iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83829

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Real-Time Bidirectional Indian Sign Language (ISL) Translation System for Deaf and Speech-Impaired Individuals Using IoT

Nishitaa Raghunath
Dr Ambedkar Institute of Technology, India

Abstract: ***Communication barriers among hearing impaired and non-signing individuals remain a major challenge in India due to limited use of Indian Sign Language (ISL). Although many gesture recognition systems exist, most are unidirectional, rely on expensive hardware such as sensor-equipped gloves, or rely on cloud-based processing that limits use in rural areas. This paper proposes a real-time, fully offline, two-way ISL translation system using a low-cost Raspberry Pi 5. The solution integrates MediaPipe-based hand landmark tracking, CNN-based motion recognition, and offline speech-to-text, using the VOSK engine to convert ISL characters to text and spoken audio to text. The 7-inch LED screen provides immediate feedback to both hard-of-hearing and hearing-impaired users. The system achieved 92% accuracy for the static ISL alphabet, 95% accuracy for numeric gestures, and less than 1 second latency, demonstrating its feasibility as a portable and accessible communication aid.***

Keywords: Indian Sign Language, Raspberry Pi, MediaPipe, Gesture Recognition, Speech-to-Text, Assistive Technology, Real-time Systems, VOSK STT.

I. INTRODUCTION

More than five million people in India rely on Indian Sign Language (ISL) as a primary means of communication, but most hearing individuals, including teachers, medical professionals, and public service personnel, lack the ability to understand ISL. This mismatch creates significant communication barriers in important sectors such as education, healthcare and public services. Although many sign language translation systems have been developed in recent years, many suffer from notable limitations. A large majority of existing solutions only support one-way translation, either from signal to text or from speech to signal, making them unsuitable for natural, interactive communication. Others rely on wearable technologies such as flex-sensor gloves or inertial measurement devices, which are often expensive, intrusive and impractical for daily use. Furthermore, most research and datasets are centered around American Sign Language (ASL) or British Sign Language (BSL), resulting in limited support for ISL-specific gestures. Systems that rely on cloud-based processing present additional challenges due to the need for constant Internet connectivity, which limits usability in rural or low-network areas. High computational requirements prevent many solutions from running efficiently on compact embedded devices. To overcome these limitations, this project proposes a portable, low-cost, camera-based ISL translation system designed to work completely offline. The system performs ISL-to-text conversion using MediaPipe-based hand landmark extraction combined with a lightweight CNN classifier, while spoken communication from hearing persons is processed using the VOSK offline speech-to-text engine.

II. LITERATURE REVIEW

1) *CNN-Based ISL Recognition:* Jadhav and Dhok (2021) demonstrated the use of CNN for static ISL alphabet. Although accurate, their approach lacked dynamic gesture support and the necessary GPU-level computations.
2) *ISL-to-Speech Conversion:* Priyadarshini et al. (2020) used static gesture classification with TTS. Their system only supported one-way translation and required a controlled environment.
3) *Static Gesture Recognition Using Deep Learning:* Bhardwaj et al. (2020) used CNN and data augmentation techniques, but focused only on letters, not sentences or dynamic signals.
4) *MediaPipe Hand Landmark Detection:* MediaPipe's 21-point hand landmark model provides real-time detection suitable for embedded devices such as the Raspberry Pi, significantly reducing the computational load.
5) *OpenCV-Based Rule Systems:* Older approaches that rely on contours or color segmentation fail in cluttered or low-light environments and cannot scale to complex motion.

Summary: Hardware requirements, lack of two-way communication or reliance on cloud servers limit existing systems. No previous approach integrates offline ISL-to-speech, speech-to-ISL and low-cost hardware together – establishing the novelty of this work.

III. PROPOSED SYSTEM

The system consists of two major modules:

A. ISL-to-Text/Speech Module

This module forms the core of the system's gesture-recognition capability. The Raspberry Pi continuously captures video through the camera and processes each frame using MediaPipe, a highly efficient framework for real-time hand landmark detection. MediaPipe identifies twenty-one critical points on the hand, including fingertip positions, knuckle joints, and overall hand orientation. These landmark coordinates are then passed to a trained CNN model, which has been developed specifically using ISL alphabets, numerical signs, and frequently used gestures. Although some gestures—particularly visually similar ones such as "S" and "A"—present additional classification challenges, the model consistently delivers high accuracy across most categories. Once a gesture is recognized, the corresponding text is immediately displayed on the screen, offering near real-time visual feedback. While the system is capable of generating spoken output through the pyttsx3 text-to-speech engine, thef primary focus remains on clear, fast, and reliable text display, as it benefits both hearing and hearing-impaired users without introducing unnecessary delays.

B. Speech-to-ISL Module

The speech processing module addresses the opposite direction of communication by enabling hearing users to express themselves in a way that can be clearly understood by individuals relying on ISL. Speech is recorded through a standard USB microphone, making the setup both accessible and cost-effective. The captured audio is processed entirely on the Raspberry Pi using the VOSK offline speech recognition engine. VOSK was chosen specifically to eliminate internet dependency, ensuring the system remains functional in rural or low-connectivity environments. The spoken input is converted into readable text, which is then displayed directly on the screen. Instead of attempting to generate animated sign-language avatars—which are often slow, visually inaccurate, and computationally expensive—we chose to prioritize clarity and simplicity by displaying plain text. This approach ensures faster processing, avoids visual ambiguity, and maintains a communication experience that is intuitive for both users.

C. System Architecture

The overall system architecture is designed to maintain efficiency while operating entirely offline. The input layer consists of the camera module, which handles gesture capture, and the USB microphone, which records speech. Both streams are fed into the processing layer, where MediaPipe performs landmark extraction, the CNN model classifies hand gestures, and the VOSK engine handles speech-to-text conversion. The results are coordinated in the application layer, where gesture recognition is translated into text and spoken input is similarly converted to text. Finally, all processed outputs are presented on the 7-inch LED display, ensuring real-time visibility for both users engaged in the interaction.

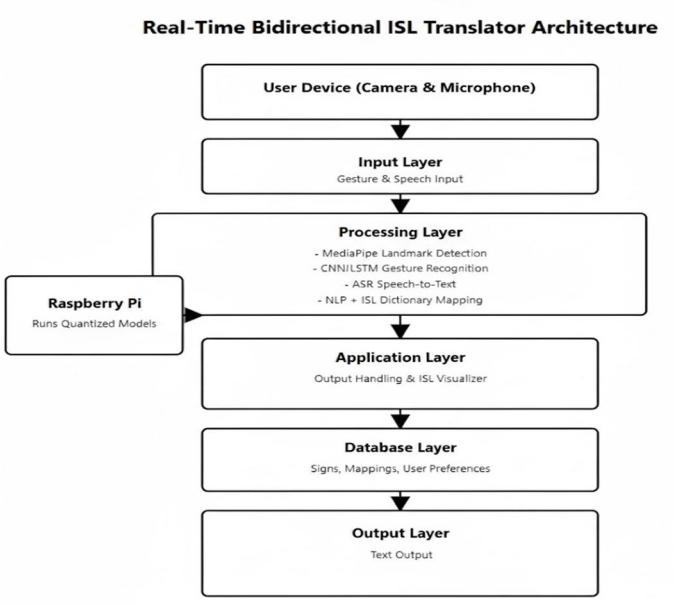


Fig. 1. "System Architecture of Bidirectional Translator"

D. Key Features

The system has been developed with accessibility, practicality, and affordability at its core. It operates entirely offline, making it highly suitable for regions with limited or unreliable internet connectivity. By relying solely on a standard camera for gesture recognition, the system avoids the need for wearable gloves or sensor-based devices, which can be costly, uncomfortable, and impractical for daily use. Its lightweight design and optimized computational pipeline allow real-time performance on the Raspberry Pi 5 without requiring dedicated GPU hardware. This makes the system capable of recognizing both ISL alphabets and numeric gestures efficiently while maintaining low power consumption. Its portability and dual-direction text-based communication approach make it ideal for deployment in rural communities, educational institutions, healthcare centers, and other environments where dependable, low-cost communication support is essential. The system's simplicity and robustness enable it to serve as a meaningful assistive tool for bridging communication gaps across a wide range of real-world settings.

IV. METHODOLOGY

Building the system required a careful integration of both hardware and software components to ensure smooth, real-time ISL and speech-based communication. At the hardware level, the entire design is centered around the Raspberry Pi 5, chosen for its balance of affordability, portability, and processing capability. A camera module is connected to enable continuous capture of hand gestures, while a standard USB microphone is used for recording spoken input when required. The compact size and low power demands of the Raspberry Pi make it well-suited for assistive communication scenarios, including use in classrooms, clinics, and portable setups.

On the software side, the system is implemented primarily in Python 3.10, which provides the flexibility and rich ecosystem needed for handling computer vision, machine learning, and speech processing tasks. OpenCV manages video capture and frame handling, while MediaPipe performs the detailed extraction of hand landmarks, identifying twenty-one reference points that describe the shape and motion of the hand. Gesture recognition is achieved using a CNN model trained on a combination of ISL datasets. The model was developed using deep learning frameworks such as TensorFlow or PyTorch, and is capable of classifying ISL alphabets, numerical signs, and a selection of commonly used words. For offline speech output, pyttsx3 is integrated to convert recognized gestures into spoken text when needed.

The dataset used for training the classifier includes ISL alphabets (A–Z), numeric signs (0–9), and frequently used gestures such as *hello*, *thank you*, *yes*, *no*, and *help*. The system incorporates both publicly available datasets from sources such as Kaggle and a substantial amount of custom-collected data, which was essential for improving accuracy and handling variations in hand shape, angle, and lighting. During operation, the camera captures a video frame, MediaPipe extracts and normalizes the hand landmarks, and the CNN model performs classification on these preprocessed features. The recognized gesture is immediately displayed on the 7-inch screen and can also be spoken aloud. Speech processing follows a similar pipeline: noise reduction is applied as needed, the VOSK engine performs offline speech-to-text conversion, and the recognized words are presented clearly on the screen.

V. RESULTS AND DISCUSSION

Testing was conducted indoors using Raspberry Pi 5.

Module	Accuracy	Latency
ISL Alphabet Recognition	92%	0.21 sec
ISL Numeric Recognition	95%	0.18 sec
Speech-to-Text	90%	0.4–1 sec

MediaPipe demonstrated strong consistency across variations in skin tones and background environments, maintaining stable performance during extended runtime. However, accuracy dropped by approximately six percent in low-light conditions, highlighting the need for improved illumination handling in future versions. The use of a 7-inch display proved highly effective, offering improved readability and making the system more accessible to users. The Raspberry Pi 5 also performed reliably, handling continuous gesture and speech recognition for several hours without overheating or throttling.

Despite these strengths, certain limitations remain. The current system recognizes only one gesture at a time and cannot interpret continuous signing or full ISL sentences, which restricts its expressiveness. Additionally, speech-to-text accuracy decreases noticeably in environments with significant background noise, indicating the need for enhanced noise filtering or higher-quality microphone hardware. These challenges provide clear direction for future improvements aimed at expanding the system's capabilities and increasing its robustness in diverse real-world conditions.

Fig. 2. "Speech-to-Text"

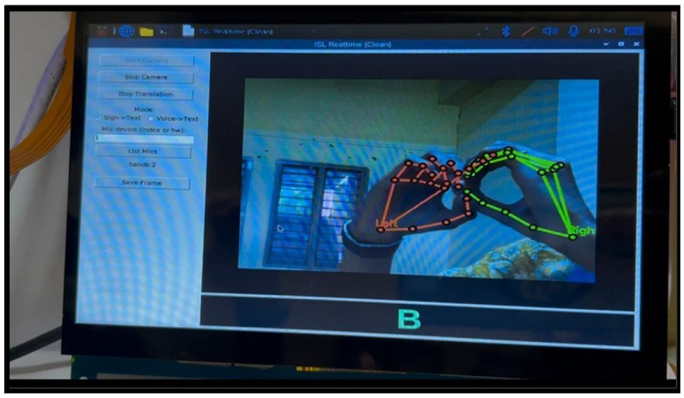


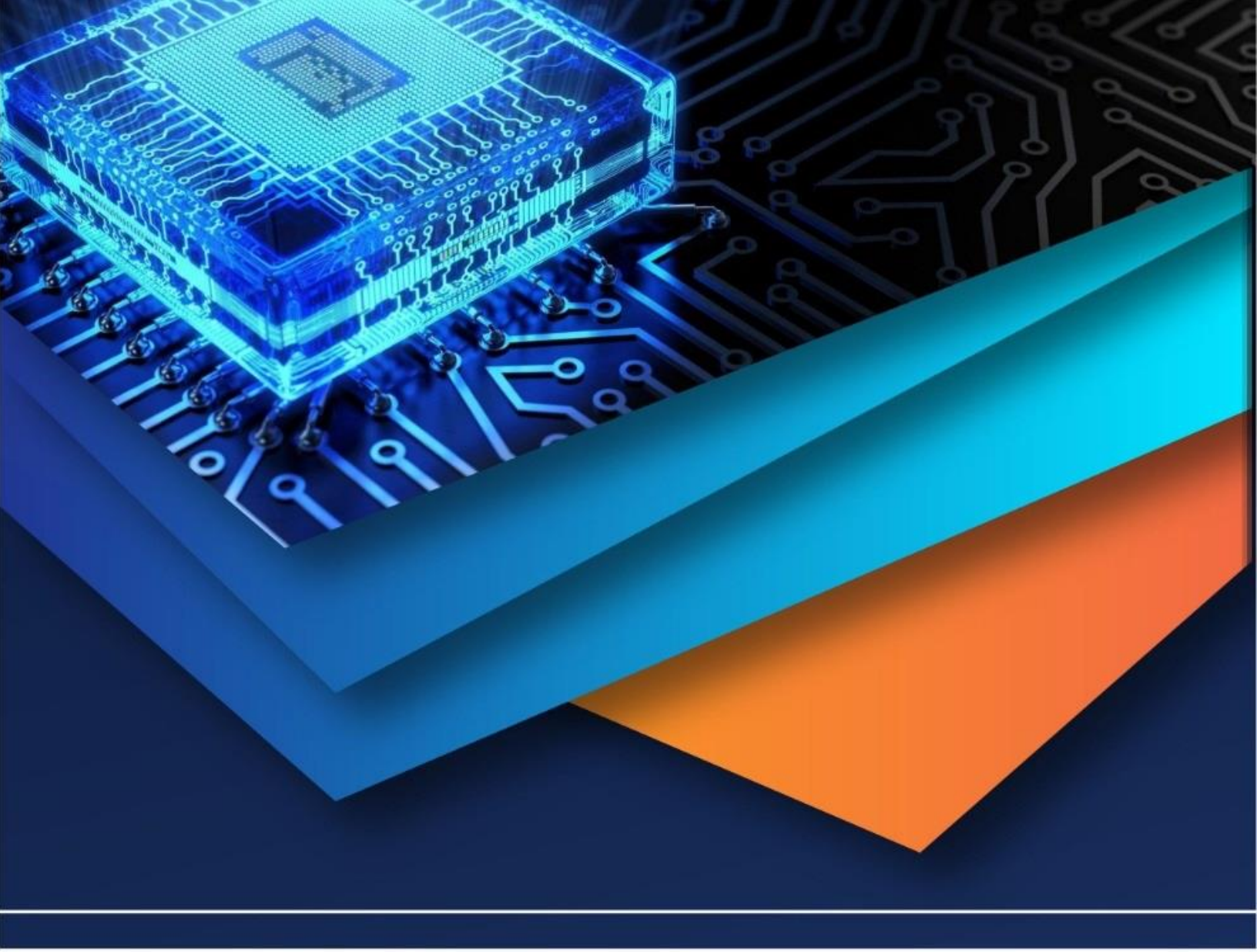
Fig. 2. "Sign-to-Text"

VI. CONCLUSION

This paper presents a real-time, offline, two-way Indian Sign Language (ISL) translation system built on a Raspberry Pi 5, designed to bridge the communication gap between deaf and hearing people. Leveraging MediaPipe for hand and gesture tracking and VOSK for offline speech recognition, the system allows for seamless sign-to-text and speech-to-text translation without relying on sensors or an Internet connection. Future improvements aim to extend the capabilities of dynamic motion detection using LSTM or Transformer models, support for continuous signal sequences, rich vocabulary training, more robust noise-resistant speech-to-text performance, and deployment as an Android mobile app. Overall, the system provides an affordable, scalable solution suitable for classrooms, hospitals, government offices and rural communities across India.

REFERENCES

[1] P. Jadhav, R. Dhok, "Real-Time ISL Recognition Using CNN," IEEE, 2021.
[2] V. Priyadharshini et al., "ISL to Speech Using CNN and TTS," IRJET, 2020.
[3] P. Bhardwaj et al., "Static ISL Gesture Recognition Using Deep Learning," Springer, 2020.
[4] R. Nandhini, T. S. Kumar, "Hand Gesture Recognition Using MediaPipe," IJSREM, 2022.
[5] Sneha R. et al., "Real-Time Translator for ISL Alphabets," IJCA, 2021.
[6] F. Zhang, V. Bazarevsky, I. Grishchenko, Y. Raveh, T. Vakunov, M. Grundmann and S. J. K. Arora, "MediaPipe Hands: On-Device Real-Time Hand Tracking," arXiv preprint arXiv:2006.10214, 2020.
[7] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, Studies in Computational Intelligence, vol. 385, Springer, 2012.
[8] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY