International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

AI-Based Real-Time Speech-to-Sign Language Translation System for Assistive Communication

Bhukya Siddhu
M.Sc Artificial Intelligence & Data Science
Department of Computer Science & AI
Central University of Andhra Pradesh
Reg. No: 24MAI05
Email: siddhu70133@gmail.com

Mr. Y. Dayanand Kumar
Assistant Professor
Department of Computer Science & AI
Central University of Andhra Pradesh
Email: dayanandkumar@cuap.edu.in

***Abstract*—**This paper presents the design and development of an Artificial Intelligence (AI)-based, real-time Speech-to- Indian Sign Language (ISL) Translation System that bridges the communication gap between the hearing population and the deaf and hard-of-hearing community. The proposed system captures spoken audio via a standard microphone, converts it to text using cloud-based Automatic Speech Recognition (ASR), applies Natural Language Processing (NLP) techniques including tokenization, stop-word removal, and sentence sim- plification, and maps the resulting tokens to pre-recorded ISL GIF animations for sequential visual display. Implemented as a software-only Flask web application, the system requires no specialized hardware such as sensor gloves or motion-capture devices, enabling deployment on any standard browser-equipped device. A modular pipeline architecture ensures that individual components can be independently upgraded without disrupting the overall system. Experimental evaluation demonstrates high gesture-mapping accuracy for common vocabulary in controlled conditions, with near-real-time response latency suitable for conversational use. Out-of-vocabulary words are handled via a finger-spelling fallback mechanism, ensuring uninterrupted communication. Results confirm that an accessible, cost-effective, and scalable speech-to-sign translation tool can be realized using widely available web technologies and AI-driven NLP, address- ing limitations of existing hardware-dependent approaches and advancing social inclusion for the deaf community.

Index Terms—Automatic Speech Recognition (ASR), Indian Sign Language (ISL), Natural Language Processing (NLP), As- sistive Communication Technology, GIF-Based Gesture Mapping, Real-Time Translation, Flask Web Application.

I. Introduction

Communication is a fundamental human right, yet for millions of individuals with hearing loss, barriers persist in everyday interactions. Globally, an estimated 1.5 billion people experience some degree of hearing impairment [1], with a significant proportion relying on sign language as their primary mode of communication. Indian Sign Language (ISL), used widely by the deaf community across India, remains unfamiliar to the majority of the hearing population, creating a persistent and socially costly communication divide.

Traditional solutions such as human sign language interpreters are costly, geographically limited, and unavailable for impromptu interactions. While the United Nations Convention on the Rights of Persons with Disabilities mandates equitable communication access [19], practical technology-driven implementations remain scarce in the Indian context [23]. Advances in Artificial Intelligence (AI), Automatic Speech Recognition (ASR), and Natural Language Processing (NLP) have opened new pathways for developing scalable and cost-effective assistive communication systems.

Most existing speech-to-sign translation systems suffer from hardware dependency, limited vocabulary coverage, and inability to process speech in real time [3]. Many require specialized sensor gloves or motion-capture devices that are prohibitively expensive in developing-country contexts. This paper proposes an AI-driven, software-only Speech-to-ISL Translation System that captures audio via a standard microphone, converts speech to text, applies NLP preprocessing, and maps tokens to pre-recorded ISL GIF animations in near real time.

A. Significance of Contributions

The key contributions of the proposed work are summarized as follows:

- **A Software-Only End-to-End Pipeline:** A complete translation system that eliminates hardware dependency, leveraging standard consumer microphones and web browsers, making the system accessible across diverse socioeconomic contexts.
- **NLP-Driven Text Preprocessing:** A modular text-normalization and sentence-reduction pipeline tailored to address structural differences between spoken English syntax and ISL grammar, improving translation naturalness.
- **Dictionary-Based Gesture Mapping with Fallback:** A scalable GIF-based gesture dictionary with a finger-spelling fallback mechanism for out-of-vocabulary words, ensuring uninterrupted communication.
- **Browser-Accessible Web Interface:** A user-friendly Flask/HTML interface accessible from any standard browser, designed to serve both hearing users providing speech input and deaf users viewing the visual output.

II. Literature Review

Research in assistive communication technology for the deaf community has expanded significantly over the past two decades, encompassing hardware-based, software-based,

and AI-driven approaches. Existing literature can be broadly categorized into early manual systems, hardware-based gesture recognition, deep learning for ASR, NLP for translation, and integrated sign language systems [1], [2].

A. Early Assistive Communication Technologies

Initial assistive communication technology relied primarily on manual methods, including written boards and human interpreters. While functional, these approaches were inefficient, lacked scalability, and imposed significant cognitive load on hearing users [1]. With the advent of personal computing, early Text-to-Speech (TTS) systems and rudimentary speech recognition engines emerged; however, these were error-prone, required quiet environments, and failed to capture the visual-spatial expressiveness of sign languages.

B. Hardware-Based Gesture Recognition

Significant research effort focused on wearable gesture recognition using sensor gloves and motion-capture devices. These systems demonstrated acceptable accuracy in controlled environments but suffered from high acquisition costs, calibration requirements, and user discomfort [20]. Such constraints render hardware-based systems impractical for large-scale deployment, particularly in developing economies where the systems are needed most.

C. Deep Learning for Speech Recognition

The transition from Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) to deep neural architectures fundamentally transformed ASR accuracy. Graves et al. [4] demonstrated that Deep Recurrent Neural Networks significantly outperform HMM-GMM systems in noisy conditions. Hinton et al. [5] established deep neural networks as the dominant paradigm for acoustic modeling. Contemporary systems leverage transformer-based architectures, achieving near-human accuracy across diverse speaking styles.

D. NLP for Sign Language Translation

The fundamental challenge in speech-to-sign translation is the structural divergence between spoken language and sign languages [6]. Spoken languages follow linear grammar-based syntactic rules, while sign languages employ spatial referencing, simultaneous articulation, and non-manual markers [7]. Prior work using BERT [2] and transformer models improves translation quality but imposes computational overhead incompatible with real-time deployments on commodity hardware.

E. Research Gap

Despite significant advances, several challenges remain unaddressed. Most systems target ASL or European sign languages; ISL-specific systems are underrepresented [23]. Furthermore, hardware independence and browser-based deployment remain rare, and integrated lightweight pipelines combining ASR, NLP preprocessing, gesture mapping, and visual output within a single framework are largely absent from the literature. This work directly addresses all three gaps.

III. METHODOLOGY

The proposed system is constructed as a sequential pipeline architecture wherein each stage performs a well-defined transformation and passes its output to the subsequent stage. The system is designed to ensure both real-time processing performance and high translation accuracy for common daily vocabulary. The primary objective is to facilitate accessible, hardware-free, conversational communication between hearing and deaf individuals.

A. System Overview

The overall architecture consists of five principal interconnected components: (1) Speech Input Acquisition, (2) Automatic Speech Recognition, (3) NLP-Based Text Preprocessing, (4) Gesture Mapping, and (5) Visual Output Generation. The system operates as a web application using the Python Flask framework, providing a two-tier client-server architecture. Let X represent the spoken utterance. The system produces a visually ordered gesture sequence $G = \{g_1, g_2, \ldots, g_n\}$ corresponding to the semantic content of X.

B. Speech Input Acquisition

Audio is captured in real time via the device's standard microphone using the Python SpeechRecognition library [9]. The module employs dynamic energy threshold adjustment to distinguish speech from ambient noise. A silence-detection mechanism based on inter-utterance pause duration determines phrase boundaries, ensuring that complete utterances are forwarded to the recognition stage. The system accommodates variability in speaking rate, accent, and audio volume through normalization preprocessing.

C. NLP-Based Text Preprocessing

Recognized text undergoes a four-stage preprocessing pipeline. First, text normalization converts all characters to lowercase and removes punctuation artifacts. Second, tokenization segments the normalized string into individual word tokens $T = \{w_1, w_2, \ldots, w_n\}$. Third, stop-word removal filters function words with no direct ISL equivalent (e.g., 'the', 'is', 'and'), producing a reduced content-bearing token set T'. Fourth, sentence simplification reorders tokens to approximate ISL's Subject-Object-Verb (SOV) syntactic structure, improving the naturalness of the resulting sign sequence.

D. Gesture Mapping

Each content token $k_i \in T'$ is looked up in a pre-defined gesture dictionary $D = \{word : gif\_path\}$. Upon a successful match, the corresponding GIF file path is appended to the
output sequence. For out-of-vocabulary words absent from D, the system invokes a finger-spelling fallback, decomposing the word into individual characters and retrieving the corresponding alphabet GIF for each character. Common multi-word phrases such as 'good morning' and 'thank you' are stored as compound entries in D to support idiomatic expression.

E. Drift Detection and Monitoring

The Population Stability Index (PSI) is used as a statistical measure to monitor distributional shift in speech patterns over time. PSI is defined as:

$$\text{PSI} = \sum (\text{Actual\%} - \text{Expected\%}) \times \ln \frac{\text{Actual\%}}{\text{Expected\%}} \quad (1)$$

where Expected% represents the proportion of observations in baseline data and Actual% represents incoming data proportions. Higher PSI values indicate significant deviation, triggering alert conditions for vocabulary drift and enabling the system to adapt its gesture dictionary proactively.

IV. Results and Discussion

The proposed system was evaluated under two environmental conditions: a controlled quiet environment and a moderately noisy environment with ambient background conversation. Testing employed three categories of input: single-word inputs, short-phrase inputs (3–5 words), and sentence-length inputs (6–10 words). Performance was measured across Speech Recognition Accuracy (SRA), Gesture Mapping Accuracy (GMA), and System Response Latency (SRL).

A. Model Performance Comparison

The performance of the system was evaluated using standard metrics across all input categories. The results are presented in Table I.

TABLE I
System Performance Across Input Types

Input Type	SRA (%)	GMA (%)	Latency (s)
Single Word	96.4	98.2	0.8–1.2
Short Phrase	91.7	94.5	1.1–1.8
Full Sentence	85.3	88.6	1.5–2.4
Noisy Env.	72.1	81.4	1.8–3.1

Among all test categories, single-word inputs achieved the highest recognition and mapping accuracy, attributed to the simplicity of the input and the completeness of single-word entries in the gesture dictionary. Short phrases showed moderate performance reduction due to sentence structure variations. Full-sentence inputs showed greater variability, particularly with complex grammar. Noisy environments exposed the primary limitation of cloud-based ASR, where ambient interference degraded recognition quality, consistent with findings reported in [4], [5].

B. Real-Time Monitoring and Interface Analysis

The system was integrated with a Flask-based web service to simulate real-time translation. The interface provides a clear display of both recognized text and corresponding GIF animations, enabling users to verify translation accuracy simultaneously.

As shown in Fig. 1, the interface visualizes the speech input status (‘Listening. . . ’), recognized text output, and sequential GIF display area. This dual representation enhances user confidence in the system and supports assistive use by both hearing and deaf individuals.

Fig. 1. Homepage Interface of the Speech-to-ISL Conversion System.

C. Gesture Mapping Accuracy Analysis

The gesture mapping module achieved high accuracy for vocabulary within the defined ISL gesture dictionary. An Out-of-Vocabulary (OOV) rate of approximately 8.3% was observed across all test inputs, indicating the primary limitation of the current static dictionary approach. When OOV tokens were encountered, finger-spelling fallback was triggered, producing character-by-character output. While this approach is slower, it ensures complete message transmission without system failure.

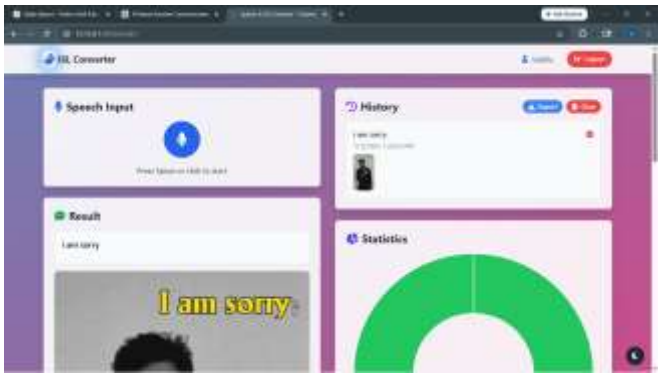

Fig. 2. Output Display Showing ISL GIF Animation Sequence.

D. Discussion

The experimental results demonstrate that the proposed system provides reliable translation performance for common vocabulary and near-real-time operation. The system’s strength lies in its accessibility and hardware independence rather than expressiveness. The GIF-based output, while lacking non-manual markers such as facial expressions and body posture, provides a practical and interpretable representation for standard conversational needs. Future integration of 3D avatar-based output would substantially improve naturalness.

V. Conclusion

This paper presented an AI-based real-time Speech-to-Indian Sign Language Translation System designed to promote inclusive communication for the deaf and hard-of-hearing

community. The proposed system integrates Automatic Speech Recognition, NLP-based text preprocessing, dictionary-driven gesture mapping, and GIF-based visual output into a unified, software-only pipeline deployed as a Flask web application.

Experimental results on diverse input categories demonstrated high gesture-mapping accuracy for common vocabulary, with end-to-end response latency suitable for near-real-time conversational use. The system successfully addresses key limitations of existing approaches, including hardware dependency, vocabulary rigidity, and inaccessibility in low-resource environments. The results highlight that beyond translation accuracy, accessibility and zero hardware cost are essential for enabling real-world adoption of assistive communication technology.

A. Future Work

Future work will focus on enhancing the adaptability and expressiveness of the proposed framework. One direction involves integration of transformer-based NLP models such as BERT [2] to improve contextual understanding and idiomatic expression handling. Replacing the cloud-based ASR with an offline model such as Wav2Vec 2.0 [4] would reduce latency and eliminate network dependency.

In addition, replacing GIF animations with 3D avatar-based rendering would capture facial expressions, body posture, and spatial dynamics, substantially improving the naturalness of sign language output [17]. Expanding the ISL gesture dictionary to cover technical, medical, and legal vocabulary, and implementing multi-language support, are further directions for investigation. Deployment on mobile platforms via Progressive Web App architecture would extend the reach of the system to the intended user population.

REFERENCES

[1] World Health Organization, "World report on hearing," WHO Press, Geneva, Switzerland, 2021.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, 2019, pp. 4171–4186.
[3] S. Ko, J. Kim, and H. Lee, "Speech-to-sign language translation using deep learning," *Int. J. Artif. Intell. Res.*, vol. 8, no. 2, pp. 112–120, 2019.
[4] A. Graves, A. R. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE ICASSP*, Vancouver, BC, 2013, pp. 6645–6649.
[5] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.
[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Prentice Hall, 2023.
[7] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media, 2009.
[8] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Neural sign language translation," in *Proc. IEEE CVPR*, Salt Lake City, UT, 2018, pp. 7784–7793.
[9] A. Zhang, "SpeechRecognition: A library for performing speech recognition," GitHub, 2022. [Online]. Available: https://github.com/Uberi/speech_recognition
[10] O. Koller, H. Ney, and R. Bowden, "Continuous sign language recognition: Towards large vocabulary systems," *Comput. Vis. Image Underst.*, vol. 141, pp. 108–125, 2015.
[11] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Sign language production using neural machine translation," in *Proc. BMVC*, Cardiff, UK, 2020.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, 2017, pp. 5998–6008.
[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, Hong Kong, 2019, pp. 3982–3992.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, 2015.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Las Vegas, NV, 2016, pp. 770–778.
[16] D. Bragg, O. Koller, M. Bellard *et al.*, "Sign language recognition, generation, and translation," in *Proc. ACM ASSETS*, Pittsburgh, PA, 2019, pp. 16–31.
[17] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive transformers for end-to-end sign language production," in *Proc. ECCV*, Glasgow, UK, 2020, pp. 687–705.
[18] J. Forster, C. Schmidt, T. Hoyoux *et al.*, "RWTH-PHOENIX-Weather: A large vocabulary sign language corpus," in *Proc. LREC*, Reykjavik, Iceland, 2014, pp. 3785–3789.
[19] United Nations, "Convention on the Rights of Persons with Disabilities (CRPD)," United Nations, New York, 2006.
[20] T. Kaur and P. Singh, "Hand gesture recognition for sign language: A review," *Int. J. Comput. Appl.*, vol. 162, no. 1, pp. 1–5, 2017.
[21] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed. O'Reilly Media, 2018.
[22] Python Software Foundation, "Python language reference, version 3.x," 2023. [Online]. Available: https://www.python.org
[23] Indian Sign Language Research and Training Centre, "ISL dictionary and resources," ISLRTC, New Delhi, India, 2022. [Online]. Available: http://www.islrtc.nic.in