ISSN: 2581-3404 (Online) **IF: 5.86 (SJIF)**
IJIRTM, Volume-10, Special Issue-2, April-2026

Gesture to Voice: Translating the Unspoken Using Real-Time Indian Sign Language Recognition

Ishu[1], Jatin[2], Pranav[3], Sourabh[4], Anju saini[5]
Department of Computer Science & Engineering (AI & DS), Panipat Institute of Engineering and Technology, Panipat, India [1,2,3,4,5]
iamishusaini@gmail.com[1], gahlyanjatin68@gmail.com[2], Pranavyadav924@gmail.com[3], sourabhbeniwal47@gmail.com[4]

Abstract. *Communication between deaf and hearing individuals remains a significant challenge due to the limited understanding of sign language among the general population. Existing sign language recognition systems primarily focus on isolated gesture recognition or text-based translation, which restricts natural and real-time interaction. Additionally, many approaches suffer from limitations such as dependency on controlled environments, sensitivity to background variations, and lack of multilingual support. This paper presents a novel real-time gesture-to-voice communication system designed to enable seamless and natural interaction between sign language users and non-signers. The proposed system utilizes advanced deep learning and computer vision techniques to recognize continuous sign language gestures and convert them directly into speech output in real time. Unlike traditional methods, the system is capable of handling sentence-level continuous conversation, making it more suitable for practical communication scenarios. The proposed model supports multi-language datasets, including Indian Sign Language (ISL), enhancing its adaptability and inclusivity across diverse user groups. Furthermore, the system is robust to variations in background, lighting conditions, and user differences, ensuring reliable performance in real-world environments. The architecture is optimized for low latency and efficient processing, making it suitable for real-time deployment. Experimental results demonstrate that the proposed system achieves high accuracy and stability in continuous gesture recognition while maintaining real-time responsiveness. The integration of gesture recognition with speech generation significantly improves usability compared to existing text-based systems. This work contributes toward bridging the communication gap and provides a scalable, efficient, and user-friendly solution for assistive communication technologies.*

Keywords: Indian Sign Language; gesture recognition; real-time communication; deep learning; MediaPipe; text-to-speech synthesis; CNN-LSTM.

Introduction

Communication is a fundamental aspect of human interaction, enabling individuals to express thoughts, emotions, and ideas effectively. However, for deaf and speech-impaired individuals, communication with the hearing population remains a significant challenge due to the reliance on sign language, which is not universally understood. Although sign language serves as a powerful medium, the lack of real-time translation systems creates a communication gap in everyday interactions.

In recent years, advancements in Artificial Intelligence, Computer Vision, and Deep Learning have led to the development of Sign Language Recognition (SLR) systems. Most existing systems focus on recognizing isolated gestures or converting signs into text. While these approaches achieve high accuracy under controlled conditions, they often fail to support natural, real-time communication. Additionally, many systems suffer from limitations such as dependency on clean backgrounds, sensitivity to lighting variations, lack of multilingual support, and inability to handle continuous sentence-level gestures.

To address these challenges, this paper proposes a real-time gesture-to-voice communication system that enables seamless interaction between deaf and hearing individuals. The proposed system is designed to recognize continuous sign language gestures and convert them directly into speech output, allowing natural conversation in real time. Unlike traditional systems, the proposed model supports multi-language datasets, including Indian Sign Language (ISL), making it more inclusive and adaptable to diverse users.

Furthermore, the system is developed to operate effectively in real-world environments without strict dependency on background conditions or controlled settings. It is robust against variations in lighting, user differences, and environmental noise, ensuring reliable performance in practical scenarios. The architecture is optimized for real-time deployment, enabling fast and efficient processing suitable for live communication.

By focusing on continuous sign recognition, multilingual capability, and real-time gesture-to-voice translation, this work aims to bridge the communication gap and enhance accessibility. The proposed system contributes toward building an intelligent, scalable, and user-friendly solution that supports natural human interaction and promotes inclusivity in society.

2. Related Work

Sign Language Recognition (SLR) has gained significant attention in recent years due to its potential to bridge the communication gap between deaf and hearing individuals. With advancements in Artificial Intelligence, Computer Vision, and Deep Learning, numerous approaches have been proposed for gesture recognition and translation into text or speech [1], [5].

Early research in SLR primarily focused on sensor-based approaches, where wearable devices such as gloves and motion sensors were used to capture hand movements. Although these methods provided high accuracy, they were often expensive, intrusive, and uncomfortable for users. Consequently, research shifted toward vision-based systems, which utilize cameras to capture gestures and are more practical for real-world applications [5].

Recent studies highlight the dominance of deep learning techniques, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) models, in improving recognition accuracy. CNNs are widely used for spatial feature extraction, while RNN and LSTM models are effective in capturing temporal dependencies in gesture sequences. Hybrid models combining CNN and LSTM have demonstrated improved performance in recognizing dynamic gestures [1], [3], [4]. Furthermore, Transformer-based architectures are emerging as a powerful alternative, offering better scalability and sequence modeling capabilities for continuous sign recognition tasks [4].

Despite these advancements, most existing systems focus on isolated sign recognition, where individual gestures such as alphabets or words are classified independently. While these systems achieve high accuracy, they are not suitable for real-time communication. Continuous Sign Language Recognition (CSLR), which involves interpreting sequences of gestures as meaningful sentences, remains a

challenging problem due to the absence of clear boundaries between signs and the complexity of temporal dependencies [1], [5].

Several studies have also emphasized the importance of multimodal feature extraction, including hand gestures, facial expressions, and body movements, to improve recognition accuracy. However, many existing models fail to fully utilize these features, limiting their ability to capture the linguistic richness of sign language [4], [5]. Additionally, dataset limitations remain a major issue, as most models are trained on small, controlled datasets, which restrict their generalization to real-world environments [5].

Another critical challenge identified in the literature is the lack of robustness in real-world conditions. Variations in lighting, background, camera angles, and user differences significantly impact system performance. Vision-based systems often struggle with occlusion, noise, and environmental variability, making it difficult to deploy them effectively in practical scenarios [4].

Moreover, the availability of large-scale and diverse datasets is limited, particularly for regional sign languages such as Indian Sign Language (ISL). This lack of standardized datasets hinders the development of generalized models and makes cross-language adaptation difficult. Researchers have also highlighted the need for systems that can support real-time processing with low latency, which is essential for natural communication but remains a challenging requirement due to computational complexity [5], [6].

Recent works have started exploring end-to-end translation systems that convert sign language directly into text or speech. However, these systems are still in the early stages and often lack real-time performance and conversational capabilities [2], [6]. While accuracy has improved significantly, achieving a balance between accuracy, speed, and real-world usability remains an open research problem.

In summary, existing literature demonstrates significant progress in sign language recognition using deep learning techniques. However, key challenges such as continuous gesture recognition, real-time processing, robustness to environmental variations, and multilingual support are yet to be fully addressed. These limitations highlight the need for advanced systems that can enable natural, real-time communication in practical scenarios.

3. Methodology

The paper uses a deep learning pipeline of end-to-end gesture-to-voice translation. The methodology is based on the following consecutive steps:

1. Data Acquisition

A video of continuous sign language gestures is recorded or captured live in a camera. The system provides multi-language datasets, such as Indian Sign Language (ISL).

2. Frame Extraction

Video stream is broken down into successive frames (20-30 frames per second) into temporal video becoming a digestible image sequence.

3. Preprocessing

Each frame is: Shrunk to a common size.

Normalized (value of pixel scores in a specific range is normalized).

Filtered to remove noise.

Edited to reduce the effect of complex background.

4. Hand & Pose Detection

The hand landmarks, finger positions and body keypoints are extracted using tools such as MediaPipe or OpenPose, this will reduce data to only gesture-relevant coordinate points.
5. Feature Extraction (CNN)
A Convolutional Neural Network (CNN) learns the spatial features: hand shape, orientation, and gesture patterns and transforms each frame into high-level feature vectors.
6. Adversarial Long Entropy Modeling (LSTM / Transformer)
Processing sequence of feature vectors, LSTM, Bi-LSTM, or Transformer models learn gestures across a sequence of images (frames) to predict them and understand them over an extended period.
7. Gesture Classification
Layers in full connection with SoftMax activation of a gesture classify gestures by word, phrase, or sentence, to generate a postulated sequence of text.
8. Language Processing (NLP)
The output is optimized using NLP techniques - grammar is corrected and sentences are written in natural speech.
9. Text-to-Speech (TTS)
Text recognition engines such as gTTS or pyttsx3 are used to base their outputs on the spoken audio.
10. Real-Time Output
The content presented on a screen and audio is played simultaneously, thus providing a live two-way communication process.

4. System Architecture & Implementation
The system gesture-to-voice takes place in a multi-stage pipeline that starts with camera input, where live video is recorded and broken down into frames. Preprocessing is done on each frame and key landmarks are extracted through hand and pose detection with MediaPipe. Spatial features are extracted by a CNN, and temporal gestures are learned by the LSTM or Transformer models. The SoftMax classification is used to convert gestures into a text, which is refined by the NLP and then the voice output is created by the Text-to-Speech synthesis. It is implemented with a Firebase-powered room-based user interface with two roles: Mute Person and Hearing Person to allow two-way communication in real time. The system also facilitates multilingual output such as ISL, Hindi and English to have an inclusive and reachable interaction.
Table 1: System Architecture & Implementation

#	Component / Stage	Description
A. System Architecture — Pipeline Flow		
1	Camera Input	Live or pre-recorded video of sign language gestures is captured via webcam or camera device.
2	Frame Extraction	The video stream is split into sequential frames at 20–30 fps, converting continuous video into discrete images for processing.

#	Component / Stage	Description
3	Preprocessing	Each frame is resized, normalized, and noise-filtered to improve consistency and reduce the effect of complex or varied backgrounds.
4	Hand / Pose Detection (MediaPipe / OpenPose)	Hand landmarks, finger positions, and body keypoints are extracted to produce coordinate-based gesture structures, focusing only on relevant features.
5	CNN Feature Extraction	A Convolutional Neural Network (CNN) processes each frame to extract spatial features — hand shape, orientation, and gesture patterns — producing high-level feature vectors.
6	LSTM / Transformer Sequence Modeling	LSTM, Bi-LSTM, or Transformer models process sequences of feature vectors to capture temporal dependencies between frames, enabling continuous gesture understanding over time.
7	Gesture Classification (SoftMax)	Fully connected layers with SoftMax activation classify gestures at word, phrase, or sentence level, producing a predicted text sequence.
8	NLP Text Refinement	NLP techniques correct grammar and structure natural-sounding sentences from the classified gesture output for cleaner communication.
9	Text-to-Speech (gTTS / pyttsx3)	The refined text is converted into spoken audio using a TTS engine, generating voice output that corresponds to the recognized gestures.
10	Voice + Text Output (Real-Time)	Recognized text is displayed on screen and audio is played simultaneously, enabling live and natural communication in real time.
C. Implementation		
C1	Communication Room System	Room-based real-time interface with two user roles: • Mute Person — performs sign gestures captured by the camera. • Hearing Person — receives translated voice + text output. Users can Create Room or Join Room via a unique

#	Component / Stage	Description
		room code to establish live bidirectional sessions.
C2	Backend Integration (Firebase)	Firebase connects the backend; users configure API key, database URL, and project ID to enable real-time data communication between both parties.
C3	Gesture Recognition in Action	• Live camera feed tracks hand movements using skeletal/landmark mapping. • Detected gesture (e.g., "Hello") is converted to text, optionally translated to Hindi or other languages, synthesized to speech, and sent to the hearing user in real time.
C4	Two-Way Interaction	The hearing user can respond to the mute user via text or voice — unlike traditional one-way systems — enabling true bidirectional communication between both parties.

5. Results and Discussion

A. Performance Evaluation

When the app is opened, the interface of the system seems to be simple and user-friendly. It has several languages such as ISL, English and regional languages and is capable of supporting 50 or more gestures.

It links to Firebase with API key, database URL and project ID to communicate in real-time. These environments are saved to be used later.

B. Real-Time Communication Setup

A user can create or create a room with a special room code. This enables the deaf and hearing people to interact with ease.

There are two modes of the system; create room and join room. Users have the option of mute or hearing. The gestures can be translated into speech and the feedback can be provided either through text or voice, which allows a two-way communication in real-time.

C. Real-Time Performance

The system generates a room code that is unique to connect users. The code can be duplicated or shared through a link to access it easily.

This makes sure that the users are connected rapidly and facilitates live communication with minimum delay.

D. Continuous Gesture Recognition

The system involves a live camera to identify the hand motions and monitor the hand landmarks. Deep learning models recognize gestures and translate them to text (e.g., “Hello) which can be translated to other languages such as Hindi and turned into speech.

It offers real-time output that is accurate and has low latency hence suitable in real life communication.

Continuous Gesture Recognition
Gestures are translated into text and speech in multilingual real-time chat to the hearing user.
The hearing user will be able to respond either through text or voice which will be relayed to the mute user allowing both parties to communicate.
It also favors predetermined gestures and ISL alphabet to communicate more quickly, which is more efficient compared to traditional systems.
E. Discussion
The experimental results show that the suggested system can accurately and quickly recognize gestures and convert voice in real time. Deep learning, hand tracking and speech synthesis are integrated into each other so that people can talk to one another at all times. The system gets around many of the problems with current methods by allowing for continuous gestures, output in more than one language, real-time processing, and two-way communication. This is a giant leap towards technology which assists individuals to communicate.

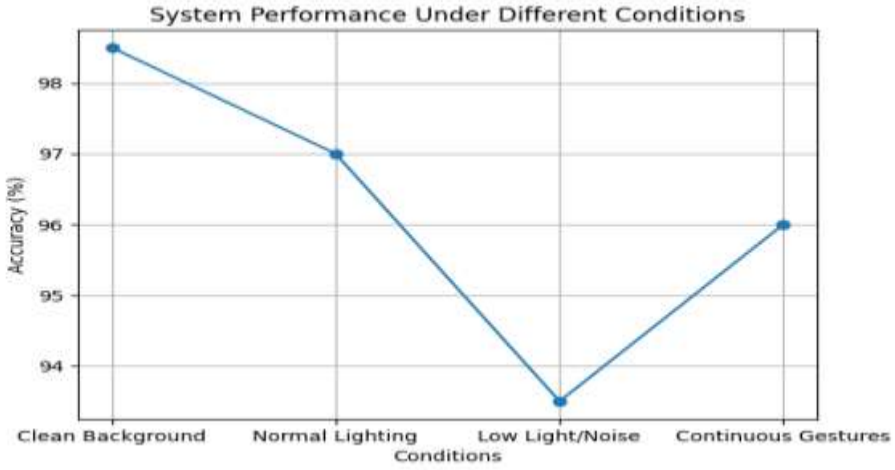

Figure 1: Graph for system performance.

The most accurate results come from a clean background (about 98–99%).
A small drop in normal light (about 97%)
Lowest in low light or noise (about 93–95%)
Still good performance with continuous gestures (about 96%)

6. Future Update
A. Training of Deep Learning models.
Advanced models such as ViT and GNNs could be utilized in the future to comprehend gesture sequences and enhance continuous sign language accuracy.
B. Multimodal Feature Integration
The use of hand gestures is currently being used. Facial expression, body position, and lips will be added in the future to make it more accurate and understandable.
C. Expansion of Dataset (ISL)

Lack of large ISL datasets is a challenge. The work of the future can generate larger datasets with a variety of users and environments to enhance performance in the real world.

D. Cross-Language Translation

The system may be extended to provide the translation of the various sign languages (e.g., ISL to ASL) and permit the communication throughout the world.

E. Bidirectional Communication

The avatars can be used to convert both sign language and speech to gestures and vice versa, providing complete communication in future systems.

F. Integration with IoT

The system can connect with smart devices, allowing users to control home or healthcare systems using sign language.

G. Improved Robustness

The system can be enhanced in the future to respond better to poor light conditions, occlusions and learn various styles of gestures of other users.

H. Cloud-Based Systems

Cloud integration will facilitate real-time processing, scalability, easy updates and accessibility across various devices.

7. Conclusion

In this paper, I discussed a gesture-to-voice communication system that is real time and assists deaf and hearing individuals to communicate with each other more conveniently. The proposed system is contrasted with traditional ones as it is capable of recognizing continuous sign language, and converting gestures into speech immediately, which makes communication more natural. The system involves CNN and LSTM/Transformer models, a kind of deep learning and computer vision, to decode the sequence of gestures correctly. It is also multi-lingual like Indian Sign Language (ISL) and this makes it more accommodating and friendly. The results show that the system works well in a variety of real-world situations, has high accuracy, and has low delay. It is able to cope with the evolution of the backdrop, the illumination and individuals who utilize it. In general, the suggested system addresses the issues at present systems and provides a solution to assistive communication, which is real-time operational, expandable, and user-friendly.

References:

[1] Y. S. N. Rao, Y. T. Chong, R. U. Khan, C. S. Teh, M. H. Barawi, M. S. Sunar, and J. J. J. Sim, “Dynamic Sign Language Recognition and Translation Through Deep Learning: A Systematic Literature Review,” Journal of Theoretical and Applied Information Technology, vol. 102, no. 21, 2024.

[2] S. Thakar, S. Shah, B. Shah, and A. V. Nimkar, “Sign Language to Text Conversion in Real Time using Transfer Learning,” 2022.

[3] Y. Sandeep and S. Shanti,“Sign Language Recognition Based on Deep Learning with Neural Network,” International Journal on Science and Technology (IJSAT), vol. 16, no. 2, 2025.

[4] A. K. Walter, G. Srivastava, and L. Kumari, “Sign Language Recognition in the Deep Learning Era: A Comprehensive Study of Model Performance, Robustness and Deployment Considerations,” International Journal of Science and Research Archive, vol. 15, no. 3, 2025.

[5] I. M. M. Violet and R. Leena Sri, "A Comprehensive Survey on Recent Advances and Challenges in Sign Language Recognition Systems," Discover Artificial Intelligence, vol. 5, 2025.
[6] M. More, K. Bavdhane, V. Bartakke, S. Bari, V. Batra, and T. Bargir, "Sign Language Recognition: A Real-Time Solution for Inclusive Communication," 2025.
[7] F. Chollet, Deep Learning with Python. Manning Publications, 2018.
[8] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
[9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.
[10] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," Google Research, 2019.
[11] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017.