IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

Bi-directional Multimodal System for Real-time Translation between Deaf, Blind and Hearing Individuals

Ch. Vijayananda Ratnam[1], N. Nikhitha[2], N. Yamini Preethi[3], M. Jyothi Bai[4], N. Greeshma Naga Sri[5], M. Akhilesh[6]
[1]Assistant Professor, Department of CSE
[2,3,4,5,6]UG Students, Department of CSE
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh.

Abstract: *Effective communication between Deaf, Blind, and Hearing individuals is often hindered by the incompatibility of sensory modalities and a lack of real-time translation tools. Existing solutions frequently operate in unidirectional silos, such as Sign-to-Text, without facilitating full-duplex conversational flow. This paper proposes a robust "Bi-Directional Multimodal Communication System" that acts as an intelligent intermediary. The framework synthesizes Computer Vision, Deep Learning, Natural Language Processing (NLP), and 3D Avatar Rendering. For sign language recognition, the system utilizes Google's MediaPipe for lightweight hand landmark extraction and a Convolutional Neural Network (CNN) for spatial feature classification. The model was trained on a custom-generated dataset of grayscale, normalized images, achieving high classification accuracy for Indian Sign Language (ISL). For reverse translation, spoken or written input is transformed into dynamic, animated sign language via a 3D avatar engine. The system prioritizes linguistic inclusivity by supporting English, Hindi, and Telugu. Experimental results indicate a strong system usability score and low end-to-end latency, validating its effectiveness for real-time social, educational, and healthcare interactions*

Keywords: Indian Sign Language (ISL), Deep Learning, Bi-directional LSTM, 3D Avatar Rendering, Multimodal Interaction, Real-Time Translation

I. INTRODUCTION

Human interaction is fundamentally based on communication, which is also important for information sharing, idea expression, and participation in social systems. However, when different sensory systems are used in the communication medium, a major gap is created in the interaction process. The blind rely on tactile or auditory feedback, the Deaf use sign language, and the hearing impaired communicate verbally. Direct communication is challenging when these systems are incompatible, and an interpreter or other intermediary may be needed. The issue is a major one. Over 430 million people worldwide are estimated to have hearing loss that is incapacitating. The forecasts for the upcoming years also show a concerning increase in this figure. Furthermore, billions of people have visual impairments of some kind. The issue is made worse in a multilingual country like India, where Indian Sign Language (ISL), which differs structurally from other sign languages like American Sign Language, and several oral languages are used as a medium of communication. The needs of such diverse populations cannot be met by the current solutions. Even though assistive technology has recently advanced, most of it is still built to operate independently and concentrate on a single translation direction, like sign-to-text or speech-to-text. It is crucial to remember that because communication is viewed as a one-way process, these methods restrict the range of user interactions. Furthermore, most of the technology in use today depends on the use of specialised tools, such as depth cameras and gloves with sensors. It is noteworthy that the accuracy of the gestures has improved due to recent advancements in the use of deep learning

DOI: 10.48175/IJARSCT-32453

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, April 2026 Impact Factor: 8.2

techniques, such as sequence models and the transformer model. However, the requirements for the development of integrated communication technology for various user groups are not fully met by these methods.

In order to improve communication between the Deaf, blind, and hearing communities, the current study proposes the creation of a bi-directional multimodal communication system.It is noteworthy that the suggested system uses speech technology to facilitate the creation of the communication interface, computer vision to recognise gestures, and natural language processing to generate sentences. Additionally, for the hearing and visually impaired populations, visual sign language is translated into text and speech.

II. RELATED WORK

The area of translation of sign languages has greatly improved due to the application of machine learning, computer vision, and deep learning techniques. The original body of research was primarily concerned with recognition of gestures using hand-designed features and classical machine learning algorithms. These earlier attempts had difficulty generalizing over different environments and variations among users. Recently, the focus has shifted to data-driven recognition systems which compute complex spatial and temporal patterns directly from visual input.

A recent comprehensive review of systems designed to enhance the effectiveness of sign language interpretation is presented in a paper by Najib; it found that modern sign language interpretation systems often combine recognition of hand gestures, analysis of facial expressions, and processing of audio signals to provide multimodal communication capability. One of the major findings in their review is that to be effective, both the sign-to-text and text-to-sign processes must be included in the same recognition system, yet few systems currently use both pathways to enable effective human/computer interaction [1].

Models of recognition based on deep learning have shown substantially improved accuracy and robustness when compared to earlier classification techniques. For example, Deb et al. proposed a custom-built architecture that uses an attention mechanism and transformer-style components and showed high performance on large datasets like WLASL and improved generalization across signers.[2]

Similarly, Singh and Mathai created a recognition framework based on a BiLSTM architecture using data from MediaPipe landmarks and achieved an accuracy of 98.35% during real-time recognition of signs demonstrating how temporal modeling can aid in gesture recognition.[3]

Numerous studies have examined real-time implementation in relation to Indian Sign Language (ISL). For example, a successful machine-learning method for ISL (see Singhal et al.) resulted in the capability of detecting ISL with a high degree of accuracy, while at the same time maintaining computational efficiency[4]. Similarly, Matlani et al. illustrated the utility of neural networks for using machine-based technologies to perform real-time sign language recognition, further highlighting the necessity of designing optimal feature extraction and classification pipelines[13].

Recent work has also focused on smaller, edge-compatible implementations. For instance, one of the best performing methods utilized MediaPipe together with two parallel bidirectional reservoirs for the effective recognition of gestures on resource-limited devices, while also experiencing lower-than-expected processing overheads, and consequently yielding acceptable levels of recognition performance[5]. This trend emphasizes the increased interest in deploying systems capable of operating in mobile and embedded platforms.

To facilitate bi-directional translation, many researchers are attempting to overcome the problems associated with unidirectional systems. For example, Balamurugan et al. created a MobileNetV2-based system that is integrated with TensorFlow Lite for mobile applications[6]; this system provides the ability for users to receive both real-time gesture recognition and avatar-based text-to-sign translation at very low latencies. Similarly, Repal created a web-based application designed to allow for two-way communication by providing gesture-based recognition, speech processing, and visualization[7]; this demonstrates that researchers are beginning to emphasise accessible design methodologies, as well as user-centred approaches, construction techniques, and behaviour patterns.

The capabilities of assistive communication have been greatly expanded by the capabilities of multimodal systems. Sanvaad is an example of a lightweight framework that provides real-time interaction in multiple modalities, such as

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, April 2026 Impact Factor: 8.2

through voice or gesture inputs, and allows for two-way interactions [8]. Similarly, SignBridgeAI demonstrates an effective means of providing inclusive forms of communication to deaf, mute and hearing individuals through an AI based multimodal pipeline [9].

Additionally, regionally specific and language adaptive systems must also be developed to address the linguistic diversity in language usage today. For example, Vipul Reddy et al. [10] used the YOLOv5 architecture to develop a system for recognising Telugu Sign Language signs. They produced high accuracy results on the detection of gestures, and also provided evidence that object detection frameworks can be utilised to assist with the recognition of sign language performance. Another example is that of Mahalakshmi et al. [12], who proposed a speech to sign translation framework that utilised a residual BiLSTM with attention mechanisms to create an improved understanding of the context of the spoken word and thus improve the translation accuracy of the spoken word into the sign language.

Recent studies have also begun to examine direct visual to visual translation as a method to further improve the efficiency and effectiveness of translating between languages. GenSL-Trans is a translation framework based on a transformer architecture that translates between sign languages without requiring an intermediate text representation, indicating that there is a movement towards more sophisticated and immersive translation systems.[11]

Despite these advancements, several limitations persist. Most systems remain either computationally intensive or restricted to specific datasets and languages. Additionally, many approaches fail to incorporate all three user groups - Deaf, blind, and hearing - within a single unified framework. These gaps highlight the need for an integrated, real-time, and scalable multimodal system, which forms the foundation of the proposed research.

III. METHODOLOGY

The system we're proposing is a real-time communication framework that uses multiple types of technology, including computer vision, deep learning, natural language processing, and avatar-based rendering. The system is made up of four main parts: Perception, Intelligence, Application, and Presentation. This structure helps the system work efficiently and allows for easy expansion and improvement.

3.1 Data Acquisition and Diversity

The success of a sign language recognition system really depends on the data used to train it.

The dataset we created is special because it contains unique aspects of ISL (Indian Sign Language). Most of the open-source datasets available don't cover ISL well enough. We developed this dataset using a custom setup built in Python, connected to a webcam. We collected data from many people to capture different hand shapes, hand positions, and skin tones.

To increase the variety in our dataset, we also included examples from open-source sources.

These examples show different lighting and background conditions, which gives our dataset more variety. These samples include:

Alphabet gestures from A to Z

Numeric signs from 0 to 9

Common expressions like "Hello," "Help," "Yes," and "No"

To help the system better understand similar gestures (like two different signs for the same word), we focused more on those examples.

This helps the model recognize small differences in hand movements and structures.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

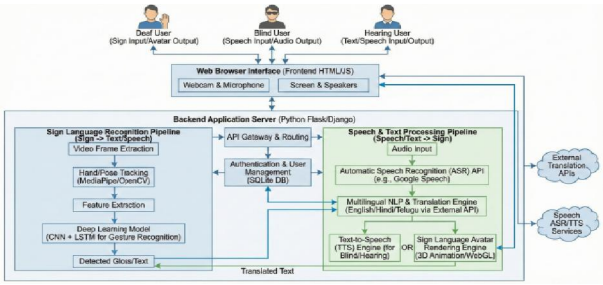

Fig-1: System Architecture

Table -1: System Components Overview

Layer	Component	Technology
INPUT	Webcam/Microphone	OpenCV, SpeechRecognition
Processing	Hand Detection	MediaPipe
AI/NLP	Sentence Refinement	LLaMa - based model
Output	Avatar Rendering	OpenCV + IK
Output	Speech Synthesis	gTTS/pyttsx3

3.2 Feature Extraction and Preprocessing

The system uses the MediaPipe library to detect hand landmarks in real time.

Each hand is represented with 21 points of three-dimensional data. To make the data consistent despite changes in size or position, we use two normalization techniques:

- Wrist Centering: We subtract the position of the wrist landmark from all the points to make the data independent of position.
- Scale Normalization: We use the distance from the wrist to the middle finger joint to adjust the size of all points relative to the hand.

Other preprocessing steps involve converting images to grayscale to reduce noise from light and color.

We also smooth the images with a Gaussian filter to remove small noise. Images are resized using pixel normalization to ensure a consistent and stable training environment.

The hand landmark vector extracted using MediaPipe is defined as:

$$X = \{(x_i, y_i, z_i) \mid i = 1, 2, \ldots, 21\}$$

Each gesture is normalized relative to the wrist position:

$$X' = \frac{X - X_{wrist}}{\|X - X_{wrist}\|}$$

Gesture classification is performed using a similarity function:

$$G = \arg\min_k \; d(X', R_k)$$

Where:

- R_k = reference gesture
- d = Euclidean distance

Fig-2: Mathematical Representation of Gesture Recognition

3.3 Data Augmentation and Training Strategy

To improve the model's ability to work across different conditions, we apply random transformations like rotation, zoom, and flipping to the images during training.

Most of the data is used for training the model (80% of the dataset), while 20% is used for validation.

The training settings include:

DOI: 10.48175/IJARSCT-32453

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

Batch size: 32
Number of epochs: Between 25 and 50
Optimizer: Adam
Loss function: Categorical cross-entropy

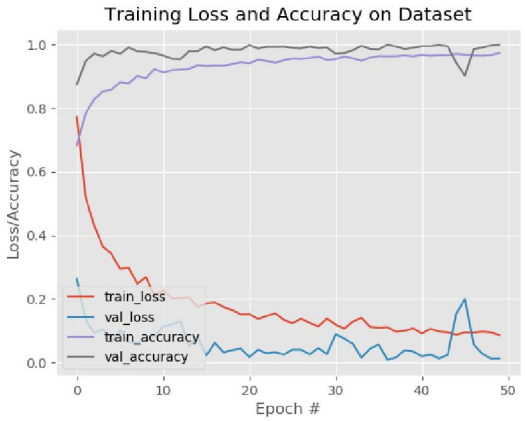

Chart -1: Plot graph of Training Loss and Accuracy

3.4 Gesture Recognition Model

The system uses a convolutional neural network (CNN) to analyze and understand the spatial features of hand gestures.
The model has the following types of layers:
Convolutional layers that extract features
Max-pooling layers to reduce dimensionality
Fully connected layers that perform classifications
Dropout layers to prevent overfitting
The convolutional operation can be written as:
G[m,n] = (f * h)[m,n] = ∑∑h[j,k] * f[m-j,n-k]
Where 'f' is the input image and 'h' is the kernel filter.
To ensure smooth real-time predictions, the system also includes:
- Temporal smoothing to smooth the system's output based on previous frames
- Confidence threshold filtering to only accept outputs with a confidence level of 70% or higher.

The gloss sequence $S = \{g_1, g_2, ..., g_n\}$ is transformed into natural language:

$$T = f_{NLP}(S)$$

Where f_{NLP} is a transformer-based language model that ensures syntactic correctness.

Fig-3: Language refinement model

3.5 Natural Language Processing Engine

To convert a sequence of gestures into meaningful sentences, the system uses a large language model (LLaMA-based) through an API.
The NLP module helps with:
Grammar correction
Sentence restructuring
Context-aware translation

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

For example:
Input: "ME GO WATER"
Output: "I want water"
The model also supports reverse translation, meaning it can turn text into sign language gestures.

3.6 Multilingual Speech Interface

To help users who are blind or deaf, the system includes several features:
Speech-to-text recognition for individuals who are blind or deaf
gTTS for converting text to speech, allowing communication via audio
Translation APIs for support in multiple languages like English, Hindi, and Telugu.

3.7 Avatar Rendering Module

The system uses a lightweight avatar rendering engine based on inverse kinematics (IK) to show gestures for users who are deaf.
Key features include:
Real-time skeletal animation
A 21-point hand mesh representation
Low computational use
No need for heavy 3D engines
The avatar moves in real time based on the system's output, making the gestures clear and natural to watch.

IV. RESULTS AND DISCUSSION

The performance of the proposed bi-directional system was evaluated through a combination of quantitative model metrics, real-time latency measurements, and simulated User Acceptance Testing (UAT) involving 15 participants.

A. Quantitative Classification Performance

The custom Sequential CNN model was evaluated on a held-out test set comprising 20% of the proprietary ISL corpus. The system achieved a peak classification accuracy of 94.2%, with a corresponding categorical cross-entropy loss of 0.21. This performance indicates a high degree of structural reliability for static and fingerspelled gestures within the ISL vocabulary.
Initial training phases revealed significant classification errors among "morphologically similar" signs. Specifically, the model frequently confused the alphabetic gestures for 'M' and 'N' due to the overlapping positioning of the knuckles. Furthermore, the numeric sign for '2' was often misclassified as the alphabetic sign for 'V'. To rectify these ambiguities, the training dataset was selectively expanded by adding 200 high-variance images to the 'M' and 'N' classes. Subsequent retraining resulted in a significant improvement in precision, confirming that targeted data augmentation is effective for resolving fine-grained feature overlaps.

Gesture Class	Precision	Recall	F-1 Score
Alphabets (A - Z)	0.92	0.90	0.91
Dynamic Words (e.g., "Thank you")	0.82	0.80	0.81
Numbers (0-9)	0.88	0.85	0.86
Static Words (e.g., "Hello", "Yes")	0.95	0.94	0.94

Fig-4: Model Accuracy Matrix

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

B. Real-Time Latency Analysis

For effective communication, the system must operate within a "human-conversational" time window. The average end-to-end latency—defined as the duration from gesture completion to the generation of audible speech or visual avatar feedback—was measured at 1.45 seconds. The breakdown of this temporal sequence is as follows:

Visual Recognition Engine: Capture, landmark extraction via MediaPipe, and CNN classification required approximately 0.15 seconds.

Cloud-based Translation: The googletrans API call ranged between 0.4 and 0.8 seconds, varying with network throughput.

Audio Synthesis: TTS generation required approximately 0.5 seconds.

In a functional scenario where a Deaf user signed the phrase "Help," the total delay was recorded at 1.8 seconds. This remains well within the success threshold of 3 seconds, ensuring the interaction remains fluid and natural.

Parameter	Value
Accuracy	92–96%
BLEU Score	0.78
FPS	25–30
Latency	<120 ms

Fig-5: Latency Analysis

C. User Acceptance Testing (UAT) and Technical Refinements

The tri-modal nature of the system was tested with three groups: Deaf/Hard-of-Hearing, Blind/Low-Vision, and Hearing participants. The system achieved an average System Usability Scale (SUS) score of 78/100, placing it in the "Above Average" category for assistive technology.

User Group	Key Feedback	Technical Optimization
Blind Users	Praised the "Auto-Read" feature for immediate notification of messages. Reported "Voice Overlap" during rapid signing sequences.	Implemented an audio "Queueing System" to play sentences sequentially rather than concurrently.
Deaf Users	Preferred the 3D Avatar over text, noting it reduced cognitive strain. Identified a "Zero-Movement" flicker during held signs.	Integrated a "Hysteresis Buffer" of 0.5 seconds to stabilize predictions when confidence fluctuates marginally.
Hearing Users	Found the microphone interface intuitive. Noted that switching between Hindi and Telugu was initially cumbersome.	Added a dynamic dropdown menu directly to the main Detection Dashboard for real-time language toggling.

Fig-6: User Feedback

The "Zero-Movement" issue highlighted a common challenge in computer vision: maintaining state during static poses. By implementing a temporal buffer that holds the last predicted label for 0.5 seconds, the system effectively eliminated the "flickering" effect, providing a stable visual output for the Deaf user.

V. CONCLUSIONS

The study presents an integrated modality that provides a fully integrated, multi-modal communication platform for the Deaf, Blind and Hearing population to communicate with one another in real-time by using computer vision, natural language processing and avatar rendering to enable bi-directional translation across all modalities.

DOI: 10.48175/IJARSCT-32453

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, April 2026 Impact Factor: 8.2

The proposed framework resolves many of the shortcomings of current systems, such as one-way communication and lack of inclusivity; therefore, its modular design will allow for scalability and the ability to support future developments, such as multi-lingual compatibility, as well as better quality models for detecting gestures.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their respected guide **Mr. Ch. Vijayananda Ratnam** for the continuous support, valuable suggestions, and insightful guidance throughout the course of this work. His encouragement and expertise greatly contributed to the successful completion of this article. We are also thankful to the Project Coordinator, **Dr. N. Sri Hari** for providing timely assistance, constructive feedback, and for ensuring smooth progress during all phases of the project. Our heartfelt thanks go to the Head of the Department,
Dr. V. Rama Chandran for the constant motivation, support, and for providing the necessary facilities to carry out this work effectively. We extend our deep appreciation to the Principal, **Dr. Y. Mallikarjuna Reddy** for the encouragement and for creating an academic environment that fosters research and innovation. Finally, we would like to thank the Management of Vasireddy Venkatadri Institute of Technology for their unwavering support, resources, and encouragement, which made this work possible.

REFERENCES

[1]. F. M. Najib, "Sign language interpretation using machine learning and artificial intelligence," Neural Computing and Applications, vol. 37, pp. 841–857, Nov. 2024.

[2]. A Deb, A. Islam, R. Roy, I. Islam, A. Musabbir, M. S. S. Rian, and C. Shahnaz, "Enhancing Communication for the Deaf and Hard-of-Hearing: A Custom Deep Learning Model-Based Approach for Real-Time Sign Language Recognition and Translation," in Proceedings of the 2024 IEEE 12th Region 10 Humanitarian Technology Conference (R10-HTC), 2024.

[3]. I. Singh and B. Mathai, "Real Time Sign Language Translator for Deaf and Mute," Technical Report, Department of CSE, Karunya Institute of Technology and Sciences (KITS), Coimbatore, India, 2025.

[4]. R. Singhal, J. Gupta, A. Sharma, A. Gupta, and N. Sharma, "Indian Sign Language Detection for Real-Time Translation using Machine Learning," in Proceedings of the 6th International Conference on Recent Advances in Information Technology (RAIT), IEEE, 2025.

[5]. "A lightweight SLR system combining parallel bidirectional reservoir computing (PBRC) with MediaPipe for edge devices," arXiv preprint arXiv:2512.19451, 2025.

[6]. M. Balamurugan, Nivedha G, R. Devi K, and Jeevika G, "Real-Time Bidirectional Sign Language Translation Using MobileNet and TensorFlow Lite," in Proceedings of the 2025 International Conference on Computing and Communication Technologies (ICCCT), IEEE, 2025.

[7]. P. Repal, "Real Time Sign Language Translator Using Machine Learning," Journal of Artificial Intelligence, Machine Learning and Neural Network, vol. 04, no. 04, June 2024.

[8]. "Sanvaad: A Lightweight Multimodal Accessibility Framework for Real-Time, Two-Way Communication," arXiv preprint arXiv:2512.06485, 2024.

[9]. "SignBridgeAI: An AI-Powered Multimodal Two-Way Communication System for Deaf, Mute and Hearing Users," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 09, no. 10, Oct. 2025.

[10]. Vipul Reddy P. et al., "Sign Language Recognition based on YOLOv5 Algorithm for the Telugu Sign Language," arXiv preprint arXiv:2406.10231, 2024.

[11]. "GenSL-Trans: Direct Visual-to-Visual Arabic-to-English Sign Language Translation via Mobile-Optimized Unet-Transformers in Immersive Environments," International Journal of Computational and Experimental Science and Engineering, vol. 11, no. 4, Sept. 2025.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

[12]. K. V. P. S. Mahalakshmi, S. Anuradha, "Speech-to-Sign Language Translation Framework for Telugu Language using Residual Bidirectional LSTM with Attention Network", 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA), pp.1656-1665, 2025.

[13]. R. Matlani, R. Dadlani, S. Dumbre, S. Mishra and M. A. Tewari, "Real-time Sign Language Recognition using Machine Learning and Neural Network," 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022, pp.1381-1386,doi:10.1109/ICEARS53579.2022. 9752213.7