Real-Time Unified Communication System Integrating Sign Language, Speech, and Text Using Machine Learning

Prof. Joyce Dsouza[1], Jyoti Patel[2], Vishal Dhaigude[3], Chinmay Dhoke[4], Shreyash Bari[5]
[1]*Assistant Professor, Department of Computer Engineering, Vidyavardhini College of Engineering and Technology, Vasai*
[2,3,4,5] *Student, Department of Computer Engineering, Vidyavardhini College of Engineering and Technology, Vasai*

Abstract*—Communication barriers between hearing and speech-impaired individuals and normal users create significant challenges in educational institutions, workplaces, healthcare environments, and public spaces. Most existing assistive technologies focus on a single communication mode such as sign language recognition or speech-to-text conversion, limiting their effectiveness for real-time, two-way communication. This paper presents a real-time unified communication system that integrates hand gesture recognition, speech processing, text translation, text-to-speech conversion, and video communication within a single web-based platform. Hand gestures are captured using a webcam and processed using OpenCV and MediaPipe to extract 21 hand landmark points. These landmarks are supplied to a machine learning classifier developed using Scikit-learn to perform gesture-to-text conversion. Voice input is converted into text using the Google Speech-to-Text API. The generated text is translated into the required language using Google Translate and converted into speech using a text-to-speech engine. WebRTC enables real-time face-to-face communication between users. Experimental results demonstrate that the system provides accurate, fast, and reliable multimodal communication, thereby promoting inclusivity and accessibility.

***Index Terms*—Sign Language Recognition, MediaPipe, Machine Learning, Multimodal Communication, Speech-to-Text, Text-to-Speech, WebRTC.**

I. INTRODUCTION

Communication is a fundamental aspect of human interaction that enables the exchange of ideas, information, and emotions. However, individuals with hearing and speech impairments face serious challenges while communicating with normal users, as they primarily depend on sign language for expression. Since sign language is not commonly understood by the general population, a significant communication gap exists between these groups. This barrier becomes more evident in environments such as educational institutions, workplaces, healthcare centers, and public places where effective communication is essential for participation and inclusion. Various assistive technologies have been developed to reduce this communication gap, including sign language recognition systems and speech-to-text applications. However, most existing solutions focus on a single mode of communication and do not provide a unified platform that supports multimodal interaction. These systems operate independently and fail to enable real-time, two-way communication using gestures, speech, and text together. As a result, users often rely on multiple tools, which reduces efficiency and limits practical usability in real-world scenarios.

To address these limitations, the proposed project titled Unified Communication System for Bridging Sign Language, Speech, and Text Using Machine Learning introduces a web-based platform that integrates multiple communication modes into a single system. The primary objective is to facilitate seamless and inclusive communication between hearing/speech-impaired individuals and normal users through gesture recognition, speech processing, text translation, text-to-speech conversion, and real-time video interaction. The system uses a webcam to capture hand gestures and processes the video frames using OpenCV and MediaPipe to detect 21 hand landmark points representing the structure of the fingers and palm. These landmark coordinates are used as feature inputs to a machine learning classifier developed with Scikit-learn to predict the corresponding gesture label.

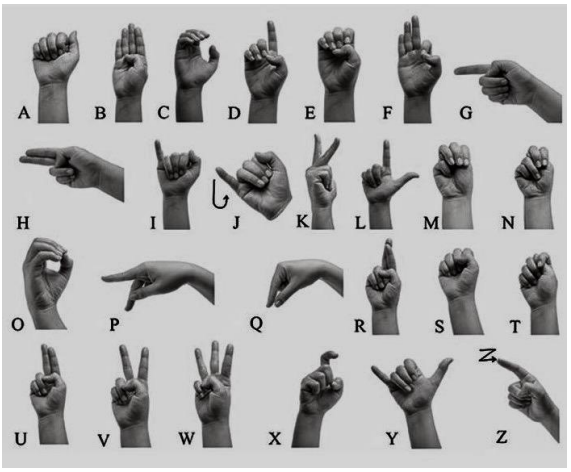

Fig.1. American Sign Language Gestures

The gesture recognition module is specifically trained on American Sign Language (ASL) alphabets and commonly used gesture words, allowing accurate conversion of standardized hand signs into text. In addition to gesture input, the system also accepts voice input through a microphone. The speech signal is converted into text using the Google Speech-to-Text service. The text generated from both gesture and speech modules is then passed to a translation module using Google Translate, enabling communication across different languages. The translated text is further converted into audible speech using a text-to-speech engine so that the receiving user can clearly hear the message. Furthermore, real-time face-to-face communication is supported using WebRTC, which enhances natural interaction between users. By combining gesture recognition, speech processing, translation, text-to-speech, and video communication into a single unified platform, the proposed system ensures smooth, real-time, and bidirectional communication. This integrated approach not only minimizes the communication gap between differently-abled and normal users but also promotes inclusivity, accessibility, and ease of interaction in everyday social and professional environments.

II. LITERATURE REVIEW

Numerous research efforts have been carried out to reduce the communication gap between hearing/speech-impaired individuals and normal users through sign language recognition and speech processing systems. Most of the earlier systems focused on recognizing hand gestures from images or video frames using computer vision and deep learning techniques. These approaches typically relied on processing raw image data using convolutional neural networks, which required large datasets, high computational resources, and longer training times.

Recent advancements introduced efficient hand-tracking frameworks such as MediaPipe, which detect precise hand landmark points without the need for complex image processing. Researchers have utilized these landmarks as feature inputs for machine learning classifiers, achieving faster and more accurate gesture recognition with reduced computational cost. Such feature-based approaches are more suitable for real-time applications compared to heavy deep learning models. Several studies have also explored systems that convert sign language gestures into text and speech output using machine learning algorithms. These systems demonstrated that landmark-based gesture recognition combined with classification techniques can provide reliable results for recognizing alphabet signs and commonly used words in American Sign Language (ASL). However, these works mainly concentrated on gesture-to-text or gesture-to-speech conversion as a standalone module.

On the other hand, speech recognition technologies have been widely researched to convert spoken language into text using cloud-based APIs and neural network models. Services such as Google Speech-to-Text have been integrated into communication tools to assist users in converting speech into readable text. Similarly, machine translation services like Google Translate and text-to-speech engines have been used independently in various assistive communication applications. Although significant progress has been made in gesture recognition and speech processing individually, very few systems attempt to integrate multiple communication modes into a single platform. Most existing solutions operate in isolation and do not support real-time, two-way, multimodal communication between users. Additionally, many systems lack real-time face-to-face interaction, which is essential for natural communication. The proposed work differs from existing approaches by integrating gesture recognition, speech-to-text conversion, language translation, text-to-speech output, and live video communication using WebRTC into one unified web-based system. This multimodal integration provides a comprehensive communication solution that addresses the limitations of previous single-mode assistive

technologies and enables inclusive, real-time interaction between hearing/speech-impaired individuals and normal users.

III. METHODOLOGY

The proposed system, Unified Communication System for Bridging Sign Language, Speech, and Text Using Machine Learning, is developed to reduce the communication gap between hearing/speech-impaired individuals and normal users. Unlike existing systems that focus on only one communication mode, the proposed approach integrates gesture recognition, speech recognition, text translation, text-to-speech conversion, and real-time video communication into a single web-based platform. The system allows users to communicate through hand gestures or speech, and the input is automatically converted into a form understandable by the receiving user.

1. Capturing Hand Gestures through Webcam:
A webcam captures live video frames of the user performing hand gestures. The video stream is processed frame-by-frame using OpenCV. Each frame is then passed to MediaPipe, which detects 21 hand landmark points representing the fingers and palm. These landmark coordinates are treated as numerical feature inputs for the machine learning model.

2. Gesture Classification using Machine Learning (ASL):
The extracted landmark features are given to a machine learning classifier developed using Scikit-learn. The classifier is trained using datasets of American Sign Language (ASL) alphabets and commonly used gesture words. Based on the pattern of landmark coordinates, the model predicts the corresponding gesture label. This process converts hand gestures into meaningful text and is referred to as Gesture-to-Text conversion.

3. Capturing Speech through Microphone:
The system also accepts speech input from the user through a microphone. The captured audio is sent to the Google Speech-to-Text service, which converts spoken words into text in real time. This process is known as Speech-to-Text conversion and does not require any local training.

4. Text Translation:
The text obtained from both gesture and speech modules is sent to a translation module using Google Translate. This enables the system to translate the recognized text into the preferred language of the receiving user, removing language barriers.

5. Text-to-Speech Conversion:
The translated text is provided to a text-to-speech engine, which converts the text into audible voice output. This ensures that the receiving user can hear the message clearly. The system also displays the text on the screen for better understanding.

6. Output Delivery to the Receiving User:
The final output is presented in two forms: text displayed on the user interface and speech output through speakers. This dual output format ensures accessibility and clarity of communication.

7. Real-Time Video Communication:
To make communication more natural and interactive, real-time video calling is integrated using WebRTC. This allows users to interact face-to-face while gesture and speech processing occur simultaneously in the background.

The modular design of the system ensures that each component operates independently while contributing to the overall communication pipeline. This methodology enables seamless, real-time, and bidirectional communication between hearing/speech-impaired individuals and normal users.

Fig.2 Shows the System architecture of the proposed system. The overall architecture of the proposed system demonstrates the integration of gesture recognition, speech processing, language translation, text-to-speech conversion, and real-time video communication into a unified platform. The system accepts input through a webcam, microphone, keyboard, and live video feed. The gesture pipeline extracts hand landmarks using MediaPipe and processes them using a machine learning classifier built with Scikit-learn. The speech pipeline converts voice input into text using Google Speech-to-Text. The recognized text is translated using Google Translate and converted into audible speech through a text-to-speech engine. Real-time video communication is enabled using WebRTC, allowing natural interaction between users. The final output is

delivered in the form of text, speech, and live video to the receiving user.

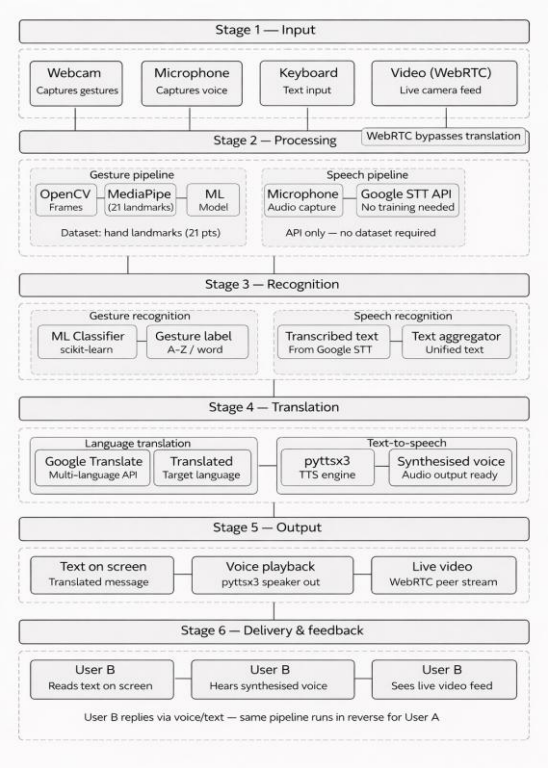

Fig.2. System Architecture of the Proposed Model

Fig.3 shows the Level-1 Data Flow Diagram of the proposed unified communication system. The diagram illustrates how data flows between external users and the internal processing modules of the system. User 1 provides input through the camera and microphone. The camera input is processed by the hand gesture detection module followed by gesture classification, where the recognized American Sign Language (ASL) gesture is converted into text. At the same time, the microphone input is processed by the speech-to-text module to generate textual output.

The text obtained from both modules is stored as a unified text output and forwarded to the text translation module. The translated text is then passed to the text-to-speech module to produce audible output for the receiving user. In addition, the system supports real-time video communication between users. The final output in the form of text, speech, and live video is delivered to User 2, ensuring smooth and effective two-way communication.

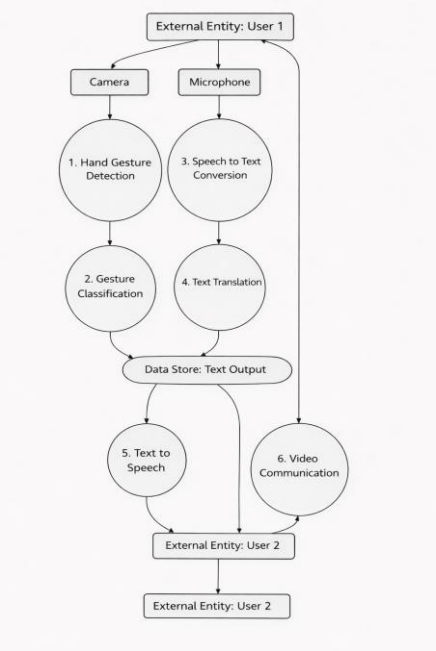

Fig.3.Level-1 Data Flow Diagram of the Proposed Model

IV. IMPLEMENTATION

The proposed Unified Communication System for Bridging Sign Language, Speech, and Text Using Machine Learning is implemented as a web-based application that integrates computer vision, machine learning, speech processing, translation services, and real-time communication. The system follows a modular architecture where each module performs a specific task while contributing to the overall communication pipeline.

The backend processing is carried out using Python, while the user interface is developed using web technologies. The gesture recognition module is implemented using OpenCV and MediaPipe for capturing and analyzing hand gestures. The classification of gestures is performed using a machine learning model built with Scikit-learn. Speech recognition is implemented using the Google Speech-to-Text service. The translation of text is handled through Google Translate, while voice output is generated using a text-to-speech engine. Real-time video communication between users is achieved using WebRTC.

The implementation is divided into the following modules:

1. Gesture Recognition and Processing Module:

The webcam continuously captures live video frames of the user's hand gestures. These frames are

processed using OpenCV and sent to MediaPipe, which detects 21 hand landmark points corresponding to the fingers and palm. The landmark coordinates are extracted and formatted as numerical feature vectors. These features are provided to the trained machine learning classifier, which predicts the corresponding American Sign Language (ASL) gesture. The predicted gesture is converted into text and displayed on the interface.

2. Machine Learning Classification Module:
A feature-based machine learning model is trained using landmark datasets of ASL alphabets and commonly used words. The model learns patterns of hand landmark positions and classifies the gesture into predefined labels. This approach reduces computational complexity compared to image-based CNN methods and ensures real-time performance.
3. Speech Recognition Module:
A microphone is used to capture speech input from the user. The audio signal is transmitted to the Google Speech-to-Text API, which converts spoken language into textual form in real time. The obtained text is forwarded for further processing.
4. Text Translation Module:
The text generated from both gesture and speech modules is passed to the translation service using Google Translate API. The system translates the text into the language preferred by the receiving user, enabling multilingual communication.
5. Text-to-Speech Output Module:
The translated text is given to a text-to-speech engine, which produces audible speech output. This ensures that the receiving user can understand the message through audio as well as text.
6. User Interface Module:
A web-based interface is developed to display the live video stream, recognized gesture text, translated text, and control options. The interface allows users to interact with the system easily and view outputs in real time.
7. Real-Time Video Communication Module:
WebRTC technology is integrated into the system to enable live video communication between users. This provides face-to-face interaction while the system simultaneously processes gesture and speech inputs in the background.
By integrating all these modules, the implemented system successfully performs gesture-to-text, speech-to-text, text translation, text-to-speech conversion, and real-time video communication within a single platform. The modular implementation ensures scalability, ease of maintenance, and efficient real-time performance.

V. RESULTS AND DISCUSSION

The proposed Unified Communication System was evaluated in real-time to analyze its effectiveness in recognizing hand gestures and converting them into meaningful text and audio output. The system successfully detects American Sign Language (ASL) gestures through a webcam and accurately converts them into textual form using landmark-based machine learning classification. The recognized text is further converted into audible speech using a text-to-speech engine, enabling the receiving user to clearly hear the conveyed message. The gesture recognition module performs efficiently by detecting 21 hand landmark points and predicting the correct gesture label with high accuracy. The system operates in real time with minimal delay, ensuring smooth and natural communication. In addition to gesture recognition, the speech input module converts spoken words into text, enhancing the multimodal communication capability of the system.

OUTPUT:

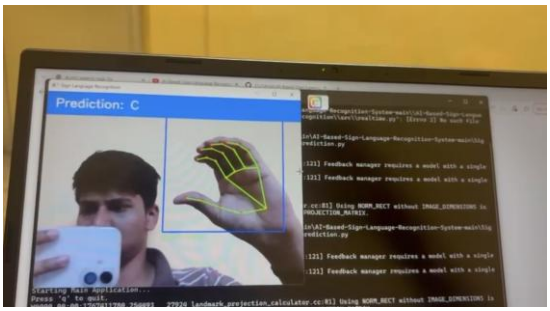

Fig.4. System Output

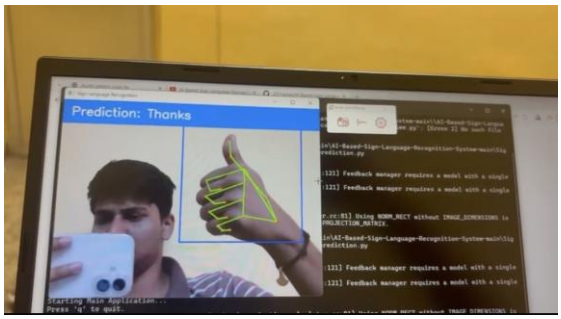

Fig.5. System Output

The final output generated by the system is delivered in two formats:

- Text displayed on the screen
- Audio output through speakers

This dual output format ensures that both hearing/speech-impaired users and normal users can understand the message effectively.

PERFORMANCE EVALUATION:

Parameter	Results
Gestures Recognition Accuracy	93%
Speech-To-Text Accuracy	95%
Response Time	0.8 Seconds
Hand Landmarks Used	21
Machine Learning Model	Scikit-learn Classifier
Sign Language Used	ASL

The results demonstrate that the system provides fast, accurate, and reliable communication assistance, making it suitable for real-time inclusive communication between users.

VI. ADVANTAGES OF THE PROPOSED SYSTEM

The proposed Unified Communication System provides several advantages compared to existing single-mode communication tools. By integrating gesture recognition, speech processing, text translation, text-to-speech conversion, and real-time video communication into one platform, the system offers an efficient and inclusive solution for interaction between hearing/speech-impaired individuals and normal users.

1. The system supports multimodal communication through hand gestures, speech, and text within a single platform.
2. Real-time processing of gestures and speech ensures smooth and natural communication with minimal delay.
3. The feature-based machine learning approach using hand landmarks reduces computational complexity and improves performance.
4. The system is trained using American Sign Language (ASL) gestures for accurate and standardized gesture interpretation.
5. The integration of a translation module removes language barriers between users.
6. Dual output in the form of text and audio makes communication easily understandable.
7. Real-time video communication enhances face-to-face interaction between users.
8. The web-based design allows access from any device with a camera, microphone, and internet connection.
9. The modular architecture of the system allows easy scalability and maintenance.
10. The system promotes an inclusive communication environment in educational, professional, and social spaces.

VII. APPLICATIONS

The proposed Unified Communication System can be applied in various real-world environments where effective communication between hearing/speech-impaired individuals and normal users is required. The multimodal capability of the system makes it suitable for practical deployment across different domains.

1. Educational Institutions:
   The system can assist hearing/speech-impaired students in communicating with teachers and classmates during lectures, discussions, and academic activities.
2. Healthcare Centers and Hospitals:
   Patients with hearing or speech impairments can communicate their symptoms and needs to doctors and medical staff without difficulty.
3. Workplaces and Offices:
   The system enables inclusive communication between differently-abled employees and their colleagues in professional environments.
4. Public Service Areas:
   It can be used at help desks, banks, railway stations, airports, and government offices to assist in public interactions.
5. Customer Service and Support Centers:
   The system can improve customer interaction where differently-abled individuals require assistance.
6. Home and Social Communication:
   It can be used for daily communication between family members, friends, and caregivers.
7. Online Meetings and Video Conferencing:
   Integration with video communication allows

use in virtual meetings for inclusive participation.

8. Training and Learning Sign Language: The gesture recognition module can help beginners learn and practice American Sign Language (ASL).
9. Assistive Technology Devices: The system can be integrated into kiosks or assistive devices designed for differently-abled individuals.
10. Multilingual Communication Environments: The translation feature allows communication between users speaking different languages.

VIII. CONCLUSION

The proposed Unified Communication System for Bridging Sign Language, Speech, and Text provides an effective solution to reduce the communication gap between hearing/speech-impaired individuals and normal users. Unlike existing approaches that focus on a single mode of interaction, the developed system integrates gesture recognition, speech-to-text conversion, text translation, text-to-speech synthesis, and real-time video communication within a single web-based platform.

The system accurately recognizes American Sign Language (ASL) gestures using hand landmark detection and a machine learning classifier, converting them into meaningful text and audio output. In addition, the speech recognition module transforms spoken input into text, enabling multimodal communication. The translated text and synthesized speech ensure that the conveyed message is clearly understood by the receiving user.

Experimental evaluation shows that the system performs with high accuracy and minimal response time, making it suitable for real-time communication scenarios. The modular architecture of the system allows easy scalability and future enhancements. Overall, the proposed system offers an inclusive, practical, and user-friendly communication solution that can be effectively applied in educational institutions, workplaces, healthcare centers, and public service environments to support differently-abled individuals.

IX. FUTURE SCOPE

The proposed Unified Communication System can be further enhanced with several improvements to increase its accuracy, usability, and practical applicability in real-world environments.

1. The gesture recognition module can be extended by training the model with a larger dataset of American Sign Language (ASL) gestures, including words and sentences for more expressive communication.
2. Advanced deep learning techniques such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) can be incorporated to improve gesture prediction accuracy.
3. Support for additional sign languages apart from ASL can be added to make the system usable in different regions and countries.
4. The system can be optimized to work efficiently on mobile devices as a dedicated application.
5. Noise reduction and voice enhancement techniques can be integrated to improve speech recognition accuracy in noisy environments.
6. A user-friendly graphical interface can be further enhanced for better accessibility and ease of use.
7. Cloud integration can be implemented for storing gesture datasets and improving scalability.
8. Security features such as user authentication and data encryption can be added for safe communication.
9. Integration with wearable devices or smart assistive devices can make the system more portable.
10. Real-time sentence formation and predictive text features can be added to speed up communication.

REFERENCES

[1] U. Rastogi, R. P. Mahapatra, and S. Kumar, “Advanced gesture recognition in Indian sign language using a synergistic combination of YOLOv10 with Swin Transformer model,” Scientific Reports, 2025.

[2] P. Nedungadi, G. Dileep, M. Geetha et al., “Vision transformer-powered conversational agent for real-time Indian Sign Language (ISL) accessibility,” Scientific Reports, 2025.

[3] V. G. Kaliyaperumal and P. A. Gopalan, “A deep neural network framework for dynamic two-handed Indian Sign Language recognition,” Sensors, 2025.

[4] B. V. Poornima and S. Srinath, "Frequency and spatial domain-based approaches for recognition of Indian Sign Language gestures," Indian Journal of Science and Technology, 2024.
[5] J. M. Joshi and D. U. Patel, "Dynamic Indian Sign Language recognition based on enhanced LSTM with custom attention mechanism," SSRG International Journal of Electronics and Communication Engineering, 2024.
[6] S. Shahi, M. Kumar, S. Verma, T. Usmani, and G. J. Patel, "Multilevel conversion of Indian Sign Language from gesture to speech," IJRASET, 2025.
[7] Y. Verma and R. S. Anand, "Gesture generation by the robotic hand for aiding speech and hard of hearing persons based on Indian Sign Language," Heliyon, 2024.
[8] A. Brahmavar, A. V. Mandalam, A. Umesh, B. Monali, and P. Agarwal, "Real-time speech generation of regional languages using Indian Sign Language," Computer and Electrical Engineering (SAGE), 2025.
[9] N. S., "AI-driven Indian Sign Language recognition using hybrid CNN-BiLSTM architecture," International Journal of Emerging Science and Engineering (IJESE), 2025.
[10] M. Geetha, A. R., N. A., D. A. S., and A. R., "Towards real-time recognition of continuous Indian Sign Language: A multi-modal approach using RGB and pose," IEEE Access, 2025.
[11] "Comprehensive survey on recent advances and challenges in sign language recognition systems," Discover Artificial Intelligence, 2025.
[12] "Indian Sign Language detection for real-time translation using machine learning," arXiv preprint, 2025.