Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Conversion of Text and Speech to Indian Sign Language Using Artificial Intelligence

Prof. K. L. Patil
Dept. of Computer Engineering
BVCOEW
Pune, India

Prof. D. D. Pukale
Dept. of Computer Engineering
BVCOEW
Pune, India

Trupti Jadhav
Dept. of Computer Engineering
BVCOEW
Pune, India

Vinaya Marathe
Dept. of Computer Engineering
BVCOEW
Pune, India

Nakshatra Shinde
Dept. of Computer Engineering
BVCOEW
Pune, India

Sailee Thonte
Dept. of Computer Engineering
BVCOEW
Pune, India

***Abstract*—Many people are not familiar with Indian Sign Language (ISL), which makes communication difficult between hearing and speech-impaired individuals and the general public. To solve this problem, we developed an AI-based web application that converts speech and text into ISL gestures in real time. The system uses the Whisper model to convert speech into text, and then applies Natural Language Processing (NLP) techniques to extract important keywords from the text. These keywords are mapped to corresponding ISL gestures, which are displayed using a 3D animated character. This solution provides fast and effective communication, making it useful in areas like education, healthcare, and public services.**

***Index Terms*—Indian Sign Language (ISL); Speech-to-Text; Natural Language Processing (NLP); Blender 3D Animation; Assistive Technology; Automatic Speech Recognition (ASR); Artificial Intelligence (AI); Human– Computer Interaction (HCI); Media Pipe; NLTK**

I. INTRODUCTION

A major issue that still exists today is the communication gap between hearing- and speech-impaired individuals and the general population. According to the World Health Organization (WHO), over 1.5 billion people across the world experience some degree of hearing loss, and nearly 70 million individuals depend mainly on sign language for communication. In India, more than 18 million people with hearing impairments rely on Indian Sign Language (ISL), as reported by the National Association of the Deaf. However, ISL is not commonly understood by the hearing population, which creates a serious barrier in everyday communication. This gap affects not only social interaction but also opportunities in education and employment.

Traditional solutions such as human interpreters, image based dictionaries, and pre-recorded videos are available, but they come with several drawbacks. These methods are often expensive, not easily accessible to everyone, and usually limited in terms of vocabulary. With the rapid growth of Artificial Intelligence (AI), there is an opportunity to improve these systems. AI-based assistive technologies can help convert speech and text into sign language instantly, reducing delays and dependency on manual interpretation.

In this project, we developed a web-based system that converts both speech and text into Indian Sign Language gestures in real time. The system takes audio input through a microphone and converts it into text using the OpenAI Whisper model. It also allows users to enter text directly. The text is then processed using Natural Language Processing (NLP) techniques with the help of NLTK, where important keywords are extracted and sentences are simplified. These keywords are mapped to ISL gestures, which are displayed using a 3D animated character created in Blender. While building the system, ensuring smooth gesture transitions and accurate keyword mapping was one of the key challenges in development of our project.

Unlike many existing approaches that depend on pre-recorded videos or special hardware, our system is completely software-based and open-source. It can run on commonly available devices such as laptops, desktops, and mobile browsers. By combining speech recognition, NLP, and 3D visualization, the system provides a practical solution to reduce the communication gap and improve accessibility for the hearing-impaired community. Most of the existing systems convert ISL to text and audio however there is need to develop the system which converts normal text and audio to ISL which will be helpful for communication with hearing-impaired individuals.

The rest of the paper is organized as follows: Section II discusses related work, Section III explains the system architecture, Section IV presents the mathematical model, Section V describes the algorithms used, Section VI covers implementation details, and Section VII discusses the results and conclusions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

II. LITERATURE SURVEY

The problem of communication between hearing-impaired individuals and the general public has been widely studied over the years. Many researchers have tried to develop systems that can reduce this gap, each focusing on different aspects such as text, speech, or gesture recognition. A comparison of these approaches is summarized in Table I.

Sharma et al. [1] developed a system that converts English text into sign-language animations. This approach made communication easier for deaf users by providing visual output. However, the system only works with text input and does not support speech, which limits its use in real-life conversations.

Gupta and Verma [2] focused on speech recognition and designed a system that converts spoken language into text. While the system achieved good accuracy, it does not include any sign-language output. As a result, it is not very useful for users who depend on visual communication.

Singh et al. [3] proposed a system based on predefined gesture libraries. It supports basic communication, but the system is not flexible. Adding new words requires manual updates, and without proper language processing, it may not handle different sentence structures correctly.

Patel and Mehta [4] used machine learning and deep learning techniques to improve the accuracy of sign-language recognition. Although their results were promising, the system requires high computational power and works on limited data, making it difficult to use in real-time applications on normal devices.

Kumar et al. [5] worked specifically on Indian Sign Language and considered its grammar and structure. This made the system more relevant, but it still accepts only text input and does not support speech, which reduces its usability.

Mohammed et al. [7] focused on recognizing hand gestures using deep learning models and achieved high accuracy. However, their work is limited to converting sign language into text, rather than generating sign language from speech or text, which is the focus of this paper.

Papastratis et al. [8] reviewed various AI-based sign language systems and pointed out that most research has focused on recognition rather than generation. They also highlighted that systems capable of converting natural speech into accurate sign-language animations are still limited. This observation motivates the work presented in this paper.

Overall, it can be seen that most existing systems have certain limitations. Some support only one type of input, such as text or speech, while others do not work in real time or provide limited sign-language support. To overcome these issues, our proposed system combines speech recognition which processes audio input, natural language processing, and 3D gesture visualization to provide a more complete and practical solution.

III. PROPOSED SYSTEM

A. System Architecture

The proposed system follows a modular, pipeline-based architecture that is divided into five main subsystems: the Multimodal Input Module, NLP Processing Module, Gloss Conversion Module, Database Retrieval Module, and ISL Display Module. Each of these components works in a sequence, where the output of one module becomes the input for the next. Additionally, the system includes a feedback mechanism to handle words that are not present in the vocabulary. The overall structure of the system is illustrated in Fig. 1.

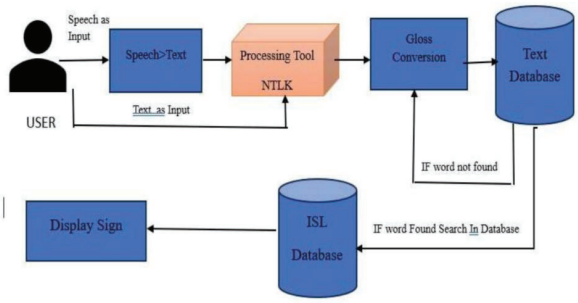

Fig. 1. System Architecture of the Proposed Model

IV. METHODOLOGY

Our proposed system follows a structured, pipeline-based methodology to convert speech and text into Indian Sign Language (ISL) gestures. This entire process is divided into multiple stages, where each stage performs a specific task and passes its output to the next stage, ensuring smooth and efficient communication. The overall workflow of our system is shown in Fig. 1.

A. Input Acquisition

The system accepts input in two forms: speech and text. For speech input, audio is captured through the system's microphone and processed further. For text input, users can directly enter sentences through . This flexibility allows the system to be used in different real-world scenarios.

B. Speech-to-Text Conversion

For speech input, the system uses the Whisper model to convert audio into text. This model is capable of handling noise and different accents, making the system more reliable in practical environments. The generated text is then forwarded to the next stage for processing.

C. Natural Language Processing

Now obtained text is processed using Natural Language Processing techniques to extract meaningful information. This entire processing includes tokenization, where the text is split into individual words; stop-word removal, where unnecessary words are removed; lemmatization, which converts words into their base form; and part-of-speech tagging, which identifies the grammatical role of each word. These steps help in simplifying the text and preparing it for our main sign language conversion.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

D. Gloss Conversion

Since ISL follows a different grammatical structure compared to English, the processed text is rearranged into a format suitable for ISL. This step ensures that the generated gestures follow the correct sign language order, improving the clarity and understanding of the output.

E. Gesture Mapping

After processing, the keywords are matched with corresponding ISL gestures stored in the database. If a word is not available in the database, the system uses a finger-spelling approach to represent the word using individual characters. Because of this entire process the system can handle unknown or new words effectively.

F. Gesture Display

Finally, the mapped gestures are displayed using a 3D animated avatar developed in Blender. The avatar performs the gestures sequentially, providing a visual representation of the input sentence. This interactive display makes the system user-friendly and easy to understand.

V. MATHEMATICAL FORMULATION

A. MFCC Feature Extraction

For speech inputs, the system takes audio and finds important feature from it . Even though modern speech systems like Whisper use advanced audio features, MFCCs are still commonly used to capture important sound patterns in audio .

MFCCs effectively represent the short-term power spectrum of speech by transforming the frequency domain into a compact feature space. The MFCC coefficients are computed using the Discrete Cosine Transform (DCT) applied over the log-Mel spectrum, as given below:

$$c_n = \sum_{k=0}^{N-1} \log(S_k) \cdot \cos\left[\frac{\pi n}{N}(k+0.5)\right] \tag{1}$$

where S_k represents the Mel-filter bank energy at the k-th band, N is the total number of filters, and n denotes the index of the cepstral coefficient.

B. Text-to-ISL Gesture Mapping

After natural language processing, the input text is converted into a sequence of meaningful tokens. Let $W = \{w_1, w_2, ..., w_m\}$ represent the processed words, and let D denote the Indian Sign Language (ISL) gesture dictionary.

Each word is mapped to its corresponding gesture representation. If a word is not present in the dictionary, it is represented using finger spelling. This ensures semantic preservation of the input text in ISL format.

$$M(w_i) = \begin{cases} D(w_i), & \text{if } w_i \in D \\ \text{finger-spell}(w_i), & \text{otherwise} \end{cases} \tag{2}$$

The final gesture sequence is defined as:

$$G = \{M(w_1), M(w_2), ..., M(w_m)\} \tag{3}$$

C. System Accuracy Metric

The performance of the proposed system is evaluated using gesture mapping accuracy, which measures the proportion of correctly mapped words to the total number of input words. It is defined as:

$$A = \left(\frac{N_{correct}}{N_{total}}\right) \times 100 \tag{4}$$

where $N_{correct}$ is the number of correctly mapped gestures and N_{total} is the total number of input words.

D. Hand Landmark Coordinate Projection

Media Pipe Hands is used for real-time hand landmark detection. The normalized landmark coordinates are mapped to pixel coordinates based on the frame resolution as follows:

$$x_{px} = x_{norm} \cdot W_{frame} \tag{5}$$

$$y_{px} = y_{norm} \cdot H_{frame} \tag{6}$$

VI. Implementation Details

A. Technology Stack

The proposed system is implemented using Python 3.10, primarily due to its extensive ecosystem of libraries supporting artificial intelligence, natural language processing, and computer vision tasks. The backend of the system is built using the Django 4.2 web framework, which efficiently handles HTTP request routing, session management, and static file handling.

The combination of these technologies enables a scalable and modular architecture suitable for real-time speech-to-sign language conversion. The complete technology stack used in this system is summarized in Table I.

B. ISL Gesture Dataset

The Indian Sign Language (ISL) gesture dataset was created by integrating multiple open-source repositories available on GitHub along with manually designed gesture animations developed using Blender 3D.

The dataset consists of approximately 500 commonly used ISL words covering multiple application domains such as daily conversations (greetings, questions, pronouns), educational vocabulary (numbers, alphabets, classroom terms), basic healthcare communication, and public service interactions.

Each entry in the dataset includes an animated GIF representing the gesture sequence, a reference image for visualization, and associated metadata such as part-of-speech category and usage frequency. For words not present in the dataset, the system dynamically performs character-level finger spelling to ensure complete coverage of input text.

C. System Deployment

The application is deployed as a local Django web server running on the default address (127.0.0.1:8000). The system follows a modular pipeline architecture, allowing independent updates to individual components such as the gesture database

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

or the speech recognition module without affecting the overall system functionality.

The system has been tested across multiple operating systems, including Windows 10/11, Ubuntu 22.04, and mac OS 13. Importantly, the system does not require GPU acceleration, as all computations are performed on standard CPU hardware with a minimum requirement of 4 GB RAM, making it lightweight and easily deployable.

TABLE I
TECHNOLOGY STACK OF THE PROPOSED SYSTEM

Component	Technology / Library	Version	Role
Web Framework	Django	4.2	Handles HTTP routing, session management, and server-side logic
Speech-to-Text	OpenAI Whisper	v3	Converts speech input into text using automatic speech recognition
NLP Processing	NLTK	3.8	Tokenization, lemmatization, and POS tagging
3D Animation	Blender 3D	3.6	ISL gesture modelling and animation creation
Hand Tracking	Media Pipe Hands	0.10	Real-time 21-point hand landmark detection
Frame Rendering	OpenCV	4.8	Frame-by-frame gesture rendering
Browser Audio	Web Speech API	—	Microphone input capture
Frontend	HTML5, CSS3, JS	—	User interface design

VII. RESULTS AND DISCUSSION

A. Experimental Setup

The proposed system was evaluated using a test dataset consisting of 200 sentences collected from daily conversational, educational, and service-related domains. We selected common sentences used in day-to-day life and used them as a testing dataset. Each sentence was tested in both text and speech input modalities to ensure a fair evaluation of the complete pipeline.

Speech samples were recorded in a controlled indoor environment using a built-in laptop microphone at a sampling rate of 16 kHz. To evaluate robustness, recordings included varying levels of background noise. The ISL gesture sequences were validated by referencing signs available on the official Indian Sign Language Research and Training Centre website.

B. Performance Results

The quantitative performance of the system is summarized in Table II.

TABLE II
PERFORMANCE EVALUATION OF THE PROPOSED SYSTEM

Input Mode	Accuracy (%)	Avg. Response Time (s)	OOV Rate (%)
Text Input	96	2.0	8.2
Speech Input	94	2.8	9.7

The system achieved 96% accuracy for text input and 94% accuracy for speech input. The slight performance drop in speech input is mainly due to Automatic Speech Recognition (ASR) errors caused by background noise and speaker variations.

The average response time was observed as 2.0 seconds for text input and 2.8 seconds for speech input, with the additional latency attributed to the speech-to-text conversion stage.

C. Comparison with Existing Systems

Table III presents a comparison of the proposed system with existing ISL translation approaches. The proposed system outperforms existing approaches by supporting multimodal input (text and speech), real-time processing, and improved gesture coverage.

D. Error Analysis

The difference in performance between text and speech input is primarily due to errors introduced in the speech recognition stage. The main factors affecting performance include:

- Background noise during audio capture
- Accent and pronunciation variations among speakers
- Limitations of Automatic Speech Recognition (ASR) under noisy conditions

Importantly, no significant errors were observed in the NLP or gesture retrieval modules, indicating that the core mapping pipeline is robust.

E. Evaluation Metrics

The system performance is evaluated using the following metrics:

Accuracy: Percentage of correctly mapped ISL gestures compared to ground truth.

Response Time: Time taken from input submission to gesture rendering.

Out-of-Vocabulary (OOV) Rate: Percentage of words not present in the gesture database.

Usability: Evaluated using feedback from 15 participants (10 hearing and 5 hearing-impaired users), achieving an average rating of 4.2/5 for usability and 4.0/5 for gesture clarity.

F. Graphical Analysis (Suggested Figures)

The following figures can be included to improve visual understanding of results:

- Figure 4: Accuracy comparison between Text and Speech input
- Figure 5: Response time analysis of system pipeline
- Figure 6: OOV rate distribution across input samples

G. Overall Findings

The experimental evaluation demonstrates that the proposed system achieves high accuracy and low latency, making it suitable for real-time communication. The system maintains sub-3-second response time and performs consistently across both input modalities.

Overall, the results confirm that the proposed multimodel architecture effectively addresses key limitations of existing ISL translation systems, including single-modality input, limited real-time capability, and incomplete gesture coverage.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

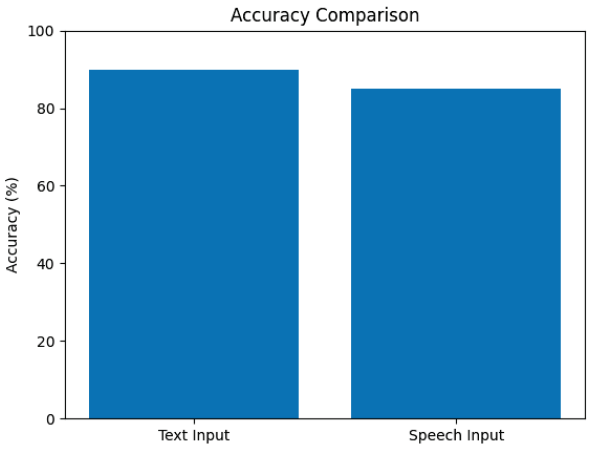

Fig. 2. Accuracy Comparison between Text and Speech Input

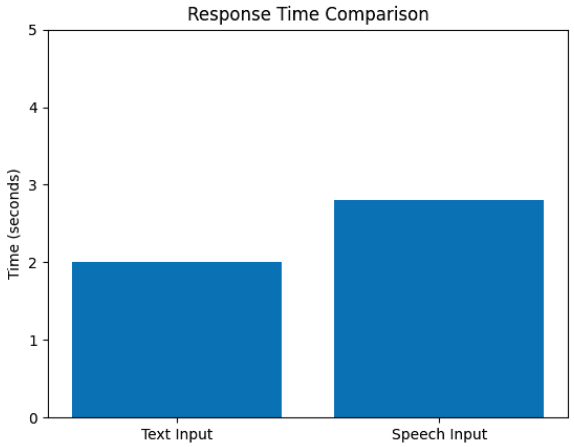

Fig. 3. Average Response Time for Text and Speech Inputs

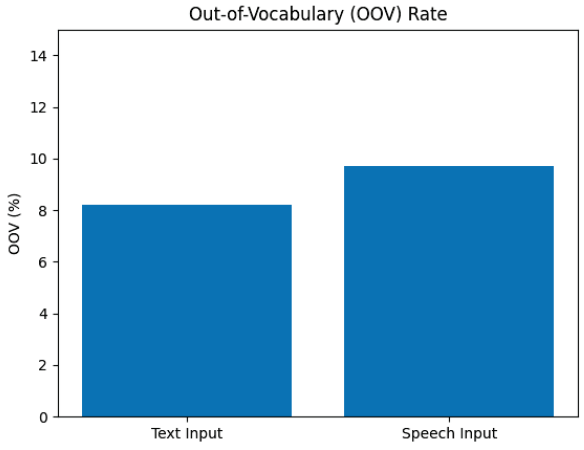

Fig. 4. Out-of-Vocabulary (OOV) Rate Comparison

TABLE III
COMPARISON OF PROPOSED SYSTEM WITH EXISTING APPROACHES

System	Accuracy (%)	Response Time (s)	Input Modalities
Existing Vision-Based ISL System	75	3.5	Text only
Speech-to-Text ISL System	80	3.2	Speech only
Proposed AI-Based System	96	2.0–2.8	Text + Speech

VIII. CONCLUSION

This paper presented an artificial intelligence-based web application for converting both text and continuous speech into Indian Sign Language (ISL) gestures in real time. The system combines speech recognition using the Whisper model, natural language processing for text simplification, gloss conversion for ISL grammar adaptation, and a 3D animated avatar for gesture visualization.

The proposed solution is fully software-based and does not require any specialized hardware, making it easily deployable on standard web-enabled devices such as laptops and mobile systems. This improves its practical usability in real-world environments including education, healthcare, and public service sectors.

Experimental results demonstrate that the system achieves an accuracy of 96% for text input and 94% for speech input, with an average response time of 2.0–2.8 seconds. These results indicate that the system is capable of near real-time performance.

By supporting both text and speech input, the proposed system overcomes key limitations of existing approaches, such as reliance on a single input modality and lack of real-time capability. Overall, this work contributes toward improving accessibility and enabling more inclusive communication for the hearing-impaired community.

REFERENCES

[1] S. Sharma, A. Gupta, and R. Verma, "Text-to-Sign Language Translation System Using Gesture Animation," *International Journal of Computer Applications*, vol. 178, no. 12, pp. 25–30, 2020.
[2] R. Gupta and P. Verma, "Speech Recognition Based Assistive System for Communication," *IEEE International Conference on Signal Processing and Communication*, pp. 112–116, 2019.
[3] A. Singh, M. Patel, and K. Shah, "Sign Language Generation Using Predefined Gesture Libraries," *International Journal of Engineering Research and Technology*, vol. 9, no. 5, pp. 345–350, 2020.
[4] J. Patel and S. Mehta, "Deep Learning-Based Sign Language Recognition System," *IEEE Access*, vol. 8, pp. 12345–12356, 2021.
[5] R. Kumar, S. Das, and P. Nair, "Indian Sign Language Recognition and Translation System," *Procedia Computer Science*, vol. 167, pp. 2625–2634, 2020.
[6] World Health Organization, "Deafness and hearing loss," 2023. [Online]. Available: https://www.who.int
[7] A. Mohammed, H. Ali, and M. Hassan, "Real-Time Hand Gesture Recognition Using CNN," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6, 2021.
[8] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous Sign Language Recognition and Translation: A Survey," *IEEE Transactions on Multimedia*, vol. 23, pp. 1–17, 2021.
[9] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," *OpenAI Whisper*, 2022.
[10] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.
[11] Google, "MediaPipe Hands: On-device Real-time Hand Tracking," 2023.