World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Gesture talks real-time sign language recognition and animation system using AI

K Kiran Babu, Srikanth Banoth, Vijaya Lakshmi Muvvala *, Mohammad Shafee and Shravan Kumar Ainala

Department of CSE (Data Science), ACE Engineering College, Telangana, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 1362-1369

Publication history: Received on 31 March 2025; revised on 08 May 2025; accepted on 10 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0593

Abstract

Effective communication between deaf and hearing individuals remains a significant challenge due to the fundamental differences in language modalities. GestureTalk presents a real-time, AI-driven communication system designed to bridge this gap by enabling seamless bidirectional interaction. The proposed solution leverages state-of-the-art technologies including Automatic Speech Recognition (ASR), Natural Language Processing (NLP), gesture detection using YOLO, and pose estimation with DWpose and MediaPipe. Spoken language is transcribed and translated into American Sign Language (ASL) gloss, then rendered as realistic 3D animated sign language via a virtual avatar. In the reverse direction, the system captures and interprets sign language gestures in real time, converting them into textual output for hearing users. Designed with real-time performance, high accuracy, and user accessibility in mind, GestureTalk serves as an inclusive interface for communication particularly suited for digital contexts such as video conferencing. This system offers a scalable and adaptable solution, contributing meaningfully to assistive technology and digital accessibility for the deaf and hard-of-hearing communities.

Keywords: Real-Time AI-Driven System; NLP; YOLO; Dwpose; 3D Animated Sign Avatar; Offering an Inclusive; Scalable Solution

1 Introduction

Communication between deaf and hearing individuals is often difficult due to differences in language. Sign language is visual, while spoken language is auditory, creating a communication gap. This project, GestureTalk, aims to bridge that gap using artificial intelligence. It enables real-time, two-way communication between deaf and hearing users. By combining speech recognition and gesture detection, it promotes equal access and inclusivity. The system enhances interaction in both physical and digital spaces.

Traditional systems that attempt to bridge this communication divide face significant drawbacks. Human interpreters, while accurate, are not always available or affordable. Text-based tools neglect the visual-spatial nature of sign language, and existing avatar applications often lack realism and accuracy. One-way solutions—either speech-to-sign or sign-to-text—fail to deliver a complete conversational experience. This project's strength lies in its bidirectional communication system, making it suitable for real-time conversations. It employs a combination of gesture detection, speech recognition, and natural language processing to ensure both parties—deaf and hearing—can understand each other with minimal latency. This addresses the limitations of earlier technologies and opens new avenues for interaction.

The system works in two main directions: speech-to-sign and sign-to-text. Spoken words are converted to text using speech recognition, then translated into ASL gloss.

* Corresponding author: Vijaya Lakshmi Muvvala.

A 3D avatar uses this gloss to perform sign language using animated gestures. For sign-to-text, the system detects hand movements using YOLO and MediaPipe. These gestures are translated into readable English text instantly. The whole process runs in real time with minimal delay.

Technically, GestureTalk utilizes powerful open-source tools and deep learning frameworks like PyTorch or TensorFlow to ensure high performance and scalability. The system is designed to run on standard hardware with GPU acceleration for smooth, real-time processing. Its architecture is modular, allowing for future expansions such as multilingual support or enhanced emotion detection in gestures. With its user-friendly interface and responsive design, the solution has practical applications in education, customer service, and online communication. Ultimately, this project is more than just a technical innovation—it is a step toward social equity and digital inclusion, empowering the deaf community to engage confidently in everyday interactions.

2 Literature review

Earlier sign language systems relied on sensor-based gloves or depth cameras to detect gestures. While accurate, these were expensive and not user-friendly. With advancements in computer vision, webcam-based approaches became more practical and affordable. This shift improved accessibility for real-world applications.

Previous methods mostly supported one-way communication, like sign-to-text or speech-to-text. Many lacked real-time performance and covered only a limited sign vocabulary. Avatar apps were too basic and didn't offer realistic gestures, making them less effective in real conversations.

Recent advances in deep learning and AI tools have significantly improved performance. The YOLO (You Only Look Once) object detection algorithm is now widely used for real-time gesture recognition. Paired with pose estimation tools like MediaPipe and DWpose, these systems can track body and hand movements accurately. Datasets like WLASL and MS-ASL have helped train models to understand a wide variety of signs. These improvements have enabled more natural and responsive systems.

Despite these advances, most systems still lack full bidirectional communication. This project addresses that gap by combining speech-to-sign and sign-to-text translation in one integrated interface. It uses speech recognition, NLP for ASL gloss translation, and 3D avatar animation for expressive signing. By supporting both directions of communication, it offers a complete solution for real-time interaction. This makes GestureTalk a novel and inclusive approach to bridging the deaf-hearing communication divide.

3 Existing system

Current systems for deaf and hearing communication mainly rely on human interpreters. While interpreters provide accurate translation, they are not always available or affordable for everyday use. This limits spontaneous and private communication in many scenarios like public services or online meetings.Some tools use text-based messaging to help bridge the gap, but they do not support sign language. These systems ignore the visual-spatial nature of sign language, making them unsuitable for deaf users who prefer or rely solely on signing. As a result, the interaction feels unnatural and incomplete.

Avatar-based apps have been developed to translate text into basic sign animations. However, most offer only a limited vocabulary and lack lifelike gestures or expressions. Their rigid and robotic movements reduce clarity and emotional expression, which are vital parts of sign language communication. Other gesture recognition systems exist but often suffer from low accuracy and limited vocabulary. Many are one-way solutions—either converting speech to signs or signs to text, but not both. These limitations prevent natural, two-way conversations, highlighting the need for a more integrated and real-time system like GestureTalk.

4 Proposed system

The proposed system, Gesture Link, enables real-time, bidirectional communication between deaf and hearing individuals using advanced AI technologies. It consists of two main modules: speech-to-sign and sign-to-text. In the speech-to-sign module, spoken input is captured using a microphone and converted into text using Automatic Speech Recognition (ASR) tools such as whisper. This text is then processed using Natural Language Processing (NLP) techniques (like SpaCy or NLTK) to generate ASL gloss, a structured form of American Sign Language.

The gloss is passed to a 3D avatar, which uses animation frameworks like DWpose and MediaPipe to perform accurate and lifelike sign gestures.

In the reverse direction, the sign-to-text module uses a standard webcam to capture the user's hand gestures. A YOLO-based model, trained on large-scale datasets such as WLASL and MS-ASL, detects and identifies these gestures accurately. Pose estimation tools further enhance the precision by tracking both hand and body movement, ensuring each sign is interpreted correctly. Once recognized, the gestures are translated into readable English text and displayed to the hearing user. The entire process is optimized for real-time performance with minimal delay, making it practical for live video calls, education, or customer service. Designed to run on standard PC setups, the system is user-friendly, portable, and scalable providing an inclusive solution that supports natural, efficient communication across hearing and non-hearing communities.

5 Methodology

The system enables real-time communication between deaf and hearing individuals using AI. Spoken input from a hearing user is converted to text using ASR (e.g., Whisper), then translated into ASL gloss via NLP.

A 3D avatar animated with DWpose displays the corresponding sign language. For deaf-to-hearing communication, sign gestures captured by a webcam are detected using YOLO with DWpose or MediaPipe. These are recognized using trained models based on WLASL/MS-ASL datasets and converted into readable text. The system ensures accurate, real-time translation in both directions, promoting inclusive and accessible digital communication.

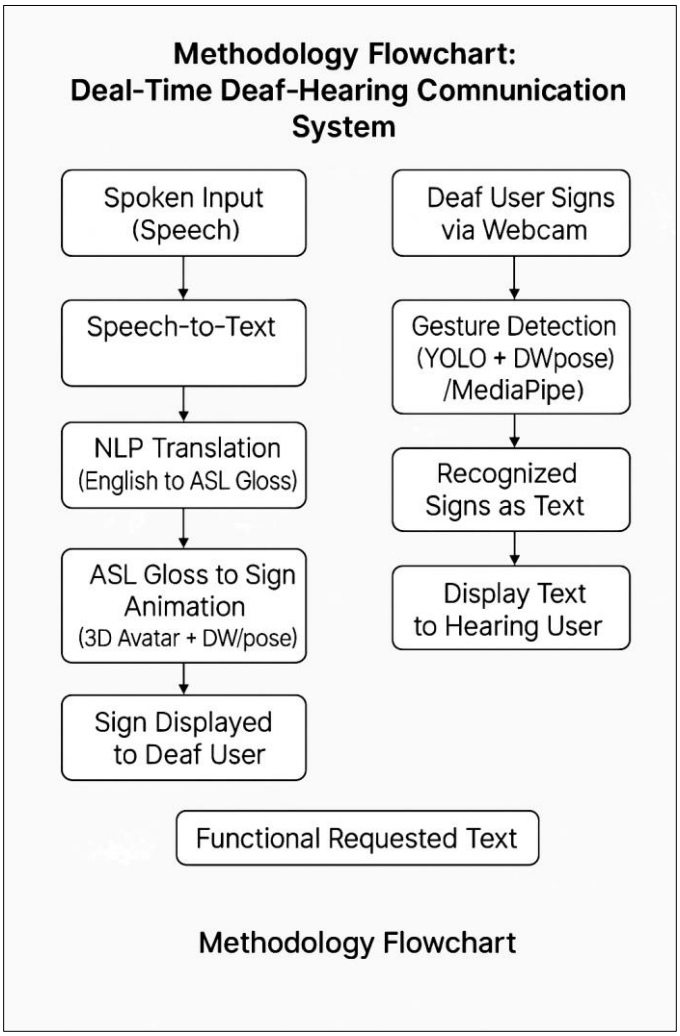

Figure 1 Methodology

5.1 System Architecture

The system uses a webcam and microphone for real-time input. Speech is converted to text using ASR, processed via NLP, and animated for deaf users. Gestures are detected with YOLO and DWpose, then translated into text for hearing users. FastAPI handles backend communication, while AI modules and optional databases manage processing, logs, and user interaction.

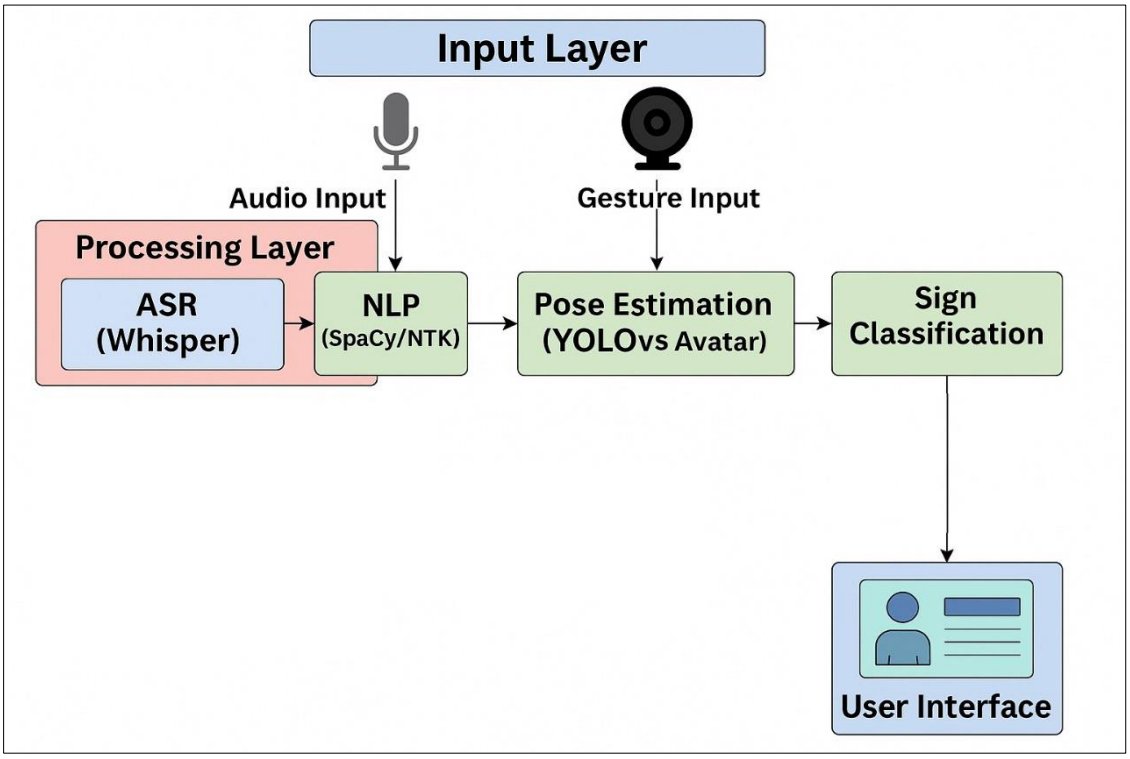

Figure 2 System Architecture

5.1.1 Overview

The proposed system architecture is designed to facilitate real-time, bidirectional communication between deaf and hearing individuals. It integrates audio input, gesture recognition, natural language processing, and avatar-based animation to ensure smooth translation from speech to sign language and vice versa. The architecture is divided into three main layers: Input Layer, Processing Layer, and Application Layer, each handling distinct functionalities.

5.1.2 Input Layer

- This layer captures data from the environment to initiate the processing flow. It includes:
- Microphone: Captures spoken language from hearing users, which is processed into text.
- Webcam: Records sign gestures made by deaf users in real time, which are interpreted through computer vision models.

5.1.3 Processing Layer

This is the core layer where data from inputs is processed through parallel pipelines:

- Speech-to-Sign Pipeline
- Automatic Speech Recognition (ASR)
- Tools: Whisper, SpeechRecognition
- Function: Converts spoken audio into text in real time.

5.1.4 Natural Language Processing (NLP)

- Tools: SpaCy, NLTK
    - Function: Analyzes and formats the text for gesture representation. (Note: ASL gloss has been omitted in this version.)
- Sign Animation Generation
    - Tools: DWpose, MediaPipe
    - Function: Animates a 3D avatar to perform corresponding sign gestures.
- Sign-to-Text Pipeline
    - Gesture Detection and Pose Estimation
    - Tools: YOLOv5, DWpose, MediaPipe
    - Function: Detects and tracks hand and body movements from live webcam input.
- Sign Classification
    - Function: Recognizes specific gestures and converts them into English text using a trained deep learning model.

5.1.5 Application Layer

- This layer handles system interaction and output visualization:
- Backend Framework:
- Flask or FastAPI is used to manage API endpoints and connect UI with processing modules.
- User Interface (UI):
- Displays sign animations or translated text.
- Provides a simple and intuitive interface for both deaf and hearing users to communicate seamlessly.

5.1.6 Modularity and Extensibility

- Each processing module (e.g., ASR, NLP, Animation, Gesture Classifier) is decoupled, allowing:
- Easy replacement or upgrade of models.
- Plug-and-play integration of alternative services (e.g., Google Speech-to-Text).
- The system can be extended to:
- Support more sign languages (e.g., BSL, ISL).
- Add voice output for sign-to-speech.
- Implement real-time video call support using WebRTC.

5.1.7 Key Features of the Architecture

- Real-Time Processing: Ensures minimal latency in speech-to-sign and sign-to-text conversion.
- Modular Design: Allows easy updates or replacement of individual components.
- AI Integration: Utilizes state-of-the-art models for pose estimation, object detection, and NLP.
- Scalability: Architecture supports future expansion to multiple languages or sign dialects.
- Accessibility: Designed with inclusivity at its core, especially for use in virtual settings like video calls.

The system architecture supports real-time communication between deaf and hearing users through two integrated pipelines. Spoken input is processed via ASR and NLP, then converted to animated signs using pose estimation. Sign gestures are captured by a webcam, recognized using YOLOv5 and DWpose, and translated into text. A Flask or FastAPI backend connects all components, while the user interface displays results instantly. The modular design ensures flexibility, real-time performance, and accessibility across various digital communication platforms.

6 Results and Discussion

Figure 3 User Interface

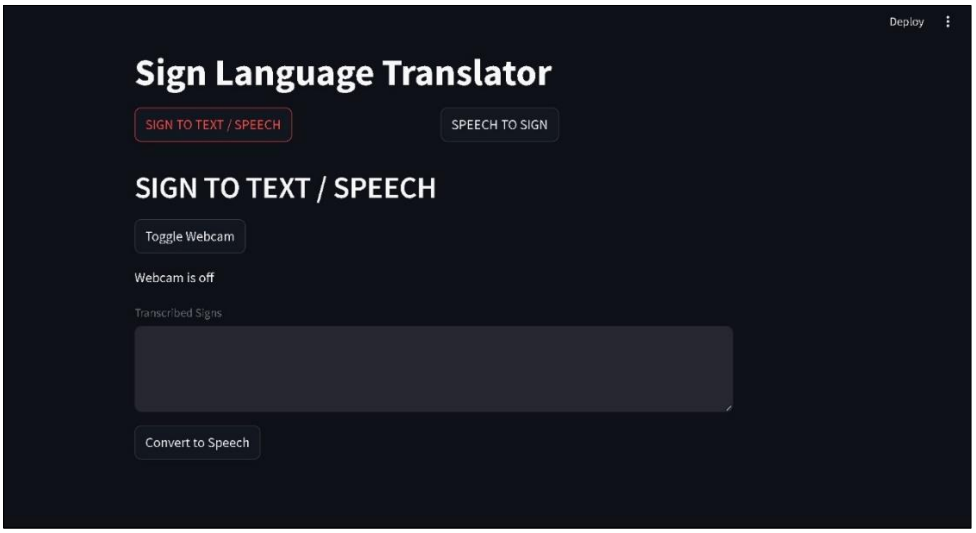

Figure 4 Sign to Text/Speech Generation

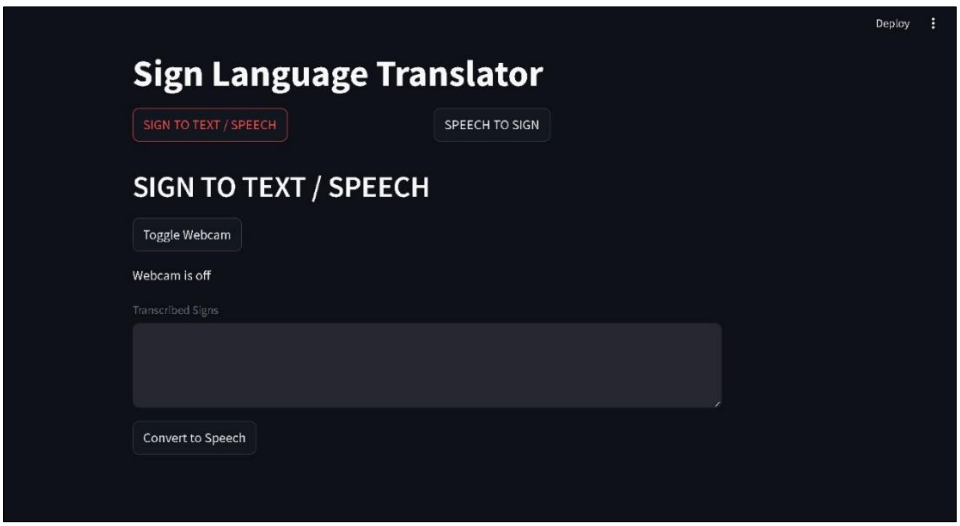

Figure 5 Speech to Sign Generation

Figure 6 Sign Response Generation (1)

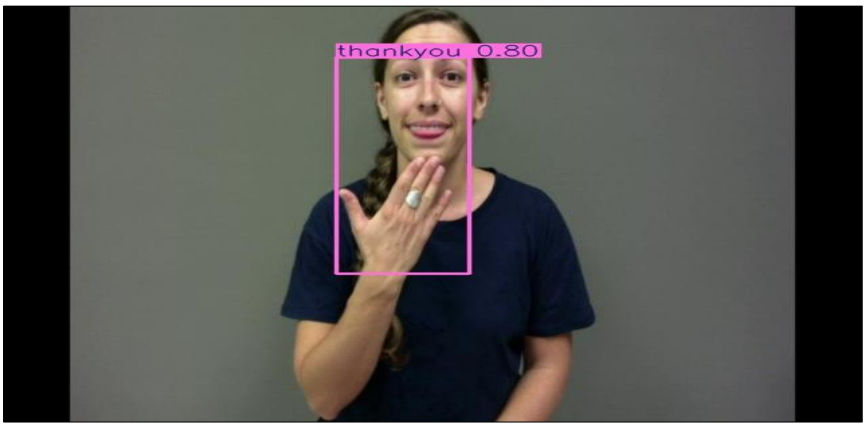

Figure 7 Sign Response Generation (2)

Figure 8 Sign Response Generation (3)

7 Conclusion

The GestureTalk system is a robust AI-driven solution that enables real-time, two-way communication between deaf and hearing individuals. Traditional tools—such as text-based chat applications or limited avatar systems—fail to fully support sign language or bidirectional interaction. GestureTalk bridges this gap by combining speech recognition, natural language processing (NLP), gesture detection, and 3D avatar animation into a seamless interface.

The system converts spoken input into American Sign Language (ASL) using a three-step process: speech-to-text via automatic speech recognition (ASR), text-to-ASL gloss through NLP, and animated sign generation using a pose-driven 3D avatar. In reverse, it uses a YOLO-based model and pose estimation (DWpose, MediaPipe) to detect and interpret hand gestures from live webcam input, translating them into readable English text.

Designed with both accuracy and real-time performance in mind, the system offers a scalable and user-friendly platform for inclusive digital communication. It performs reliably on standard computing hardware and supports practical deployment in video calls, online education, customer service, and more.

By facilitating direct interaction without the need for human interpreters, this project empowers the deaf community and promotes accessibility in digital spaces. Its modular architecture also provides a strong foundation for future improvements, such as multilingual sign support and emotion-aware animations.

Compliance with ethical standards

Disclosure of conflict of interest

There is no conflict of interest.

References

[1] Li, Dongxu, et al. (2020).Word-Level American Sign Language Recognition (WLASL). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[2] Shi, Y., Tian, Y., Wang, L., et al. (2023). DWpose: Dense Whole-Body Pose Estimation. arXiv preprint arXiv:2307.11879. https://arxiv.org/abs/2307.11879

[3] Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Augenstein, I. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005ss.11401. https://arxiv.org/abs/2005.11401

[4] Lee, M. Y. (2023). Building Multimodal Language Models with Retrieval-Augmented Generation. arXiv preprint arXiv:2305.03512. https://arxiv.org/abs/2305.03512

[5] Kassahun, Y., & Ramel, J.-Y. (2020). 3D Avatar-Based Sign Language Synthesis. In International Conference on Human-Computer Interaction. https://link.springer.com/chapter/10.1007/978-3-030-49062-1_15