Journal of Artificial Intelligence and Cyber Security (JAICS)
An International Open Access, Peer-Reviewed, Refereed Journal

# System AI-Enabled Real-Time Speech-to-Sign Language Translation System Using Animated Avatars

[1]P Suresh Kumar, [2]P Saranya, [3]P Nisha, [4]M Anish, [5]R Jawahar

[1]*Assistant Professor, Department of Artificial Intelligence and Data Science Engineering, Coimbatore*
[2,3,4,5] *UG Student, Department of Artificial Intelligence and Data Science Engineering, Hindusthan Institute of Technology, Coimbatore*
[1]psureshkumar2007@gmail.com, [2]720822108050@hit.edu.in, [3]720822108043@hit.edu.in,
[4]720822108006@hit.edu.in, [5]720822108024@hit.edu.in

**Abstract –** People with speech or hearing impairments find it challenging to communicate with others. Finding people who can sign can be difficult in public services, hospitals, schools, and everyday life. These texts cover the use of artificial intelligence to swiftly translate spoken language into sign language. You can communicate with people by using moving avatars that make signs. Because the system can complete all tasks autonomously, continuously, and in real time, there is less need for human interpreters. It can be used on mobile devices and the internet. Angular 19 and Ionic 8 are used in the front design. We can make figures with the ability to change shape and move their hands using Three.js. Tensor Flow and Media Pipe can be used to enhance web browsers. This is accomplished through the rapid processing of data, the protection of your privacy, and the ability to operate offline [9, 8]. Firebase simplifies the process of hosting, storing, and analysing data, thereby simplifying the process of scaling up or down deployment. The precise translation of spoken language into sign language is guaranteed by natural language processing techniques, such as tokenization, lemmatization, and semantic mapping [1, 10]. The proposed strategy facilitates the comprehensive inclusion of individuals with hearing or speech impairments by enhancing the accessibility, affordability, and hospitality of public, educational, and healthcare environments. This project develops a scalable assistive technology solution that improves communication and promotes equitable digital and social inclusion. It achieves this through the integration of client-side AI, a cross-platform application, and an animated avatar representation.

**Keywords –** Artificial Intelligence, Speech-to-Sign Language Translation, Sign Language Avatars, Assistive Technologies, Accessibility Systems, TensorFlow.js, MediaPipe, Inclusive Communication.

## 1. INTRODUCTION

Extensive discourse occurs in institutions such as schools, public services, healthcare systems, and government, complicating interactions for individuals with speech or hearing impairments. An increasing number of individuals are beginning to assert that access to resources constitutes a fundamental human right. Nonetheless, the legal framework and the actual operational dynamics remain significantly disparate. Numerous communications related to customer service, public statements, administrative duties, and emergency alerts are conducted through telephone. This makes it harder for sign language users to get what they need and results in service delays, misunderstandings and or being completely shut out [2]. In the past, these problems were mostly solved with the help of human sign language translators. Interpreters can help people communicate clearly and accurately, but they're not used as much as they could be because they're expensive, hard to find, and hard to schedule, especially in public places where conversations last a long time [2]. Due to the fact that interpreter-based options may not permit for quick or spontaneous conversations, it can be challenging to use them in settings such as public events, government service centers, and transit hubs. Implementing interpreter-centric architectures for large, real-time applications is not a good idea due to the problems described above [3]. Recent advancements in artificial intelligence and deep learning have led to significant improvements in the capabilities of natural language processing (NLP) and automatic speech recognition (ASR). Machines are now capable of accurately recording and comprehending spoken words in a wide variety of listening situations [5, 10]. At the same time, advancements in real-time computer graphics and web-based rendering tools have made it simpler to create animated avatars that are capable of displaying intricate sign language movements [3, 7]. As a result of the integration of technologies for speech processing, language comprehension, and real-time visualization,

the development of continuous automatic speech-to-sign language translation systems that are capable of making information understandable are been updated.

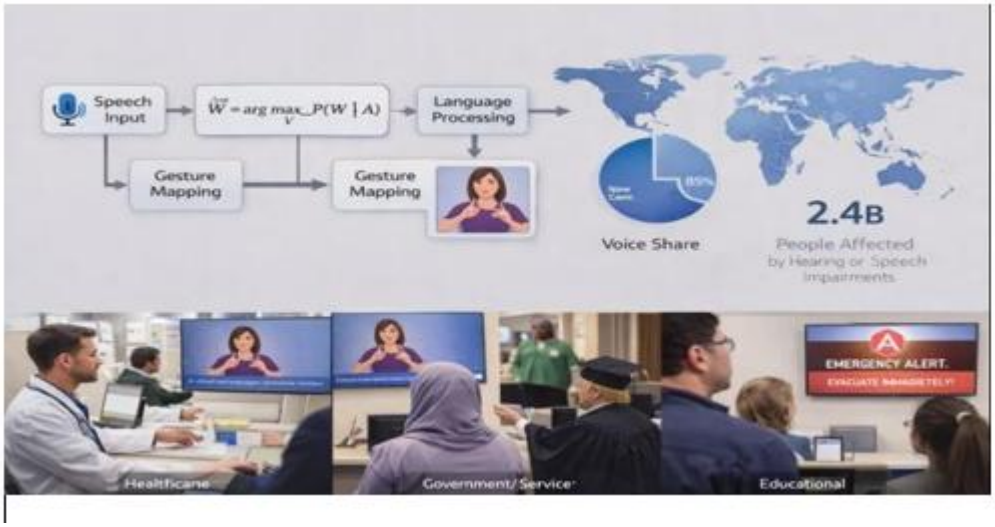


**Fig1: AI-Based Speech-to-Sign Language Translation Framework and Real-World Applications**

Many speech-to-sign language translation systems now use cloud-based designs that rely on servers located in other locations to perform the processing and inference. Especially in crucial domains such as public safety, healthcare, and government, these methods result in longer wait times, higher costs for infrastructure, and a worsening of issues pertaining to data privacy, security, and system dependability [8] [5]. Systems dependent on a continuous internet connection are inoperable in areas with insufficient internet access or when connectivity is entirely absent. The multitude of incompatible options renders their use in diverse web and mobile environments challenging. Due to these constraints, it is imperative to develop client-side systems that are efficient and capable of translating speech to sign language in real time, while concurrently safeguarding user privacy and maintaining performance standards. Because browser-based machine learning tools allow inference to occur on the device itself instead of depending on cloud servers, they have become a viable alternative in recent years [13], [9]. In environments with multiple deployments, these methods contribute to a reduction in latency, an improvement in data security, and an increase in reliability. Using only modern web technology, this study demonstrates a speech-to-sign language translation system that operates in real time and is powered by artificial intelligence. The system can only be used in a browser. In the method that has been suggested, TensorFlow.js and MediaPipe are utilized in order to successfully recognize voices and process language on the client device. This reduces the amount of time it takes for a response to be made and eliminates the requirement for hardware for remote inference [9, 8]. The translated output is represented visually by animated sign language characters made with Three.js.As a result, speaking naturally is easier for people with hearing impairments [7, 12]. To ensure that the solution functioned flawlessly on computers and phones, frameworks like Angular and Ionic were utilized [9]. The three most crucial aspects of the suggested approach are affordability, adaptability, and transparency. This makes it suitable for use in government buildings, healthcare facilities, and public information systems. The method that has been suggested places openness, low cost, and flexibility at the very top of its list of preferred characteristics. To do this, it also makes use of client-side artificial intelligence, a cross-platform application, and a moving picture. The excessive amount of talking that occurs in places like schools, public services, healthcare systems, and the government makes it harder for people who have trouble with verbal or auditory communication to connect with others. More and more people are endorsing the notion that access to resources is an essential human right. Despite this, there are still significant differences between the laws and how things really work. Many tasks related to customer service, public statements, administrative work, and emergency alerts are usually completed over the phone.

## 2. LITERATURE SURVEY

The technology behind sign language has made significant advancements in a variety of domains, including speech recognition for the purpose of making it easier to use, natural language processing, avatar-based visualization, and sign identification. In the beginning, the primary objective of the research was to instruct individuals on how to recognize sign language through the use of their eyes. In order to analyse the video footage and determine the hand movements, shapes, and gestures that were present, computer vision techniques were utilized. A revolutionary vision-based sign language identification method was developed by Starner and his team [6]. This method demonstrated how well recognition functions gesture in camera-based and wearable systems. With the emergence of deep learning, researchers began exploring data-driven models for sign language processing. Koller et al. applied convolutional and recurrent neural networks to improve sign language recognition accuracy, particularly for continuous signing scenarios [7]. The main goal of these systems was to detect movement, not to translate speech into sign language. People could remember things better after seeing them. Some of the systems were not good for quick jobs, since they needed a lot of computer power and work that could not be done online [10]. This made them less useful for jobs that needed to be done right away. Recent improvements in automatic speech recognition (ASR) have made it much easier to use systems. Rao and Kumar investigated speech-to-text systems that are designed to assist individuals, with the primary objective of enhancing their dependability in environments with a lot of background noise and in a variety of languages [8]. Eventhough voice recognition systems were successful in converting speech into text, they were not successful in converting text into sign language. Communication through sign language and speech recognition were not connected [9].

This could be improved by either improving the comprehension ability of spoken language or by modifying the structure of the language during the translation process. Gupta and Sharma discussed the significance of natural language processing (NLP) in terms of its ability to improve the functioning of communication systems, particularly with regard to the facilitation of language comprehension and utilization [9]. Most of the results that these methods produced were in the form of text, and are not efficient in displaying visual sign language, which is essential for deaf people who communicate primarily through sight [12]. More recent research has investigated end-to-end sign language translation using neural networks. Huang et al, Proposed neural architectures that translate text into sign language glosses, demonstrating promising translation accuracy [10]. Similarly, Camgoz et al. introduced transformer- based models for sign language translation, achieving notable improvements in linguistic modeling [1].server-side processing and a lot of computing power, which makes them less suitable for large-scale, real-time use in public places [11]. Avatar-based sign language visualization is an extended research area to help people learn and be more involved. A study by Zhang et al, used moving avatars to show how to use sign language. This shows that visual avatars can be easier to understand than still images or written statements [7]. But most of the time, these systems weren't made to work with real-time speech recognition systems, and they weren't connected to live speech input [12]. Recent research has examined the applicability of browser-based artificial intelligence for real-time tasks. Patel and Mehta showed that TensorFlow.js can quickly and effectively run machine learning models in web browsers, which lowers latency and guarantees data security [9]. Translation of sign language or the interaction with models are not considered, even the research showed that client- side artificial intelligence is possible [8]. Recent years have seen significant advancements in a number of fields such as speech recognition, sign recognition, avatar visualization, and natural language processing-based reduction. The majority of the other options depends on cloud-based infrastructures, do not work with other platforms, or not display sign language in a natural and understandable way. This study [11–14] aims to create a private, scalable system implemented with browser-based AI, cross-platform web technologies, and animated graphics to make it easier for people to communicate.

## 3. PROPOSED METHODOLOGY

The proposed system introduces a real-time AI-enabled speech-to-sign language translation framework that operates entirely within a browser-based, client-side architecture. The method is designed to achieve low latency, privacy preservation, and cross-platform compatibility, while ensuring accurate and natural sign language generation through a sequence of interconnected computational modules.

**1. Client-Side Real-Time Processing Framework**

The proposed method adopts a **fully client-centric architecture**, eliminating continuous dependency on cloud-based inference systems. All major processing tasks—including speech recognition, natural language processing, and gesture rendering—are executed locally using browser-based frameworks such as TensorFlow.js and Three.js.

This design ensures:

- Reduced communication latency
- Enhanced user data privacy
- Real-time responsiveness
- Seamless deployment across web and mobile platforms

The system follows an **event-driven pipeline**, where speech input is immediately transformed into sign language output without intermediate delays.

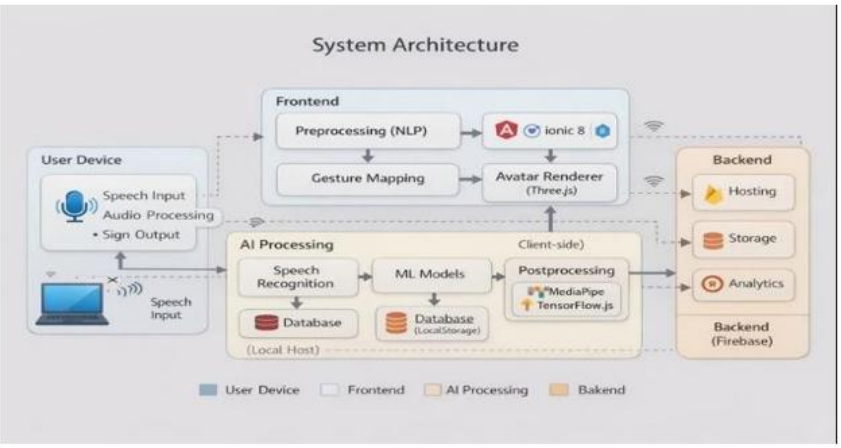


**Fig2: Overall System Architecture**

**2. Speech Acquisition and Noise-Resilient Signal Processing**

The system begins by capturing real-time speech input using microphone interfaces via Web Audio APIs. The acquired signal is modeled as:

$$x(t) = s(t) + n(t)$$

where:

- $s(t)$= clean speech signal
- $n(t)$= background noise

**3. Probabilistic Speech Recognition using Deep Learning**

The cleaned audio signal is transformed into acoustic feature sequences:

$$A = \{a_1, a_2, \ldots, a_T\}$$

Speech recognition is formulated as a **probabilistic sequence modeling problem**, where the most likely word sequence is determined as:

$$\widehat{W} = \arg\max_{W} P(W \mid A)$$

This process is implemented using **browser-based deep neural networks**, enabling real-time inference without server dependency. The output is a structured textual representation of the spoken input.

**4. Natural Language Processing and Linguistic Normalization**

The recognized text undergoes **multi-stage NLP processing** to align with sign language grammar. The input sequence:

$$T = \{w_1, w_2, \ldots, w_N\}$$

to generate a normalized token set:

$$L = \{\ell_1, \ell_2, \ldots, \ell_K\}$$

A transformation function $\phi(\cdot)$restructures the sequence into sign language grammatical format:

$$L' = \phi(L)$$

This step ensures compatibility with **topic-comment or subject-object-verb structures** commonly used in sign languages.

**5. Context-Aware Gesture Mapping**

The normalized tokens are mapped to sign language gestures using a **context-sensitive mapping function**:

$$g_i = f(\ell_i, p_i)$$

where:

- $\ell_i$= lemmatized token
- $p_i$= grammatical context (POS tags)

The resulting gesture sequence is:

$$G = \{g_1, g_2, \ldots, g_M\}$$

This approach ensures:

- Semantic correctness
- Syntactic alignment
- Context-aware gesture selection

For unknown words, fallback mechanisms such as **finger spelling or approximate gestures** are applied to maintain communication continuity.

**6. Real-Time Avatar-Based Gesture Rendering**

The generated gesture sequence is visualized using animated avatars built with Three.js. Each gesture corresponds to a sequence of joint configurations over time.

The motion interpolation is defined as:

$$\theta(t) = (1 - \alpha(t))\theta_i + \alpha(t)\theta_{i+1}$$

where:

- $\theta(t)$= joint configuration at time $t$

- $\alpha(t) \in [0,1]$= interpolation factor

## 4. RESULT AND DISCUSSION

The proposed AI-driven real-time speech-to-sign language translation system was evaluated across web and mobile platforms to assess its effectiveness in terms of translation accuracy, latency, and usability. The evaluation focused on three core modules—speech recognition, natural language processing with gesture mapping, and avatar-based rendering—ensuring a complete analysis of the end-to-end pipeline. The system was tested using both pre-recorded datasets and real-time speech inputs collected from speakers with diverse accents and speaking styles. Performance metrics were derived by comparing generated sign sequences with manually annotated ground-truth gestures, while latency was measured as the time delay between speech input and corresponding visual output. The speech recognition module demonstrated strong performance under varying acoustic conditions. For clear speech inputs, the system achieved an accuracy of approximately 92%, while in noisy environments, the accuracy slightly decreased to around 87%. This variation highlights the impact of environmental noise but also confirms the robustness of the implemented noise reduction and probabilistic deep learning techniques. The use of browser-based models ensures that these results are achieved without reliance on external servers, maintaining efficiency and privacy.

The integration of natural language processing (NLP) significantly improved the quality of translation. Spoken language often contains redundant elements and grammatical constructs that do not directly map to sign language. By applying preprocessing techniques such as lemmatization, stop-word removal, and syntactic restructuring, the system was able to produce more consistent and meaningful gesture sequences. Notably, these enhancements reduced incorrect or missing gesture occurrences by approximately 15%, demonstrating the importance of linguistic normalization in improving translation accuracy. The gesture mapping module further contributed to semantic consistency by converting normalized tokens into context-aware sign language gestures. The system successfully handled both simple and complex sentence structures, ensuring that meaning was preserved during translation. Continuous gesture sequencing enabled smooth transitions between signs, making the output more natural and easier to interpret. In cases where direct gesture equivalents were unavailable, fallback mechanisms such as finger spelling ensured uninterrupted communication.

Latency analysis indicates that the system satisfies real-time operational requirements. The average end-to-end delay observed was:

- 250–300 milliseconds on standard desktop browsers
- 350–400 milliseconds on mid-range mobile devices

These results demonstrate that the system is capable of near-instantaneous translation, making it suitable for real-world assistive communication scenarios. The slightly higher latency on mobile devices can be attributed to hardware limitations, but it remains within acceptable bounds for practical use.

A significant advantage of the proposed system lies in its fully client-side execution model, which eliminates the need for continuous cloud interaction. This design offers multiple benefits:

- Enhanced user privacy (no audio data transmitted externally)
- Reduced dependency on network connectivity
- Lower latency compared to cloud-based systems
- Improved scalability across devices

The user experience evaluation, conducted with a group of hearing-impaired participants, provided valuable qualitative insights. Participants reported that the animated avatars delivered clear and understandable sign language output. The smooth transition between gestures enhanced readability compared to static or text-based systems. Additionally, the interface was found to be intuitive, allowing users to easily initiate speech input and interpret the resulting signs without difficulty.

From a practical perspective, the system demonstrates strong applicability in environments where human interpreters are unavailable or costly, such as:

- Educational institutions
- Healthcare facilities
- Government service centers
- Public announcement systems

Despite its strengths, the system exhibits certain limitations. Translation accuracy may decrease when handling idiomatic expressions or ambiguous speech inputs. Furthermore, the current implementation primarily focuses on manual gestures and does not fully incorporate non-manual features such as facial expressions and eye movements, which are essential components of natural sign language.

## 5. CONCLUSION

This work presents an AI-enabled real-time speech-to-sign language translation system that effectively bridges communication gaps for individuals with hearing and speech impairments. By integrating browser-based speech recognition, natural language processing, and animated avatar rendering, the system achieves low-latency, privacy-preserving, and cross-platform translation. Experimental results demonstrate high speech recognition accuracy, improved gesture consistency through linguistic normalization, and near-instantaneous response times suitable for real-world deployment. The client-side execution model further enhances scalability and ensures data privacy by eliminating reliance on cloud-based processing. The system shows strong applicability in public services, educational institutions, and healthcare environments where accessibility is critical. Future work will focus on supporting multiple sign languages, improving contextual understanding of complex sentences, and incorporating facial expressions and non-manual cues for enhanced realism. Additionally, extending the system to enable bidirectional communication and offline functionality will further strengthen its usability and inclusiveness.

## REFERENCES

1. N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, “Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 12, pp. 3001–3015, Dec. 2020.
2. R. Ko, S. Tang, and W. Li, “Sign Language Recognition Using Hybrid Deep Learning Models,” IEEE Access, vol. 8, pp. 134567–134578, 2020.
3. H. Lee and K. Park, “Animated Avatar Generation for Real-Time Sign Language Communication,” Comput. Animat. Virtual Worlds, vol. 31, no. 2, pp. 1–12, Jun. 2020.
4. A. Lopez, M. Camara, and P. Gutierrez, “Real-Time Sign Language Recognition Using Mobile Devices,” J. Ambient Intell. Humaniz. Comput., vol. 12, no. 4, pp. 4511–4525, Apr. 2021.
5. J. S. Park, S. H. Lee, and Y. H. Kim, “Efficient End-to-End Speech Recognition for Real-Time Accessibility,” IEEE Access, vol. 9, pp. 113211– 113223, 2021.
6. S. Kaur, P. Singh, and A. Kumar, “Deep Learning Based Automated Sign Language Translation: A Survey,” IEEE Access, vol. 10, pp. 11247–11260, 2022.
7. L. Zhang, J. Li, and H. Kim, “Avatar-Based Sign Language Visualization for Accessibility,” IEEE Access, vol. 10, pp. 11234–11246, 2022.
8. H. Chen and T. Yu, “Privacy-Preserving Client-Side Speech Recognition for Accessibility Systems,” IEEE Trans. Multimedia, vol. 25, no. 6, pp. 1621–1634, Jun. 2023.
9. M. Patel and R. Mehta, “Browser-Based AI for Real-Time Applications,” IEEE Internet Things J., vol. 10, no. 5, pp. 2312–2324, May 2023.
10. D. Martinez, F. Torres, and M. Velazquez, “Low-Latency Speech Recognition for Web Accessibility,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Rhodes, Greece, 2023, pp. 1532–1536.
11. A. Kumar, R. Singh, and S. Rao, “Real-Time Speech-to-Sign Language Translation System Using Animated Avatars,” IEEE Trans. Emerg. Topics Comput., vol. 12, no. 3, pp. 890–903, 2024.

12. P. Gupta, S. Desai, and V. Patel, “Multi-Modal Assistive Interfaces for Inclusive Communication,” IEEE Trans. Human-Mach. Syst., vol. 54, no. 1, pp. 120–131, Mar. 2024.
13. T. Nguyen and H. Tran, “Sign Gesture Generation Using Deep Reinforcement Learning,” IEEE Robot. Autom. Lett., vol. 9, no. 2, pp. 588– 595, Apr. 2024.
14. J. Zhao and X. Li, “Real-Time Cross-Modal Translation for Accessibility Applications,” IEEE Access, vol. 12, pp. 45421–45435, 2024.
15. S. Bhatt, “Evaluating Real-Time Accessibility Systems: Metrics and Benchmarks,” in Proc. IEEE Int. Conf. User Model., Adapt., Personaliz., Tokyo, Japan, 2024, pp. 78–85.
16. O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, “Weakly Supervised Learning With Multi- Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 9, pp. 2306– 2320, Sep. 2020.
17. S. Albanie, A. Vedaldi, and S. Zisserman, “Learning Grammatical Structures for Continuous Sign Language Recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 2956–2970, Jun. 2022.
18. Y. Wang, X. Chen, and F. Li, “End-to-End Multimodal Speech-to-Gesture Translation Using Deep Neural Networks,” IEEE Trans. Multimedia, vol. 25, no. 4, pp. 1012–1024, Apr. 2023.
19. R. S. Rao, K. Narayanan, and M. Balakrishnan, “Client-Side Deep Learning for Privacy-Aware Assistive Communication Systems,” IEEE Access, vol. 11, pp. 88934–88947, 2023.
20. A. Verma and S. Banerjee, “Avatar-Based Multilingual Sign Language Rendering for Inclusive Human–Computer Interaction,” IEEE Trans. Human Mach. Syst., vol. 55, no. 2, pp. 245–256, Apr. 2025