Innovative AI Solution for Real-Time Translation of Dynamic Sequential Sign Language

Harsh Tyagi, Gaurav Yadav, Saksham Pandey, Jaivardhan Saini, Ms. Vani Rastogi
Meerut Institute of Engineering & Technology, Meerut

***Abstract*–This paper introduces a system to convert Indian Sign Language (ISL) gestures into text and speech,helping to address communication challenges for the Mutism impaired. The system employs state-of-the-art computer vision and natural language processing techniques with MediaPipe to perform pose detection and key point extraction from 30-second video segments on upper-body landmarks. These keypoints are processed by the Long Short-Term Memory (LSTM) network for categorizing gestures to ISL words or phrases. For perfect coherence with the grammar of English, it has the refinement output to be carried out through the transformer-based language model Gemma Model. It operates on four different stages, that is video capture, extraction of key points, classification through LSTM to ISL gestures, and conversion to speech from text using edge_tts library. The system's ability to combine pose estimation with sequential modeling accurately interprets the ISL gestures, bringing out grammatically correct text and speech. The system manifests high accuracy in gesture-to-text translation and real-time performance, allowing for flawless communication between mutism-impaired individuals and others. Such innovation enhances inclusivity in that it converts ISL gestures into accessible text and speech, contributing to broader participation in society and breaking through barriers of communication.**

***Index Terms*– Gemma, ISL, LSTM, Mediapipe, Mutism, Transformer**

I. INTRODUCTION

Communication is fundamental to human interaction, enabling individuals and groups to share ideas, emotions, and intentions through mutually understood signs and rules. It is essential for personal growth, societal coexistence, and problem-solving in environments that prioritize care and creativity. Effective communication fosters respect, trust, and collaboration, which are vital for sustainable development. However, the lack of effective communication remains a significant barrier, particularly for individuals with physical challenges such as hearing or speech impairments. Communication barriers, including inattentiveness, arguments, language differences, and physical limitations, contribute to misunderstandings and conflicts in relationships. According to global studies, millions of individuals face such challenges, with hearing and speech impairments being among the most common disabilities. The need for bridging the communication gap between mustism-impaired individuals and the normal population has been a recurring focus of research.

Existing studies have explored methods to enhance communication, particularly in the context of gesture recognition. Gesture recognition has gained significant attention as a means of understanding and interpreting human actions. For instance, research on scene segmentation using deep learning achieved a classification accuracy of 53.8%, emphasizing the potential of deep learning techniques for image-based tasks. However, such methods often lack the specificity required for sign language interpretation.

Other studies have investigated unsupervised learning techniques to categorize human actions. For example, Juan Carlos demonstrated the application of probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) for action recognition. While effective for general human actions, these methods do not address the unique structural and grammatical complexities of sign languages.

Further advancements in Indian Sign Language (ISL) recognition have shown promising results[3][4]. Techniques leveraging unsupervised learning with Gaussian models and autoencoders have successfully recognized ISL alphabets, achieving up to 98.20% accuracy with large training sets. These studies highlight the importance of robust datasets and effective feature extraction methods. Despite such progress, converting ISL

words and phrases into meaningful sentences remains an area requiring further exploration to ensure seamless communication for hearing-impaired individuals.

This research aims to address these gaps by presenting a system that translates ISL gestures into text and speech. The system combines pose estimation, sequential modeling, and transformer-based language models to provide an end-to-end solution for gesture-to-speech conversion. By leveraging modern tools such as MediaPipe[5][3] for key point extraction and LSTMs[1][3] for sequential classification, this approach seeks to bridge the communication gap effectively and inclusively.

II. METHODOLOGY

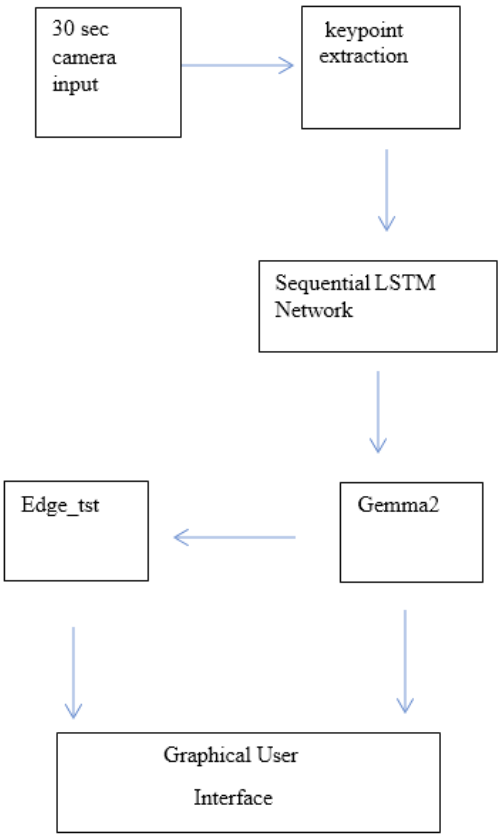

Fig1. Architecture

The objective of this study was to design a system capable of translating Indian Sign Language (ISL) gestures into meaningful text and speech. The system combines pose estimation, sequential modeling, and language refinement techniques to achieve accurate gesture interpretation and natural language generation. The methodology is divided into four main stages:

1. Segmenting Captured ASL Signed Gestures

The process begins with capturing 30-second video segments containing ASL gestures. These videos are segmented into individual frames, and MediaPipe is used to extract upper-body key points, including hand, arm, and facial landmarks. This segmentation isolates the relevant gestures, converting them into a structured time-series representation.

2. Extracting Features from the Segmented Gestures

The extracted key points represent spatial and temporal features of the gestures. These features serve as inputs for the classification model. By focusing on the Region of Interest (ROI)—such as hand movements and facial cues—the system ensures that only meaningful data is processed, reducing noise and improving accuracy

3. Classifying the Gestures

The time-series key point data is passed through a Long Short-Term Memory (LSTM) network. The LSTM model is trained to recognize ISL words and phrases by learning sequential patterns in the gestures. This supervised learning approach leverages a dataset of annotated ISL gestures, enabling the system to output corresponding ASL text.

4. Synthesizing Text and Speech

The classified text, typically in ASL grammatical structure, is refined using a transformer-based language model (Gemma Modal(lite version og gemini)). This step translates ISL text into meaningful English sentences by understanding context and grammar. The refined text is then converted into speech using a Text-to-Speech (TTS) system, providing a complete gesture-to-speech pipeline

Segmentation of Captured ISL Signed Gestures

The segmentation process in this study aims to convert video input into smaller, structured data representations that facilitate analysis. Segmentation involves dividing the video frames into distinct segments, enabling the identification of meaningful components like hand shapes, motion paths, and boundary outlines (e.g., curves, arcs, and lines).

Proposed Network

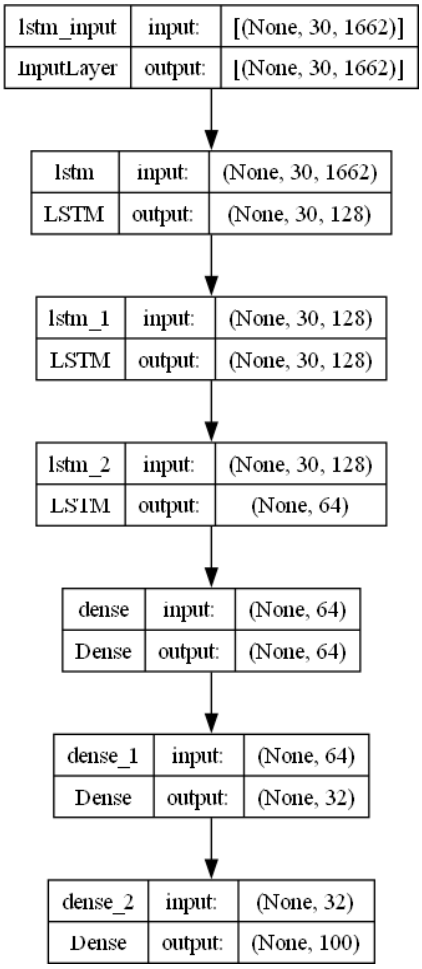

Fig2.LSTM network

Fig3. Proposed Architecture

Video Frame Selection: The captured 30-second ASL video segments are divided into individual frames. MediaPipe is then used to extract key landmarks representing the upper body, focusing on hands, arms, and facial features.

Feature Localization: MediaPipe assigns unique spatial coordinates to each detected landmark. This data forms the basis for understanding gesture patterns, isolating each hand gesture as an analyzable unit.

Binary Representation: To further refine segmentation, key points are analyzed for motion and spatial relationships, reducing data complexity while retaining gesture-relevant information. Frames are cropped and resized to a standardized format to ensure consistency.

Feature Extraction of the Segmented Images

ASL gestures rely heavily on upper body movements, particularly involving the head, shoulders, hands, and elbows, while other body parts are less significant. To extract meaningful data, the system identifies key points representing high-contrast regions such as edges and corners. These features are selected to ensure they are non-redundant, informative, and application-specific.

Regions of Interest (ROI) Extraction:

Rectangular Regions of Interest (ROI) are used, characterized by four corner points connected to form a boundary. The ROI focuses on the most informative sections of the frame, such as hand movements, to facilitate gesture analysis.

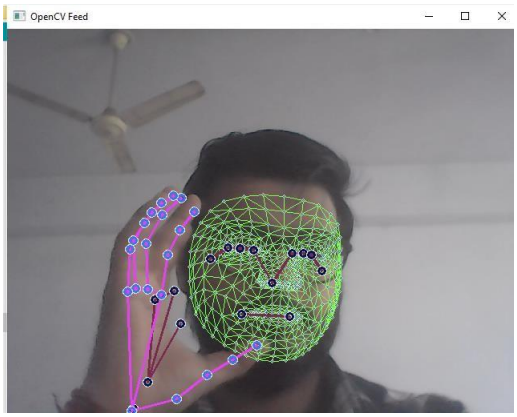

Fig.4 Key point detection with mediapipe

LSTM Network

The system leverages a Long Short-Term Memory (LSTM) network to classify ISL gestures based on the extracted body key points. These key points, which represent the head, shoulders, elbows, and hands, are obtained from MediaPipe and converted into a NumPy array for further processing.

Procedure of Workflow:

Key Point Representation: Each gesture is represented as a sequence of key points over time, capturing the movement patterns critical to ISL gestures.

Input to LSTM: The NumPy array of sequential key points serves as the input to the LSTM network. The LSTM effectively models the temporal dependencies between the frames, learning the dynamic patterns associated with each ISL gesture.

Classification: The LSTM network processes the input sequence and outputs a corresponding text word for each gesture. For example, a sequence representing the gesture for "hello" is classified as the text word "hello."

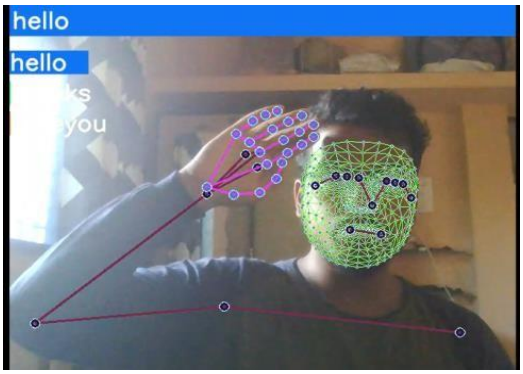

Fig.5 Sign Prediction

Text and Speech Synthesis of Classified Gestures
Once the gestures are classified into text, the next step is to synthesize text and speech for communication.

Text-to-Speech (TTS):

Text Processing: The classified gesture text is refined using Gemma Modal, which ensures grammatical correctness and meaningful sentence construction.

Speech Synthesis: The refined text is input into a TTS engine(edge_tts) capable of converting it into audible speech. The TTS system:

- Converts text into phonemic representations.
- Transforms phonemes into waveforms.
- Outputs waveforms as human-like speech.

III. RESULTS AND DISCUSSION

Sample videos of various Indian Sign Language (ISL) gestures were recorded using a camera for a duration of 30 seconds per segment. Each video captured specific ISL gestures performed by individuals, providing a dataset for training and testing the model. A total of approximately five hundred (500) video samples were collected, with each sign represented by 15 to 00 samples to ensure a robust dataset. This approach was adopted to enhance the model's ability to accurately classify gestures and minimize misclassification rates.

Segmentation and Key Point Extraction

The recorded videos were split into 30-second segments. Using MediaPipe, key points corresponding to body landmarks such as hands, elbows, and shoulders were extracted from each video frame. This conversion of video input into body key points was crucial for reducing computational complexity and focusing on the essential features for gesture recognition.30-second video is segmented because any sign language doesn't have static signs means If a person is communicating in ISL he/she will move its body for next sign and so on.

Sequential Classification
The extracted key points were converted into NumPy arrays and passed through a Long Short-Term Memory (LSTM) neural network. The LSTM model effectively captured the temporal dependencies in the gesture sequences, classifying the gestures into corresponding ISL text labels with high accuracy.

Text Refinement and Speech Synthesis
The classified text labels were refined using the Gemma Modal, which converted the raw outputs into meaningful sentences. This step ensured the linguistic correctness and contextual accuracy of the generated text. Finally, the refined text was converted into speech output using a Text-to-Speech (TTS) engine, enabling auditory communication of the ISL gestures.

Fig6.Final Output

IV. DISCUSSION

The developed system demonstrates a promising approach to recognizing and converting Indian Sign Language (ISL) gestures into meaningful text and speech outputs. The integration of advanced techniques such as video segmentation, key point

extraction using MediaPipe, and sequential classification with an LSTM model highlights the potential for achieving high accuracy in gesture recognition.

Key Observations:

Effectiveness of Video-Based Input: Using 30-second video segments as input proved to be highly effective for capturing the dynamic nature of ISL gestures. Unlike still images, video input preserves temporal information, which is critical for accurately interpreting sequential movements.

Key Point Extraction and Computational Efficiency: The MediaPipe framework efficiently extracted body key points, significantly reducing the complexity of processing raw video frames. By focusing on relevant landmarks, the system minimized redundant data, making it computationally efficient without sacrificing accuracy.

Temporal Dependencies with LSTM: The LSTM model excelled in recognizing gesture sequences by capturing temporal dependencies between key frames. Its ability to classify gestures into text labels accurately validates the suitability of recurrent neural networks for this task.

Text Refinement with Gemini Modal: The use of the Gemini Modal for refining raw text outputs added a layer of linguistic and contextual correctness, ensuring that the generated sentences were not only accurate but also coherent. This step was vital for bridging the gap between gesture recognition and human communication.

V. CONCLUSION & FUTURE PROSPECTS

This ISL recognition system leverages video inputs, MediaPipe for key point extraction, and LSTM for sequential classification, with Gemini Modal ensuring contextually accurate text outputs. It lays the groundwork for future advancements, focusing on real-time functionality, scalability, and enhanced accessibility for hearing and speech-impaired individuals.

REFERENCES

[1] Sheth, Pranav, Sanju Rajora, and Yogeshvari Makwana. "Sign Language Recognition Application Using LSTM and GRU (RNN)."

[2] Khan, Rafiqul Zaman, and Noor Adnan Ibraheem. "Hand gesture recognition: a literature review." *International journal of artificial Intelligence & Applications* 3, no. 4 (2012): 161..

[3] Ilham, Amil Ahmad, and Ingrid Nurtanio. "Dynamic Sign Language Recognition Using Mediapipe Library and Modified LSTM Method." *International Journal on Advanced Science, Engineering & Information Technology* 13, no. 6 (2023).

[4] Vijayalakshmi, P., and M. Aarthi. "Sign language to speech conversion." In *2016 international conference on recent trends in information technology (ICRTIT)*, pp. 1-6. IEEE, 2016.

[5] Lugaresi, Camillo, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang et al. "Mediapipe: A framework for building perception pipelines." *arXiv preprint arXiv:1906.08172* (2019).

[6] Team, Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot et al. "Gemma 2: Improving open language models at a practical size." *arXiv preprint arXiv:2408.00118* (2024).