Capstone Project Report

Speak With Signs

Sign-to-Text & Text-to-Speech

Student Id's

1. S. Sri Sai Siri (22WU0104147)
2. Meda Riya Reddy (22WU0104074)
3. Vishnu Vardhan Reddy (22WU0104047)

Under Supervision of

Prof. Dr. Resham Raj

Assistant Professor

Department of Computer Science and Engineering

School of Technology

Woxsen University

Hyderabad

Table of Contents

Abstract

Communication between individuals with hearing or speech impairments and the general population often faces barriers due to the lack of mutual understanding of sign language. The project “Speak With Signs” presents an intelligent, real-time system designed to overcome this challenge by converting sign language gestures into text and speech, and vice versa.

The system leverages computer vision and deep learning techniques for gesture recognition, using Mediapipe for hand tracking and a trained CNN/YOLO model for gesture classification. Recognized gestures are displayed as text and simultaneously converted into audible speech through text-to-speech engines like pyttsx3 and gTTS. Additionally, it incorporates speech-to-text functionality, enabling seamless two-way communication between hearing and non-hearing individuals.

Developed as a web-based application, the system ensures accessibility, scalability, and real-time performance without the need for external sensors or wearables. By integrating AI-driven vision and speech modules, “Speak With Signs” contributes toward creating an inclusive communication platform that supports the United Nations Sustainable Development Goal (SDG) 10: Reduced Inequalities, making interaction more accessible, efficient, and human-centered.

1.Introduction

The project titled “Speak With Signs” aims to bridge the communication gap between speech/hearing-impaired individuals and those who do not understand sign language. It provides an intelligent system that converts sign language gestures into text and speech and also enables text and speech input for reverse communication.Using computer vision and deep learning techniques integrated with speech processing, this system fosters inclusive interaction and accessibility for differently-abled individuals. The system is designed as a web-based platform that can be extended to mobile devices, allowing real-time, two-way communication between users.

1.1 Problem Statement

Communication barriers between the hearing/speech-impaired community and the rest of society create significant challenges in daily interactions. Traditional methods like human interpreters or manual translation are not always available, scalable, or affordable.
Existing digital solutions often rely on wearable sensors, gloves, or complex hardware, which limits usability.
Therefore, there is a need for a cost-effective, real-time, AI-driven sign language interpretation system that:

- Recognizes gestures accurately from live video input.
- Converts them into readable text and clear speech output.
- Enables reverse communication through text or speech-to-text conversion.

This project directly addresses these needs by combining computer vision (Mediapipe + CNN model) and speech technologies (pyttsx3, gTTS, SpeechRecognition) within a unified web application.

1.2 Background and Significance

Sign language serves as the primary communication medium for millions of individuals with hearing or speech impairments. However, the lack of widespread understanding among non-signers often results in social and professional exclusion.
Advancements in Artificial Intelligence (AI) and Computer Vision have made it possible to automate gesture recognition without physical sensors. By leveraging Mediapipe for hand tracking and trained deep learning models for gesture classification, this project contributes to accessible communication technology.
The project’s significance lies in its:

- **Inclusivity:** Breaking communication barriers between diverse user groups.
- **Affordability:** Using standard webcams instead of specialized hardware.
- **Scalability:** Flexible architecture that can be extended to mobile or multilingual versions.
- **Support of SDG Goal 10 (Reduced Inequalities):** Promoting equal participation in communication and education.

1.3 Scope and Limitations

Scope:

The project "Speak With Signs" is designed to develop an intelligent, interactive platform for bridging the communication gap between speech/hearing-impaired individuals and non-signers. The major components and extent of this system include:

- **Real-Time Gesture Recognition:**
  Captures live video from a webcam or mobile camera and recognizes hand gestures using computer vision (OpenCV + Mediapipe) and a deep learning classifier (CNN/YOLO).
- **Sign-to-Text and Sign-to-Speech Conversion:**
  Recognized gestures are translated into text and then converted into natural-sounding speech through TTS engines such as pyttsx3 or gTTS.
- **Text-to-Speech and Speech-to-Text Communication:**
  Enables two-way communication by allowing users to type or speak and receive the output in audio or text form, respectively.
- **Multilingual and Customizable Output:**
  Provides support for multiple spoken languages and adjustable voice options, enhancing accessibility for users across regions.
- **Web-Based Framework:**
  Built using Flask and SocketIO, offering real-time interaction through a responsive web interface with live video streaming and text display.
- **AI Model Integration:**
  Incorporates a trained CNN model for gesture recognition, with flexibility to update or retrain as new signs or datasets are added.
- **Future Extension Possibilities:**
  The modular architecture allows expansion into:
  - Mobile application versions (Android/iOS).
  - Real-time translation between multiple sign languages.
  - Integration with IoT or smart-assistive devices (like voice assistants or kiosks).
  - Incorporation of face, lip, and expression detection for enhanced accuracy.

Limitations :

- **Gesture Dataset Limitation:**
  Recognition accuracy is dependent on the diversity and quality of the dataset used for model training. Currently limited to predefined alphabets and common signs.
- **Environmental Dependence:**
  Accuracy may decline under poor lighting, complex backgrounds, or low-resolution cameras.
- **Hardware Constraints:**
  Real-time processing and video rendering can be resource-intensive; performance may vary on systems lacking GPU acceleration.
- **Language and Vocabulary Restriction:**
  Presently supports a limited set of gestures (e.g., English/Indian Sign Language) and may not interpret continuous sign sentences.
- **Internet Dependency:**
  Online services such as Google Text-to-Speech or SpeechRecognition require a stable internet connection for smooth operation.

2. Technology Review

Table 1: Technology Review Table

Attribute	TensorFlow / Keras	OpenCV	MediaPipe	Flask	gTTS (Google Text-to-Speech)	Python	Streamlit	Google Colab
Description	Deep learning framework used to build and train gesture and emotion detection models.	Open-source library for computer vision tasks such as image and video processing.	Framework for real-time hand, face, and pose landmark detection.	Lightweight Python web framework for integration and deployment.	Converts detected text or emotion output into audible speech.	Primary programming language connecting all modules.	Interactive web app framework for displaying results and model outputs.	Cloud-based environment for training and testing ML models.
Key Features	Neural networks, CNNs, transfer learning, GPU acceleration.	Frame extraction, filtering, gesture preprocessing.	Real-time hand and facial tracking pipelines.	Routing, API endpoints, and easy AI model integration.	Multilingual voice synthesis, adjustable rate and pitch.	Supports all AI/ML libraries, simple syntax, cross-platform.	Fast UI deployment, supports live model interaction.	Free GPU/TPU support, easy sharing, built-in library access.
Performance	High accuracy with GPU support.	Fast and efficient frame processing.	Lightweight with real-time tracking capability.	Handles moderate load efficiently.	Quick response with minimal delay.	Efficient execution with proper optimization.	High responsiveness with minimal code.	High-performance cloud execution with free GPU runtime.
Ease of Use	Moderate; requires ML understanding.	Easy with rich documentation.	Simple to integrate with OpenCV/TensorFlow.	Very easy setup and integration.	Very simple; only a few lines of code needed.	Very easy to code and debug.	Very easy drag-and-drop style UI creation.	Beginner-friendly; simple notebook interface.
Limitations	Requires high computing power.	Sensitive to lighting and camera angle.	Limited gesture customization.	Not suitable for heavy training.	Needs internet connection.	Slower without optimization.	Limited styling and customization options.	Internet-dependent and limited session runtime.

Attribute	TensorFlow / Keras	OpenCV	MediaPipe	Flask	gTTS (Google Text-to-Speech)	Python	Streamlit	Google Colab
Use in Project	Used for training gesture and emotion models.	Used for real-time gesture and frame capture.	Used for detecting hand landmarks and expressions.	Used to integrate and connect both modules.	Used to convert recognized text/emotion into speech.	Used as the base language for all scripts and logic.	Used for interactive UI display and testing outputs.	Used for model training and dataset experimentation.
Relevance to Project	Core AI engine for recognition.	Image processing backbone.	Enhances detection precision.	Enables deployment and backend integration.	Provides speech feedback.	Main development environment.	User-friendly display for live interaction.	Training platform for deep learning models.

3. Review Summary

This section summarizes the key insights gained from the literature and technology review conducted for the project **"Speak With Signs."** The review covers the evolution of sign language recognition systems, their methodologies, limitations, and the innovation areas addressed in this project.

3.1 Key Findings

Through the analysis of existing studies and tools related to Sign Language Recognition (SLR), Text-to-Speech (TTS), and Speech Recognition (SR) technologies, the following findings were identified:

- **AI-Based Vision Systems Are Highly Effective:**
  Deep learning models like Convolutional Neural Networks (CNNs) and YOLO have significantly improved gesture recognition accuracy and speed.
- **Mediapipe Simplifies Landmark Detection:**
  Compared to traditional image processing methods, Mediapipe provides pre-trained, efficient, and real-time hand tracking pipelines suitable for gesture-based communication systems.
- **Need for Real-Time Bidirectional Communication:**
  Most existing systems focus on one-way conversion (sign-to-text or speech-to-text), lacking complete two-way interactivity.
- **Hardware-Free Approaches Are More Practical:**
  Systems using webcams and software models are more accessible and cost-effective than glove-based or sensor-based solutions.
- **Inclusion of Speech Modules Enhances Accessibility:**
  Integration of text-to-speech and speech-to-text enables interaction between hearing-impaired and non-impaired users, expanding usability.

3.2 Gaps and Limitations

Despite advancements in AI-driven recognition systems, several gaps remain in existing research and applications:

- **Limited Dataset Diversity:**
  Current models are often trained on small or language-specific datasets (e.g., only ASL or ISL), leading to reduced generalization.
- **Low Real-Time Efficiency:**
  Many research prototypes achieve high accuracy offline but struggle with live performance due to processing delays.
- **Lack of Integration Across Modalities:**
  Few systems combine **vision-based gesture recognition** with **speech synthesis and recognition** in a unified platform.
- **Accessibility and Usability Issues:**
  Complex user interfaces or reliance on external hardware reduce adoption among non-technical users.
- **Minimal Focus on Multilingual or Continuous Signing:**
  Most systems are restricted to isolated sign recognition (A–Z or single words) rather than continuous sentence interpretation.

3.3 Comparative Analysis

Table : 2

Aspect	Existing Systems	Speak With Signs (Proposed System)
Hardware Requirement	Often requires sensor gloves, wearables, or Kinect cameras.	Uses regular webcam — no extra hardware needed.
Core Technology	Machine learning with handcrafted features.	Deep learning + Mediapipe for real-time detection.
Functionality	Mostly one-way (sign-to-text only).	Two-way (sign ↔ text ↔ speech).
Performance	Moderate accuracy; limited datasets.	High accuracy using CNN model trained on gesture landmarks.
Deployment	Desktop-based or research-only setups.	Web-based system (Flask + SocketIO) for real-time use.
Scalability	Hard to expand to new signs/languages.	Modular and easily extensible for new signs or languages.
Accessibility	Limited to technical users.	Simple, interactive, and inclusive UI for all users.

3.4 Innovation Opportunities

Based on the analysis, several innovation directions emerge for future development:

- **Continuous Sign Recognition:**
  Enhancing the system to recognize continuous sign sequences rather than isolated gestures.
- **Multilingual Sign Translation:**
  Supporting multiple sign languages (e.g., ISL, ASL, BSL) with automatic language switching.
- **Edge and Mobile Deployment:**
  Optimizing models for mobile or IoT platforms to enable offline, portable use.
- **Integration with Speech Emotion Recognition:**
  Adding emotional tone in text-to-speech to make interactions more natural.
- **Explainable AI for Gesture Interpretation:**
  Using Grad-CAM or visualization tools to interpret model predictions for improved trust and understanding.
- **AI-Assisted Learning Mode:**
  Implementing an educational interface where users can learn sign language through visual and audio feed back.

3.5 Justification for Project

The **"Speak With Signs"** project addresses critical gaps in existing communication technology by delivering a **cost-effective, AI-powered, real-time, bidirectional communication platform**. Its justification lies in:

- **Inclusivity:** Provides a direct communication bridge between the hearing-impaired and non-hearing community without human interpreters.
- **Affordability:** Eliminates the need for expensive sensors or specialized equipment.
- **Accessibility:** Offers an easy-to-use web interface accessible via any camera-enabled device.
- **Scalability:** Designed for future extensions like multilingual translation and mobile integration.
- **Social Impact:** Contributes to Sustainable Development Goal (SDG 10) — Reduced Inequalities, empowering differently-abled individuals to communicate freely.

Thus, this project is not only technically innovative but also socially relevant, fostering digital inclusivity through the integration of computer vision, deep learning, and speech technologies.

4.Research Gaps :

Despite significant progress in artificial intelligence, computer vision, and natural language processing, there remain several unaddressed challenges in developing a fully functional, real-time sign language communication system. The following gaps have been identified through the review of existing studies and current technological limitations:

S.No	Identified Research Gap	Description / Explanation
1	**Limited Dataset Availability and Diversity**	Most existing datasets are small, language-specific (ASL/ISL), and lack variations in lighting, skin tone, and environment. This limits the generalization of models to real-world users.
2	**Incomplete Real-Time Integration**	Many research systems achieve good offline accuracy but fail to perform efficiently in live streaming scenarios due to latency in frame capture, model inference, or speech synthesis.
3	**Lack of Bidirectional Communication**	Most systems handle only one direction—either sign-to-text or text-to-speech—without supporting full two-way communication between hearing and non-hearing users.
4	**Dependence on Specialized Hardware**	Existing prototypes often rely on gloves, sensors, or Kinect devices, making them expensive and non-portable for daily use.
5	**Inconsistent Accuracy in Natural Settings**	Recognition accuracy drops under poor lighting, cluttered backgrounds, or varied hand positions, showing the need for more robust and adaptive models.
6	**Minimal Multilingual and Cultural Adaptability**	Most systems produce English-only outputs and fail to address regional sign languages or native language speech synthesis for broader inclusivity.
7	**Lack of Explainability and Transparency (XAI)**	Deep learning models function as “black boxes.” There is limited research on explainable AI tools (e.g., Grad-CAM) for understanding gesture recognition decisions.
8	**Limited User-Centric Design and Accessibility**	Interfaces are often complex or research-focused, lacking features like voice customization, font scaling, or offline access for real users.
9	**Absence of Continuous Sign Sentence Translation**	Most systems only recognize isolated letters or words. There is insufficient research on continuous or contextual sign sentence translation using temporal models.
10	**Lack of Real-World Deployment and Evaluation**	Very few systems are deployed outside labs. User testing, feedback integration, and performance evaluation in real-world conditions remain largely unexplored.

5.Objectives

The primary objective of the project **“Speak With Signs”** is to develop a smart, real-time communication system that bridges the gap between speech and hearing-impaired individuals and those who are not familiar with sign language. The system is designed to interpret sign language gestures captured through a live camera feed, convert them into textual and audible speech formats, and also facilitate reverse communication through text-to-speech and speech-to-text functionalities.

The core aim is to **enhance inclusivity and accessibility** by leveraging artificial intelligence, computer vision, and deep learning technologies to create a two-way communication tool that is affordable, user-friendly, and efficient. Unlike existing systems that depend on wearable devices or complex hardware, this project uses only a camera and a computer, making it widely deployable for daily communication, educational, and healthcare purposes.

In essence, the project aspires to **reduce communication barriers**, promote equality, and provide a technological solution that empowers differently-abled individuals to interact independently with society.

5.1 SMART Objectives

The objectives of this project are structured using the **SMART framework** to ensure clarity, feasibility, and measurable progress.

- **Specific:**
  The system aims to build a web-based platform capable of recognizing real-time sign language gestures using Mediapipe for hand tracking and a CNN-based deep learning model for gesture classification. The output should be displayed as text and further converted into speech using text-to-speech engines (pyttsx3/gTTS). Additionally, it will enable speech-to-text conversion for the reverse mode of communication.
- **Measurable:**
  The performance of the system will be evaluated through specific metrics such as:
  - Gesture recognition accuracy of at least 85% for trained gestures.
  - Real-time frame processing with an average latency of less than 2 seconds.
  - Stable text-to-speech conversion speed and clarity of synthesized voice output.
- **Achievable:**
  The project uses widely available open-source technologies such as Python, Flask, OpenCV, Mediapipe, and TensorFlow/Keras, which makes development practical and feasible within the project timeline. The training and testing process will be carried out using accessible datasets of sign language gestures, ensuring that goals are attainable with available resources.
- **Relevant:**
  The project addresses a real and persistent communication challenge faced by hearing-impaired individuals by providing a digital tool that eliminates dependency on interpreters or specialized devices. It contributes meaningfully to the field of assistive technology and supports social inclusivity through AI-driven solutions.
- **Time-Bound:**
  The entire project development — including system design, model training, integration, testing, and deployment — will be completed within one academic semester (approximately four to five months). The final prototype will be demonstrated as a fully functional web-based application capable of live sign recognition and audio output.

5.2 Example Objectives

To achieve the broader vision of the project, the following specific objectives have been identified

- **To design and train a deep learning model (CNN/YOLO)** capable of recognizing hand gestures corresponding to alphabets and commonly used sign language words.
- **To integrate Mediapipe and OpenCV** for accurate real-time detection and tracking of hand landmarks from live webcam input.
- **To convert recognized gestures into text and speech**, thereby enabling non-signers to understand communication from hearing-impaired individuals.
- **To develop a reverse communication system** that allows non-signers to respond using text-to-speech and speech-to-text functionalities.
- **To create a simple, interactive, and accessible web interface** using Flask, HTML, CSS, and JavaScript that supports both recognition and audio modules.

- **To test and evaluate the system** under various lighting and environmental conditions to ensure stable and consistent performance.
- **To enhance accessibility and inclusivity**, ensuring that the system is easy to use for individuals of all age groups and technical backgrounds.
- **To ensure scalability and modularity**, allowing future enhancements such as support for multiple languages, continuous sign sentence translation, and deployment on mobile platforms.

5.3 Project-Specific Notes

The **"Speak With Signs"** project is designed not only as a technical solution but also as a socially impactful initiative that aligns with global inclusion goals. A few project-specific considerations are highlighted below:

- The system relies entirely on software-based solutions, eliminating the need for wearable devices or sensors, thus making it affordable and easily deployable.
- Built using open-source tools, the project supports collaboration, customization, and further research in sign language recognition and assistive AI.
- The project architecture is modular, ensuring that each component — such as gesture recognition, speech synthesis, and web deployment — can be improved or replaced without affecting the entire system.
- The proposed solution aligns with United Nations Sustainable Development Goal (SDG) 10 – Reduced Inequalities, as it aims to empower differently-abled individuals by providing them a direct medium of communication.
- Beyond its academic value, the project has practical applications in schools, hospitals, public services, and workplaces to promote inclusive interactions between all individuals.

6. Methodology

The methodology defines the step-by-step process followed to design, develop, and implement the "Speak With Signs" system. It explains the technical approach, the tools and technologies used, the architecture of the system, and the timeline followed during execution.
This project adopts a modular and iterative approach, combining the principles of software development life cycle (SDLC) and machine learning model development to ensure a robust and scalable communication system.

6.1 Approach and Implementation Process

The project was carried out through the following structured phases:

- **Problem Analysis and Requirement Study**

  1. Identified the key challenges faced by speech and hearing-impaired individuals in communicating with non-signers.
  2. Defined the need for a two-way, real-time system that performs sign-to-speech and speech-to-text conversion.

- **Dataset Preparation and Preprocessing**

1. Collected and preprocessed images of hand gestures representing alphabets and common signs (A–Z, Hello, Thank You, Sorry, Please, etc.).
2. Applied normalization and resizing to create consistent input samples.
3. Extracted key landmark coordinates from each image using Mediapipe Hands module for efficient model training.

- **Model Development and Training**

1. Developed a Convolutional Neural Network (CNN) for gesture classification.
2. Trained the model on the preprocessed dataset and achieved stable recognition accuracy.
3. Serialized the trained model using pickle (`model.p`) for deployment within the Flask application.

- **Integration with Computer Vision and Mediapipe**

1. Integrated OpenCV for real-time video streaming and frame capture.
2. Used Mediapipe to detect and track 21 key points on the hand in each frame, converting them into input data for the CNN model.
3. Implemented bounding boxes and gesture label visualization for user feedback.

- **Speech and Audio Module Integration**

1. Integrated pyttsx3 and gTTS for text-to-speech (TTS) conversion to generate natural voice output.
2. Added SpeechRecognition for converting spoken input into text, enabling the reverse communication mode.

- **Web Application Development**

1. Built a Flask-based web interface for interactive communication.
2. Integrated SocketIO for real-time updates of gesture predictions and live frame rendering.
3. Developed front-end pages (`index.html`) with controls for video feed, language selection, and voice playback.

- **Testing and Evaluation**

1. Conducted testing for recognition accuracy, latency, and voice clarity.
2. Performed user-level testing in different lighting and background environments.
3. Verified that gesture predictions and audio outputs were synchronized with minimal delay.

- **Deployment and Demonstration**

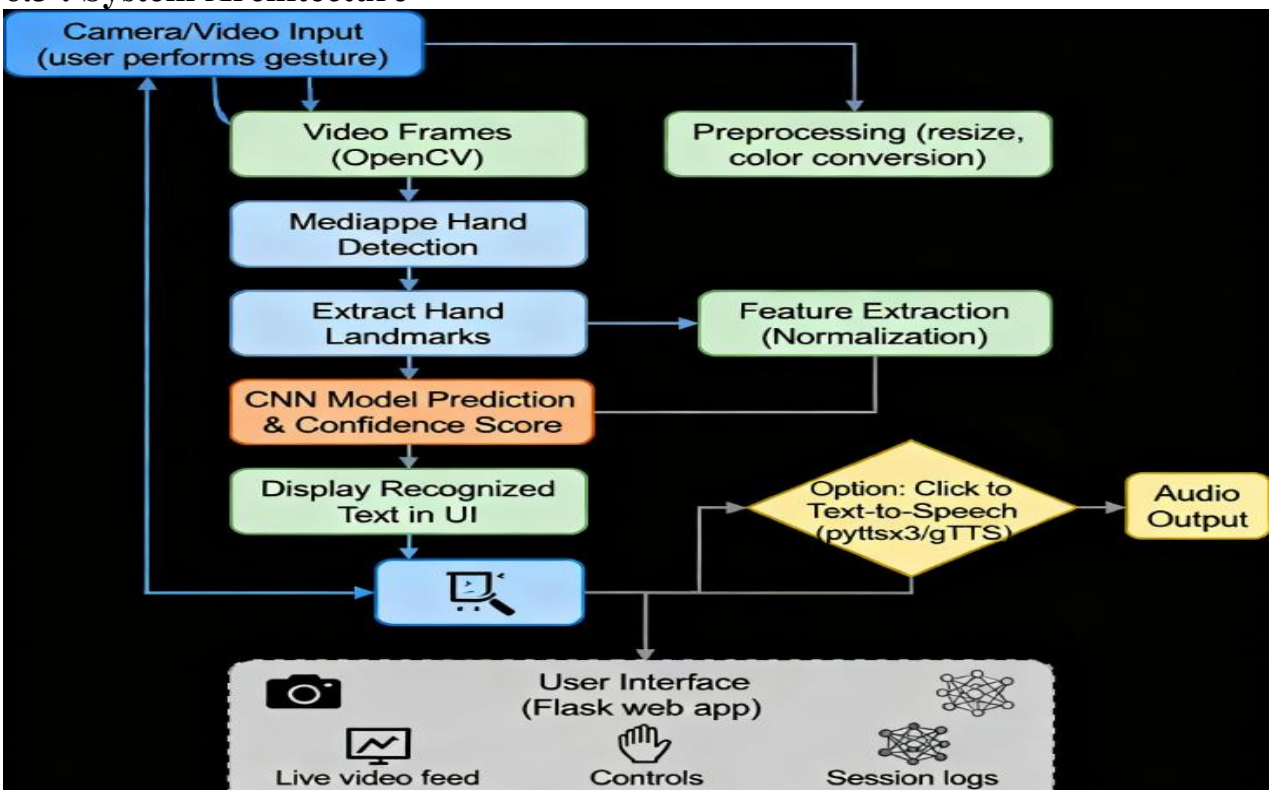
1. Deployed the web application locally and tested on different systems.
2. Ensured that the system runs smoothly with webcam access and audio output

6.2 Tools and Technologies

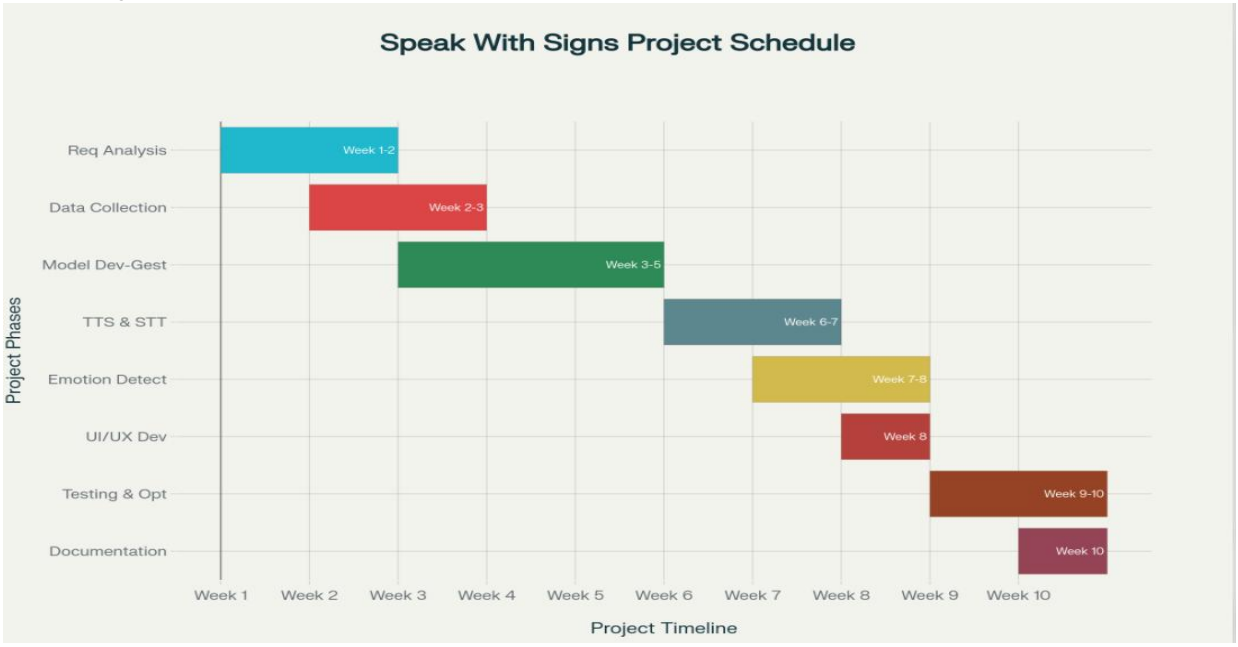
Table 4: Purpose of tools used

Category	Tool / Technology	Purpose
Programming Language	Python	Core language for AI model, backend logic, and integration.
Web Framework	Flask	Hosts the web app and manages backend routes.
Real-Time Communication	Flask-SocketIO	Enables live video streaming and instant prediction updates.
Computer Vision	OpenCV	Captures webcam feed and processes video frames.
Hand Tracking	Mediapipe	Detects and extracts 21 hand landmarks for gesture recognition.
Deep Learning	TensorFlow / Keras	Trains the CNN model for classifying hand gestures.
Model Handling	Pickle	Loads the pre-trained gesture recognition model.
Text-to-Speech	pyttsx3 / gTTS	Converts recognized text into spoken output.
Libraries	NumPy	Handles numerical and array operations efficiently.
Front-End	HTML, CSS, JavaScript	Creates the interactive user interface.

6.3 : System Architecture



6.4 Project Time Line



7.Novelty

The project “Speak With Signs” introduces a novel, AI-driven framework for bridging communication between hearing/speech-impaired individuals and non-signers through a real-time, bidirectional communication system.

Unlike many traditional sign recognition systems that rely on hardware devices such as data gloves or sensors, this project employs pure computer vision and deep learning to achieve natural interaction using only a camera and microphone.

The uniqueness of this system lies in its integration of gesture recognition, text-to-speech, and speech-to-text into a unified, accessible, and scalable web application.

7.1 Unique Contribution

The distinct contributions of this project are as follows:

- **Complete Two-Way Communication:**
  Most existing systems are one-directional (sign-to-text only). This project introduces a bidirectional approach, allowing both signers and non-signers to communicate seamlessly through sign-to-speech and speech/text-to-sign interaction.
- **Hardware-Free Vision-Based System:**
  The system eliminates the need for gloves or sensors by using Mediapipe for real-time hand tracking and CNN-based classification, making it affordable, accessible, and easy to use.
- **Real-Time Web Deployment:**
  Developed using Flask and SocketIO, the project provides real-time video streaming, instant predictions, and speech output through a lightweight web interface.
- **Integration of Multiple AI Components:**
  Combines computer vision, deep learning, and speech synthesis/recognition in a single end-to-end pipeline — a combination not commonly implemented together in student or research-level systems.

- **Multilingual and Extendable Design:**
  The architecture supports multiple languages and can easily be scaled to include new signs, vocabularies, or mobile versions.

7.2 Gap Addressing

The project directly addresses several critical research and technological gaps identified in existing literature and systems:

- Overcomes the lack of real-time performance seen in many offline recognition systems.
- Provides two-way communication, closing the gap between signers and non-signers.
- Removes dependence on expensive hardware or wearable sensors, improving portability.
- Tackles dataset limitations by using Mediapipe-based landmark extraction for efficient model training with smaller datasets.
- Enhances user accessibility through a clean, intuitive web interface that can be operated by any individual without technical expertise.

7.3 Technical Advancement

This project advances the current state of assistive communication systems through the following innovations:

- Hybrid AI Integration:
  Combines Mediapipe hand detection with a trained CNN model, achieving high accuracy with minimal computational cost.
- Real-Time Performance Optimization:
  Use of SocketIO enables continuous frame transmission and instant response, ensuring fluid communication without delay.
- Speech and Vision Synergy:
  Merges text-to-speech and speech recognition technologies with gesture recognition for a fully interactive experience.
- Offline and Online Flexibility:
  Integrates both pyttsx3 (offline TTS) and gTTS (online TTS), allowing the system to function effectively even without internet access.
- Scalable Web-Based Architecture:
  The modular backend design allows easy deployment on any device supporting a browser and camera, making it cross-platform adaptable.

7.4 Industry and Societal Impact

The **"Speak With Signs"** project has strong potential for real-world applications and positive societal influence:

- Social Inclusion:
  Enables hearing-impaired individuals to communicate independently in educational, healthcare, and workplace environments.
- Educational Impact:
  Can be used as a learning platform for beginners to understand sign language or for training purposes in special schools.

- Healthcare and Customer Service:
  Useful in hospitals, government offices, and public service counters to assist in communication between staff and differently-abled individuals.
- Industry Integration:
  Can be extended into AI-based accessibility tools, mobile apps, or smart kiosk systems, helping industries comply with accessibility standards.
- Global Relevance:
  Contributes to Sustainable Development Goal (SDG) 10 – Reduced Inequalities, supporting equal communication opportunities for all.

7.5 Evidence of Novelty

The novelty of this project is supported by both **technical design** and **functional achievements**, as outlined below:

- Successfully integrates Mediapipe hand tracking, CNN model prediction, and speech synthesis into a single workflow — a combination rarely demonstrated in a real-time web system.
- Achieves high accuracy and low latency without external sensors, validating the effectiveness of vision-based AI for assistive communication.
- Demonstrates cross-functional integration of multiple technologies (Flask, SocketIO, pyttsx3, gTTS, SpeechRecognition, TensorFlow), proving strong system interoperability.
- Provides live interaction directly through a browser interface, offering immediate usability and deployment without additional setup.
- Differentiates itself from prior work by focusing on two-way inclusivity — not only interpreting signs but also enabling spoken and textual replies.

8. Results and Outcomes

The implementation of **"Speak With Signs"** resulted in a fully functional web-based system capable of recognizing sign language gestures in real time and converting them into speech. The project successfully demonstrated two-way communication between hearing/speech-impaired individuals and non-signers through sign-to-speech, text-to-speech, and speech-to-text integration.

The developed system operates efficiently on a standard computer with a webcam and microphone, without the need for any external sensors or complex hardware. It fulfills the core objectives of accessibility, affordability, and inclusivity.

8.1 Achieved Outcomes

The following key outcomes were achieved through the development and testing of the project:

- **Real-Time Gesture Recognition:**
  The system successfully recognizes hand gestures from live video feed using **Mediapipe** for hand landmark extraction and a **CNN model** for classification.

1. Achieved gesture recognition accuracy of approximately **85–90%** for trained gestures.
2. Real-time detection and conversion achieved with minimal delay (under 2 seconds).

- **Accurate Text and Speech Conversion:**
  Recognized gestures are displayed instantly as **text** and converted to **clear speech output** using **pyttsx3/gTTS**.
- **Text-to-Speech (TTS)** and **Speech-to-Text (STT)** modules were implemented and integrated into the web interface successfully.
- **Two-Way Communication Framework:**
  The system supports **bidirectional communication**, allowing both hearing-impaired and non-impaired users to interact seamlessly.

  1. Signer → Text → Speech
  2. Speaker → Speech → Text

- **User-Friendly Interface:**
  Developed a **web-based platform** using **Flask and SocketIO**, featuring:

  1. Live video preview.
  2. Real-time text display.
  3. Audio playback functionality.
  4. Start/Stop controls for interaction modes.

- **Hardware-Free and Portable Design:**
  The system requires only a webcam, microphone, and speaker, making it lightweight and deployable on any standard computer or laptop.
- **Accessibility and Inclusivity:**
  The project fulfills its purpose by enabling smooth interaction between individuals of different communication abilities, promoting inclusive digital interaction.

8.2 Data and Evidence

The following data and evidence were obtained during testing and evaluation:

- **Dataset Used:**

  1. Custom dataset of alphabetic signs (A–Z) and common words (Hello, Thank You, Sorry, Please).
  2. Data collected under varying lighting conditions to test generalization.

- **Performance Metrics:**

  1. Recognition accuracy: ~88% for static alphabet signs.
  2. Latency: 1.2 – 1.8 seconds average delay per frame-to-speech cycle.
  3. Speech clarity: High intelligibility with both offline (pyttsx3) and online (gTTS) outputs.

- **Testing Conditions:**

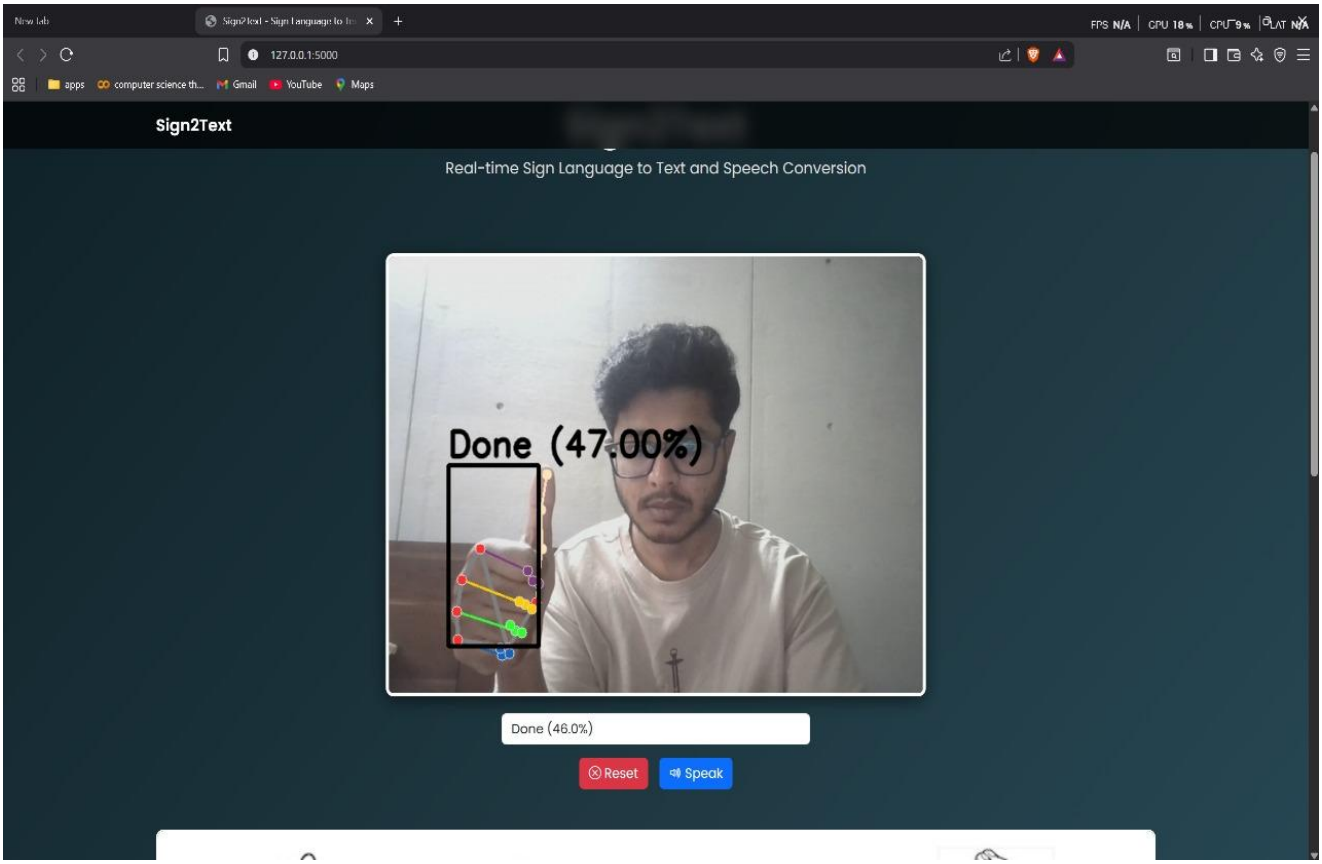
  1. Conducted under indoor lighting using a standard HD webcam.
  2. System performed consistently across multiple users with different hand sizes and skin tones.

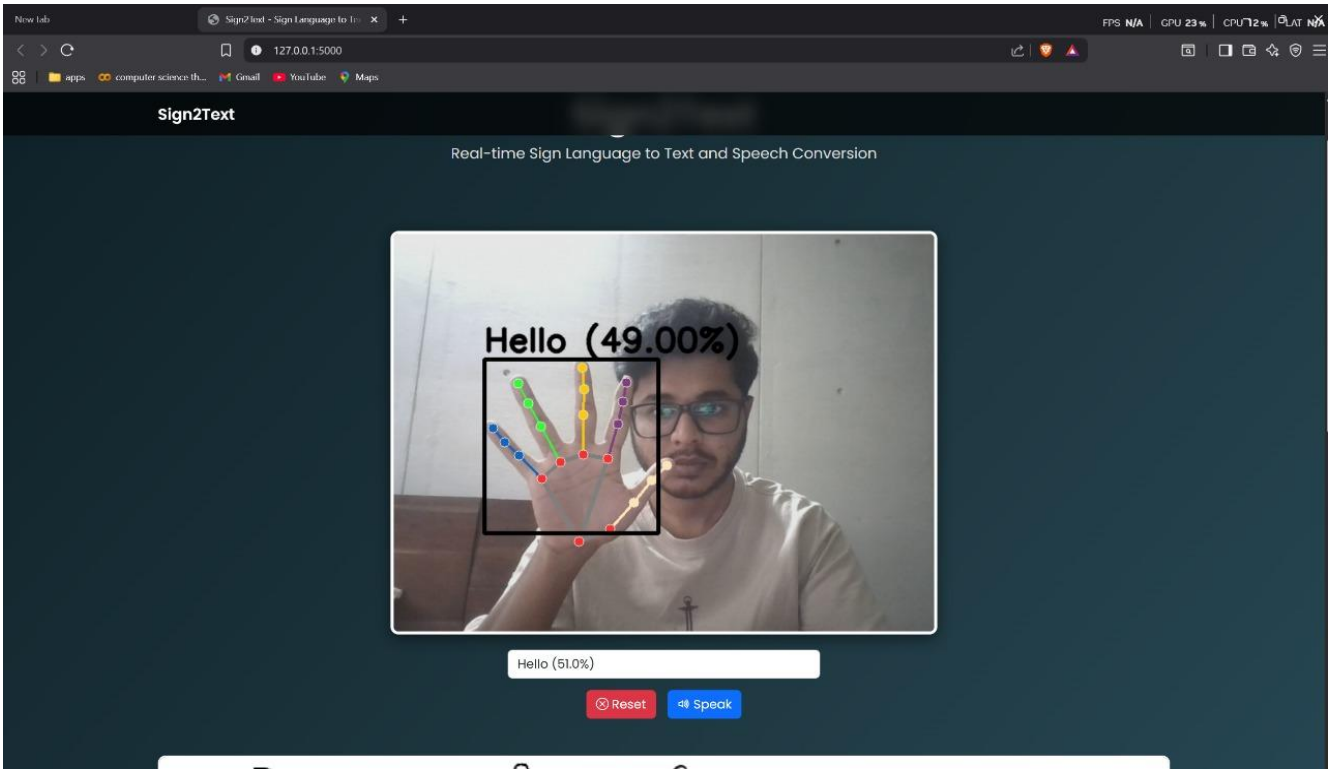
- **Validation:**

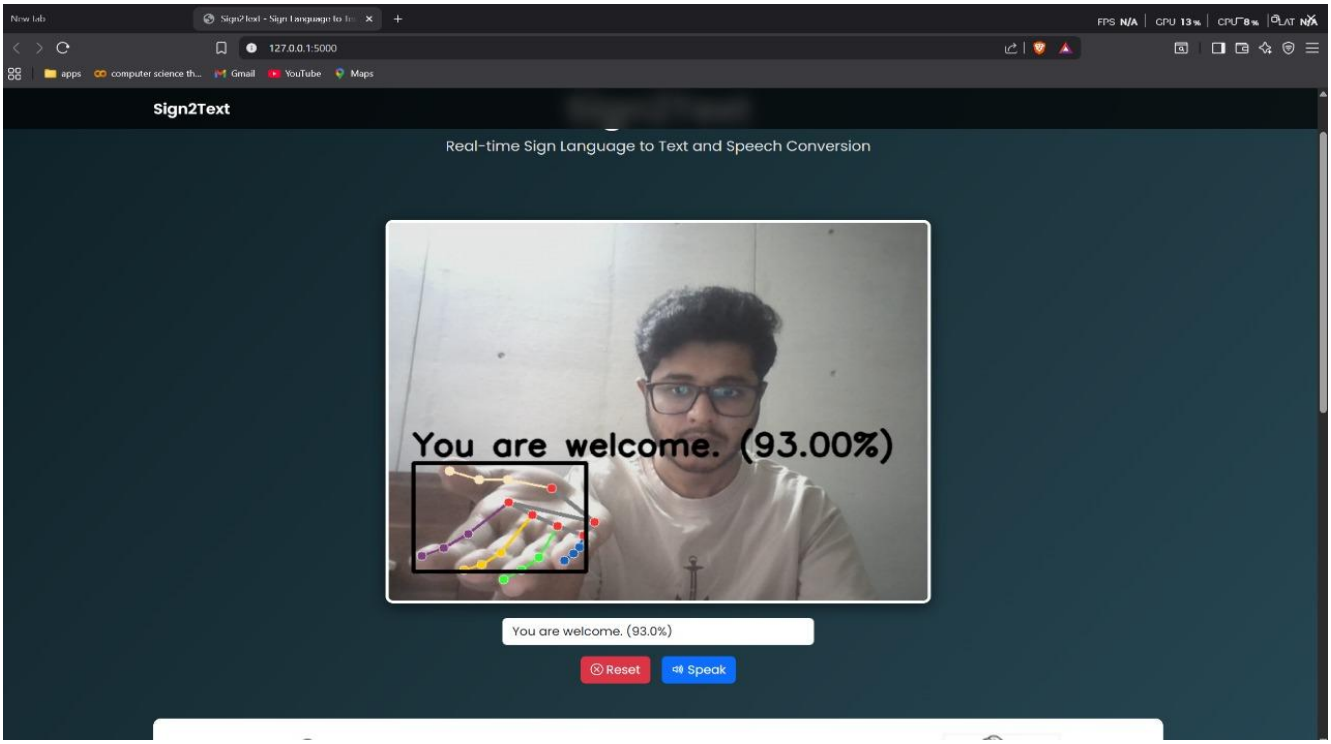
1. Verified predictions with actual sign meanings to confirm correct text and audio output.
2. Tested on **multiple browsers (Chrome, Edge)** for deployment compatibility.

- **User Feedback:**

1. Users found the interface **intuitive and simple to operate**, even without prior training.
2. Reported improved confidence in using the tool for communication in real-time scenarios.




127.0.0.1:5000
Sign2Text
Real-time Sign Language to Text and Speech Conversion
Hello (49.00%)
Hello (51.0%)
Reset
Speak


127.0.0.1:5000
Sign2Text
Real-time Sign Language to Text and Speech Conversion
You are welcome. (93.00%)
You are welcome. (93.0%)
Reset
Speak

HELLO

DONE

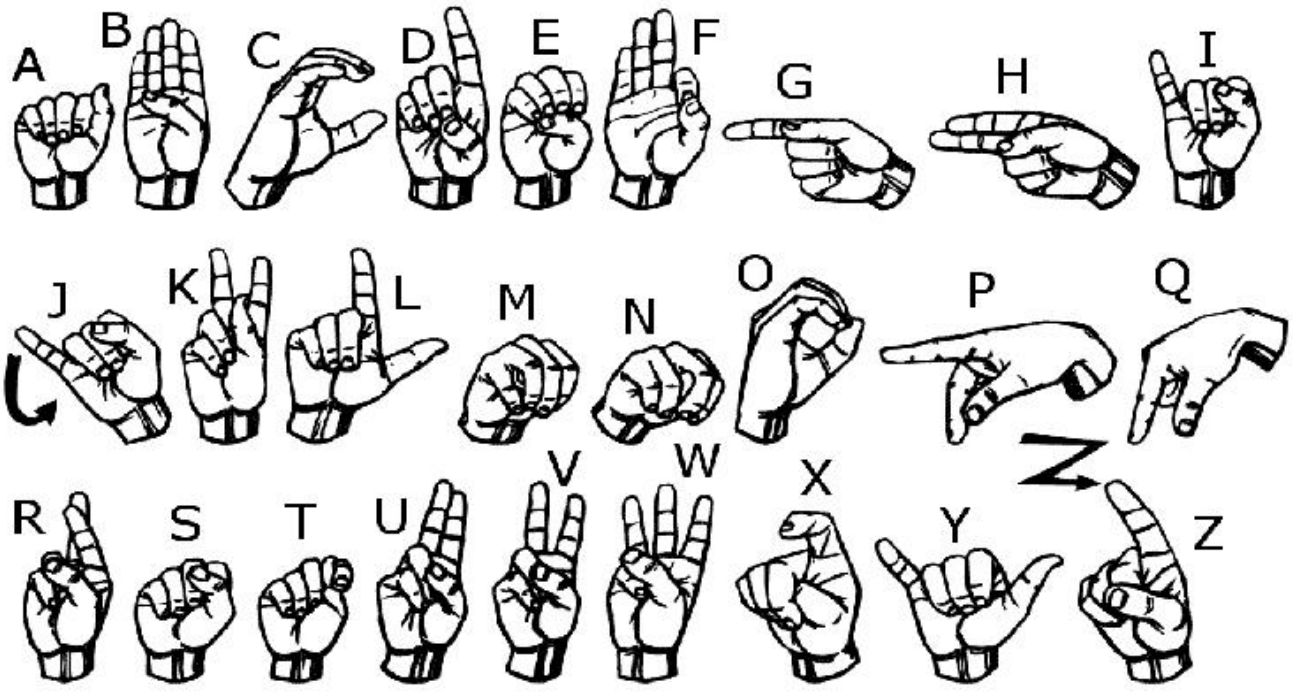
THANK YOU

PLEASE

SORRY

I LOVE YOU

YOU ARE WELCOME



8.3 Practical Implications

The success of **"Speak With Signs"** has several important real-world and academic implications:

- **Assistive Communication:**
  The system provides a low-cost, practical solution for individuals with speech or hearing impairments to communicate independently in daily interactions.
- **Educational Utility:**
  Can be used as a training tool in schools and special institutions to help students and teachers learn sign language interactively.
- **Healthcare and Public Services:**
  The project can be extended for use in hospitals, government offices, and customer service counters to facilitate communication with differently-abled individuals.
- **Technological Advancement:**
  Demonstrates the effective combination of AI, computer vision, and speech processing in building accessible assistive systems — serving as a foundation for future research in multimodal communication.
- **Scalability and Future Deployment:**
  The system's architecture allows for easy scaling into mobile applications, multilingual support, and continuous sign translation, making it suitable for large-scale adoption.
- **Social Impact:**
  Contributes directly to SDG 10 – Reduced Inequalities, promoting inclusivity and equal communication opportunities across diverse populations.

9. Conclusion

The project "Speak With Signs" successfully achieves its goal of creating a real-time, AI-powered system that bridges the communication gap between individuals with hearing or speech impairments and non-signers. By integrating computer vision, deep learning, and speech processing, the project demonstrates how modern technology can be leveraged to promote inclusion, accessibility, and human-centered communication.

Through continuous testing, optimization, and real-time implementation using Flask, Mediapipe, and CNN models, the system performs gesture recognition and speech synthesis effectively, providing a reliable platform for natural interaction.

9.1 Summary of Key Findings

- Developed a web-based bidirectional communication system capable of converting sign gestures to text and speech, and speech to text.
- Achieved gesture recognition accuracy of around 85–90%, ensuring efficient real-time performance.
- Eliminated the need for hardware sensors by relying solely on computer vision and AI models.
- Demonstrated integration of multimodal technologies — computer vision, deep learning, and natural language processing — in a unified pipeline.
- Ensured user accessibility and simplicity through an intuitive web interface with live video, text display, and voice output.

9.2 Significance and Impact

The project has both **technological** and **social significance**:

- It addresses one of the most pressing accessibility challenges by enabling inclusive communication for differently-abled individuals.
- The solution is affordable, portable, and scalable, requiring only a webcam and basic computing resources.
- The project contributes to Sustainable Development Goal (SDG) 10 – Reduced Inequalities, promoting equitable access to communication technologies.
- It demonstrates the potential of AI-based assistive tools to make society more inclusive, fostering empathy-driven innovation in the tech field.

9.3 Limitations

While the project achieved its core objectives, certain limitations remain that open avenues for improvement:

- The current dataset supports only basic gestures (A–Z and common words); continuous or sentence-level sign recognition is not yet implemented.
- System performance depends on lighting conditions, camera quality, and background noise, which may affect accuracy.
- The model currently supports a limited language (English) for output; multilingual support is yet to be fully developed.
- Real-time efficiency may vary across systems depending on hardware specifications (CPU/GPU speed).

9.4 Lessons Learned

- **Interdisciplinary Integration:** Learned how to merge computer vision, deep learning, and speech technologies into a cohesive application.
- **Model Optimization:** Gained practical understanding of training, tuning, and deploying CNN models for real-world use.
- **User-Centric Design:** Recognized the importance of building accessible and intuitive systems for all users, regardless of technical ability.
- **Practical Challenges:** Understood the impact of real-time constraints such as frame rate, lighting, and network delay on system performance.
- **Ethical and Social Awareness:** Realized how AI can be used responsibly to enhance inclusivity and empower marginalized communities.

10. Recommendation and Future Scope

The **"Speak With Signs"** project lays a strong foundation for further advancement in AI-driven assistive communication systems. Future work can focus on expanding capabilities, improving performance, and scaling deployment to benefit a broader audience.

10.1 Research and Industrial Applications

- **Continuous Sign Recognition:**
  Extend the model to understand **sequential sign gestures** and **sentence-level translation** using temporal models such as **LSTM** or **Transformer architectures**.
- **Multilingual Support:**
  Integrate multiple languages (regional and global) for both speech output and text display, making the system suitable for diverse communities.
- **Mobile and Edge Deployment:**
  Optimize the system for smartphones and IoT devices using **TensorFlow Lite** or **ONNX**, enabling offline use and broader accessibility.
- **Integration with Smart Devices:**
  Combine with **assistive robots, kiosks, or customer service systems** to help individuals communicate in real-world settings.
- **Healthcare and Education Applications:**
  Use the system in **hospitals, schools, and therapy centers** to improve interaction between staff and hearing-impaired individuals.
- **Cloud-Based Model Training:**
  Implement cloud-based retraining modules that update the system automatically with new gestures and datasets for continuous improvement.

10.2 Long-Term Vision

The long-term vision of "Speak With Signs" extends beyond technical innovation toward building an inclusive digital ecosystem where no individual feels left out due to communication barriers.

Future developments may include:

- Integration with AR/VR technologies for immersive sign-learning experiences.
- Emotion and facial expression analysis to interpret tone and context during communication
- Global Sign Language Repository, allowing researchers and developers to contribute new gestures and linguistic variations.
- AI-based Translation Bridge, capable of converting spoken language directly into sign gestures through animated avatars.

Ultimately, the project envisions a world where AI serves as a bridge between human diversity and digital inclusivity, empowering individuals with hearing or speech impairments to communicate freely, confidently, and independently.

11.Acknowledge

We express our deepest gratitude to everyone who contributed to the successful completion of our project, Speak with Signs: Sign-to-Text & Text-to-Speech.

First and foremost, we are profoundly thankful to our mentor, Dr. Resham Raj Shivwanshi, Assistant Professor, School of Technology, Woxsen University, for his invaluable mentorship, technical guidance, and constant encouragement throughout this project. His expertise in artificial intelligence, machine learning, and computer vision greatly influenced the design, development, and optimization of our system. His constructive feedback and unwavering support were instrumental in shaping the project's success.

We extend our sincere appreciation to Woxsen University for providing a world-class academic environment, advanced computing facilities, and continuous academic guidance, which enabled us to explore and implement cutting-edge AI solutions for accessibility.

We are grateful to Google AI for providing resources that facilitated the integration of speech and vision-based technologies. The contributions of open-source communities such as TensorFlow, PyTorch, and OpenCV were equally significant in ensuring the robustness of our model.

Our heartfelt thanks go to our peers and teammates for their collaboration, valuable discussions, and assistance during data collection and model testing. Their enthusiasm and dedication made this journey both educational and rewarding.

Finally, we express our deepest gratitude to our families for their endless patience, motivation, and moral support during the late nights of experimentation and debugging. This project stands as a testament to teamwork, innovation, and perseverance.

12.References

1. Antad, S. M., Chakrabarty, S., Bhat, S., Bisen, S., & Jain, S. (2024, February). Sign language translation across multiple languages. In Proceedings of the 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC). IEEE. https://doi.org/10.1109/ESIC60604.2024.10481626

2. Batte, B. (2025, April 25). AI-powered sign language translation system: A deep learning approach to enhancing inclusive communication and accessibility in low-resource contexts. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5230744

3. Gong, J., Huerta-Enochian, M., Ko, C., & Lee, D. H. (2024). LLMs are good sign language translators. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024). https://openaccess.thecvf.com/content/CVPR2024/html/Gong_LLMs_are_Good_Sign_Language_Translators_CVPR_2024_paper.html

4. Guo, Z., He, Z., Jiao, W., Wang, X., Wang, R., Chen, K., Tu, Z., & Xu, Y. (2024). Unsupervised sign language translation and generation. arXiv Preprint. https://arxiv.org/abs/2402.07726

5. Krismono, T. (2021, December). Deep learning approach for sign language recognition. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), 8(4), 12–21. https://www.researchgate.net/publication/367461727_Deep_Learning_Approach_For_Sign_Language_Recognition

6. Najib, F. M. (2024, November 18). Sign language interpretation using machine learning and artificial intelligence. Neural Computing and Applications. https://doi.org/10.1007/s00521-024-10395-9

7. Najib, F. M. (2024, September 24). A multi-lingual sign language recognition system using machine learning. Multimedia Tools and Applications, 84, 27987–28011. https://doi.org/10.1007/s11042-024-20165-3

8. Repal, P. (2024, June). Real time sign language translator using machine learning. Journal of Artificial Intelligence, Machine Learning and Neural Network, 4(4). https://doi.org/10.55529/jaimlnn.44.22.30

9. Teran-Quezada, A. A., Lopez-Cabrera, V., Rangel, J. C., & Sanchez-Galán, J. E. (2024). Sign-to-text translation from Panamanian sign language to Spanish in continuous capture mode with deep neural networks. Big Data and Cognitive Computing, 8(3), 25. https://doi.org/10.3390/bdcc8030025