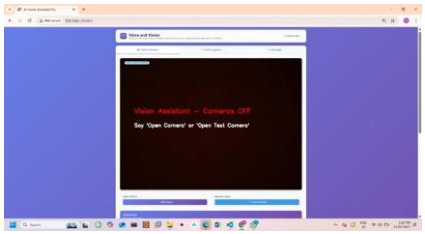
Text Accessibility

- The text to speech feature can read out any printed or digital text after scanning it.
- This helps people who have trouble seeing or reading.
- It supports them in doing everyday activities on their own, like monthly bills , labels or paragraphs.

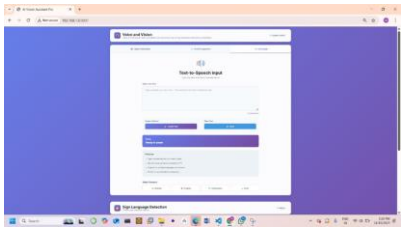
Snapshots



Home page



Object detection



Text to Speech

Sign Language Detection

5. Conclusion

Voice and Vision project has built a helpful mobile app that uses smart technology to support people with hearing or vision difficulties. The app makes daily tasks easier and helps them communicate more comfortably, giving them more confidence and independence. The key accomplishment is the seamless integration of crucial assistive feature from real time sign language interpretation to dynamic object detection into one portable platform. This unified approach directly resolves the major problem of fragmentation and complexity in older assistive technologies.By effectively combining visual and auditory intelligence, the system successfully bridges the communication gap between the mute and hearing communication while providing essential safety and navigational supports for the visually impaired. Ultimately, Voice and Vision is a significant step toward personalized accessibility, demonstrating that comprehensive two-way support is technically feasible within a single, accessible device . The system functions as a unified digital companion that promotes natural, inclusive interaction in all daily environments.

6. Future Scope

In the future, this application can be further improved by adding a GPS-based navigation system that guides visually impaired users through speech, obstacle warnings, and suggestions for safe routes.The sign-language recognition model can be made better by collecting samples from more people with different hand shapes and lighting conditions. If the model gets used to these variations, it will respond better. If it incorporates advanced sign language translation models, then it will allow deaf individuals to have natural, two-way conversations with people in their surroundings. The inclusion of multilingual speech-to-text and text-to-speech functionality will further make access easier for users from different regions and, therefore, almost eliminate problems resulting from language barriers. It will truly be an all-in-one support tool if it includes gentle haptic feedback to help with navigation and if it is compatible with wearable devices and capable of working in offline mode. Cloud technologies can be used for supporting heavy processing, saving large amount of user data.

References

[1] A. Kolkur, A. Yattinmalgi, G. Korimath, S. Chikkamath, N. S. R., and S. Budihal, "Deep Learning Based Indian Sign Language Recognition for People with Speech and Hearing Impairment," *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)* , Bangalore, India, 2024, pp. 1–5, doi: 10.1109/InC460750.2024.10649094.

[2] P. S. Reddy, N. S. Kumar, B. Teja, A. B. Prasad, S. Hariharan, and V. Kekreja, "Gesture Recognition in Indian Sign Language Using Deep Learning Approach," *2024 International Conference on Computing and Data Science (ICCDS)* , Chennai, India, 2024, pp. 1–6, doi: 10.1109/ICCDS60734.2024.10560428.

[3] M. Jha, H. P. Rajawat, H. Suhagiya, and L. Kurup, "Indian Sign Language Recognition Using LSTM and MediaPipe," *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)* , Mumbai, India, 2023, pp. 1–6, doi: 10.1109/ICACTA58201.2023.10393604.

[4] S. Rithesh Manikandan, G. L. Prasuna, G. Sivanesh, and K. Divya Lakshmi, "Gesture Guide: Empowering Deaf and Mute Individuals with ISL Recognition and Guidance System using Hand Signs," *2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* , Tirunelveli, India, 2024, pp. 462–465, doi: 10.1109/ICICV62344.2024.00078.

[5] I. A., K. P., A. A., R. R., and M. K., "Sign Comm: A Real-Time Indian Sign Language Recognition System Using Deep Learning for Inclusive Communication," *2024 International Conference on Emerging Research in Computational Science (ICERCS)* , Coimbatore, India, 2024, pp. 1–8, doi: 10.1109/ICERCS63125.2024.10895381.

[6] J. David, M. S. S., and S. Revathy, "Indian Sign Language Recogniser with Text and Speech Translation," *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)* , Tirunelveli, 10.1109/ICOEI65986.2025.11013145.

[7] L. H. Medida *et al.*, "An Advanced Auditory Response for Object Detection Using Deep Learning," *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, Namakkal, India, 2023, pp. 843–847, doi: 10.1109/ICECAA58104.2023.10212126.

[8] V. V. Reddy S., B. Sathyasri, J. J. Jeya Sheela, M. C., S. Vanaja, and S. S., "Smart Blind Assistant Using Deep Learning for the Visually Impaired Users," *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2023, pp. 680–684, doi: 10.1109/ICAISS58487.2023.10250718.

[9] P. Agrawal, K. Sharma, K. Dhage, I. Sharma, N. Rakesh, and G. Kaur, "Speech-to-Text Conversion and Text Summarization," *2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP)*, Bali, Indonesia, 2024, pp. 536–541, doi: 10.1109/TIACOMP64125.2024.00094.

[10] M. G. Gonzales, P. Corcoran, N. Harte, and M. Schukat, "Joint Speech-Text Embeddings for Multitask Speech Processing," *IEEE Access*, vol. 12, pp. 145955–145967, 2024, doi: 10.1109/ACCESS.2024.3473743.

[11] A. Tomar, P. Jain, V. Panwar, V. K. Gupta, V. Shukla, and Vidisha, "A Unified Communication Model for Multiple Sensory Disabilities Using Convolutional Neural Networks," *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, Ghaziabad, India, 2024, pp. 1–5, doi: 10.1109/ACET61898.2024.10730536.