Real-Time Bidirectional Translation System Between Text and Indian Sign Language Using Deep Learning and NLP Techniques

1st Niyati V. Gaonkar
Dept. of Information Technology
V.E.S.I.T, Mumbai, India
niyatig26@gmail.com

2nd Vishal R. Gori
Dept. of Information Technology
V.E.S.I.T, Mumbai, India
vishalgori2208@gmail.com

3rd Anket G. Kadam
Dept. of Information Technology
V.E.S.I.T, Mumbai, India
anketkadam61@gmail.com

4th Soham H. Nimbalkar
Dept. of Information Technology
V.E.S.I.T, Mumbai, India
soham.nimbalkar08@gmail.com

5th Mr. Manoj Sabnis
Dep. HOD, Dept. of Information Technology
V.E.S.I.T, Mumbai, India
manoj.sabnis@ves.ac.in

***Abstract*—In this paper, we present a real-time translation system that bridges the communication gap between the hearing and non-hearing communities. Our system converts English text to Indian Sign Language (ISL) and vice versa, using Natural Language Processing (NLP) techniques and deep learning-based gesture recognition. The system supports video-based gesture recognition for ISL and provides accurate text translations in real-time. This study addresses the technical challenges involved, including feature extraction from gestures and translating complex ISL sentences using neural networks like LSTM.**

Keywords- Indian Sign Language (ISL), Sign Language Translation, Gesture Recognition, Deep Learning, LSTM Model, Mediapipe Holistic, Text-to-Sign Conversion, Dynamic Gesture Segmentation, Fingerspelling, Natural Language Processing (NLP), Pose Estimation, Hand Landmark Tracking, Real-Time Sign Language Recognition, Data Augmentation, Accessibility Technology

I. Introduction

Indian Sign Language (ISL) plays a crucial role in facilitating communication for the hearing-impaired community across India. However, a significant challenge persists in bridging the communication gap between individuals who rely on ISL and those who use spoken languages. In India, the availability of certified interpreters is minimal, with only around 339 Level-C/DISLI certified interpreters available nationwide. This limitation restricts communication and access to services for the deaf community, affecting education, employment, healthcare, and daily interactions.

Given the advancements in artificial intelligence (AI) and natural language processing (NLP), there is a growing interest in developing automated systems that can translate between spoken languages and sign languages. While various solutions exist for languages like American Sign Language (ASL), very few focus on ISL, despite its unique structure and wide usage.

Our project aims to address this gap by developing a real-time bidirectional translation system that converts English text to ISL and ISL gestures to text. This system uses deep learning models for gesture recognition and natural language processing (NLP) techniques for sentence restructuring and translation. By offering a mobile-based, user-friendly application, we provide a tool that can assist in day-to-day communication without relying on human interpreters.

The proposed system integrates cutting-edge technologies such as Long Short-Term Memory (LSTM) neural networks for gesture recognition, and Mediapipe for detecting key points in gestures. Through this innovation, we enable real-time communication between hearing and non-hearing individuals, ensuring smoother interactions and greater inclusivity. This paper discusses the development process, technical challenges, and the results obtained during testing.

II. Overview

Motivation For Work: The motivation behind this project stems from the communication barriers faced by the hearing-impaired community in India. Indian Sign Language (ISL) is the primary mode of communication for many deaf individuals, but there are only a limited number of trained interpreters available across the country. This shortage restricts the daily interactions of hearing-impaired individuals with the hearing population in crucial areas like education, healthcare, and employment.

Traditional solutions, such as relying on human interpreters or text-based messaging systems, are often slow, expensive, or impractical for real-time communication. In today's digital world, where real-time and seamless communication is a necessity, there is a strong demand for automated systems that can effectively translate between spoken and sign languages. By bridging this gap, we aim to provide the hearing-impaired with more independence in their day-to-day lives.

Furthermore, the use of artificial intelligence (AI), natural language processing (NLP), and deep learning has opened up

new possibilities for creating systems capable of recognizing gestures and translating them into text, and vice versa. Our project leverages these technologies to create an efficient, real-time bidirectional translation system between text and ISL.

This project is also motivated by the opportunity to bring this technology into mobile applications, making it accessible to a broader audience. With smartphones becoming increasingly ubiquitous, a mobile-based ISL translation tool can significantly impact communication for millions of people across the country.

Problem Statement: The lack of real-time communication tools for the hearing-impaired presents a significant barrier to social and professional interactions. Despite the existence of technologies for other sign languages, such as American Sign Language (ASL), Indian Sign Language (ISL) remains largely underrepresented in the development of automated translation systems.

Key problems that need to be addressed include:

- **Real-time Translation:** Current tools for ISL translation are either non-existent or not equipped to handle real-time communication. This limitation hinders spontaneous and effective conversations between hearing-impaired individuals and others.
- **Bidirectional Translation:** There is no widely available tool that can effectively translate both text to ISL and ISL to text. Many available tools focus on one-way translation, limiting their usefulness.
- **Accurate Gesture Recognition:** ISL involves not only hand gestures but also facial expressions and body postures, which are essential for conveying meaning. Existing gesture recognition systems often struggle to accurately interpret these complex movements, especially in varied environments.
- **Grammar Differences:** ISL follows a different grammatical structure from English, making direct translation impossible. A system that translates text to ISL must account for these structural differences to produce grammatically correct and meaningful translations.

The proposed system aims to address these problems by using deep learning and NLP techniques to create an accurate, real-time bidirectional ISL translation tool. By developing this tool, we hope to create a solution that is not only technically feasible but also accessible to the broader hearing-impaired community in India.

III. Literature Review

The concept of translating between sign language and spoken language has been explored through various approaches, with significant advances made in recent years due to the rise of AI and deep learning techniques. Previous research has focused on both American Sign Language (ASL) and Indian Sign Language (ISL), leveraging different machine learning and image processing models to interpret gestures and translate them into textual or spoken languages.

In order to give the model a more realistic and vibrant appeal, Gupta and Sinha (2018) attempted to create a system that effectively transforms the entire input sentence into a single visual instead of representing several words through a GIF or photo. Since the ISL lexicon is still limited and will eventually need to expand, much more work can be done on this track. [1].

In order to create character animation, YIN Qinran and CAO Weiqun suggested a technique for projecting human motion data from a video onto a two-dimensional character. We propose a geometric calibration method based on the tree structure to correct motion reorientation of the bones, obtain a good skeleton-driven deformation effect, and generate high quality animation in the same posture. We statistically analyze common two-dimensional human body movements and classify the basic posture of the human body. We design and implement the method of human body posture recognition based on the skeleton information on images and videos. Auto-produced animation that requires less user participation can be created using this technique. [2].

In order to facilitate clear communication and self-expression, Ms. Satvika Reddy Satti and Mrs. P. Pavithara used technologies such as the Webkit Speech Recognition API for input, the Natural Language Processing Toolkit for text processing, and Blender 3D to create sign language animations for converting text or audio into sign language [3].

By creating SIGNPOSE, a 3D pose estimation model for continuous sign animations based on deep learning, Krishna and Vignesh (2020) investigated gesture recognition in sign languages. Their system's inability to support ISL limited its usefulness in translating spoken language to ASL. [4]

Using lemmatization and root-word extraction approaches to align sentences with ISL grammar, Kumar and Roy (2020) concentrated on NLP integration for translating English text to ISL. [5].

Using computer vision machine learning techniques and a 3D creation suite, Lan Thao Nguyen and Aeneas Stankowski created a technique to produce SL from 2D video recordings. By constructing bones between pairs of designated locations, a skeleton was constructed from the identified body markers. This was then applied to a rigged avatar after being exported as BVH data. The BVH skeleton's corresponding bone rotations were replicated in the avatar's bone constraints. [6].

Using smart glove flex sensors, a Raspberry Pi, a microcontroller, Python, and C, Alumona T. L., Okorogu V.N., and Nnabude E. F. created a sign language recognition system that can recognize and understand hand motions and

translate them into speech. The system's goals of developing a flex sensor system, using a microcontroller, building a signal database, adding a text-to-voice/audio subsystem, and integrating all parts to guarantee complete operation were all accomplished. [7].

In their work, Abey Abraham and Rohini V. suggested that back-propagation neural networks be used to forecast the demands of mute persons by associating their varied needs with the values of flex sensors. By modifying their weights, the neural network model produces an accurate prediction. But only American Sign Language was compatible with the equipment. [8].

Limitations:

- **Focus on ASL:** Many existing systems are primarily designed for ASL, making adaptation to ISL challenging due to structural and grammatical differences.
- **Gesture Recognition Accuracy:** Many systems experience low accuracy rates in real-time scenarios, particularly under varied lighting conditions or with different signers.
- **Complex Sentence Structures:** Most tools struggle to effectively translate complex or idiomatic expressions, often resulting in oversimplified translations that can lead to miscommunication.
- **Limited Vocabulary:** Many systems do not encompass the full range of vocabulary and nuances present in ISL, limiting their practical applications in everyday communication.

These limitations highlight the need for continued research and development to address the challenges and enhance the robustness, reliability, and accessibility of IoT based accident detection systems.

IV. Methodology

Our project relies on a combination of natural language processing (NLP) for the text-to-ISL conversion and deep learning for gesture recognition in the ISL-to-text conversion. The theoretical foundation of the project can be broken down into several key components:

A. *Text-to-ISL Conversion:*

Parsing and POS Tagging: English sentences are parsed using an NLP pipeline that assigns part-of speech (POS) tags to each word. The POS tags help identify the grammatical structure of the sentence, which is then reordered to match ISL's Subject-Object Verb (SOV) structure. For example, in English, "She ate an apple" becomes "She apple ate" in ISL.

Lemmatization: The process of reducing words to their base or root form is critical in ISL, as it tends to use simpler forms of words. For example, "running" is reduced to "run" to match the vocabulary used in ISL.

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WPS	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	` or "
POS	Possessive ending	*'s*	"	Right quote	' or "
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	[, (, {, <
PRPS	Possessive pronoun	*your, one's*	)	Right parenthesis	],), }, >
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	. ! ?
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	: ; ... – -
RP	Particle	*up, off*			

Fig. 1. List of Parts of speech in english language

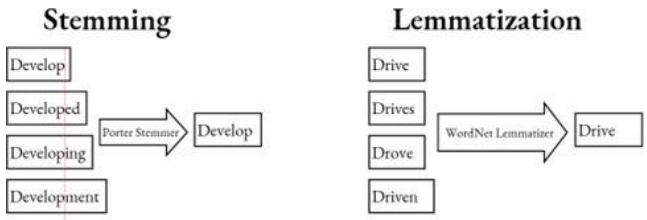

Fig. 2. Stemming and Lemmatization

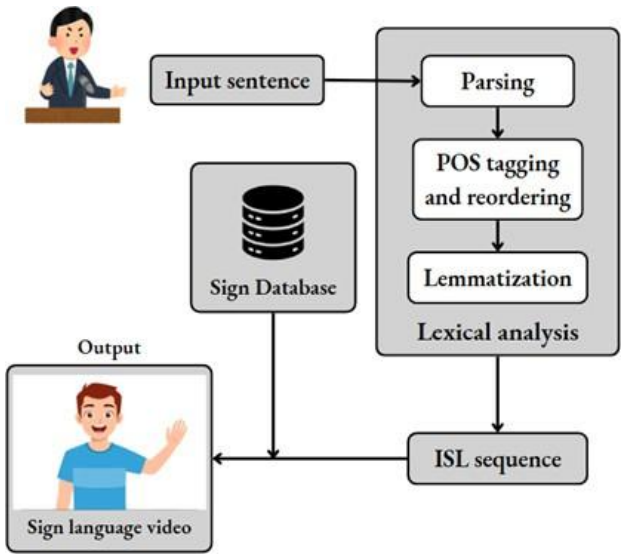

Fig. 3. Text to sign language conversion

Stop-Word Removal: ISL typically omits function words such as ”a,” ”the,” and ”of,” which do not carry substantial meaning. The system removes these words before mapping the remaining words to their corresponding ISL signs.

Sign Mapping and Gesture Formation: The key challenge is converting text into gestures. The system uses a pre-existing database of ISL signs to match each word in the sentence to its corresponding sign. If a word is not available in the database, the system falls back on finger-spelling, where each letter of the word is signed individually.

B. ISL-to-Text Conversion

Gesture Recognition: The core of ISL-to-text conversion involves recognizing ISL gestures from video inputs. Using the Mediapipe library, the system extracts keypoints from the hands, face, and body. These keypoints are critical for accurately identifying each sign.

LSTM Neural Networks: Long Short-Term Memory (LSTM) networks are used to process the time-series data of gestures. LSTMs excel at recognizing sequences, making them ideal for interpreting gestures that involve multiple frames. The input sequence consists of keypoints detected by Mediapipe, and the LSTM processes this data to predict the most likely gesture being performed.

Text Generation: After the LSTM identifies the sign, the system maps it to the corresponding English word or phrase. The output is then structured into grammatically correct English sentences based on SVO order.

C. Handling Ambiguities:

ISL, like all languages, contains signs that may have multiple meanings depending on context. To resolve ambiguities, the system incorporates contextual analysis based on the surrounding words, similar to how humans disambiguate meanings in communication.

D. Software Architecture:

The system architecture integrates frontend and backend components, data pipelines, and databases. Below is a detailed breakdown of the software architecture:

- **Frontend**: The frontend is built using NextJs , which provides cross-platform compatibility, making the app available on both Android and iOS devices. The user interface allows users to input text for conversion into ISL and to record ISL gestures for translation into text. The interface is designed for simplicity, ensuring ease of use for individuals with varying levels of technological proficiency.

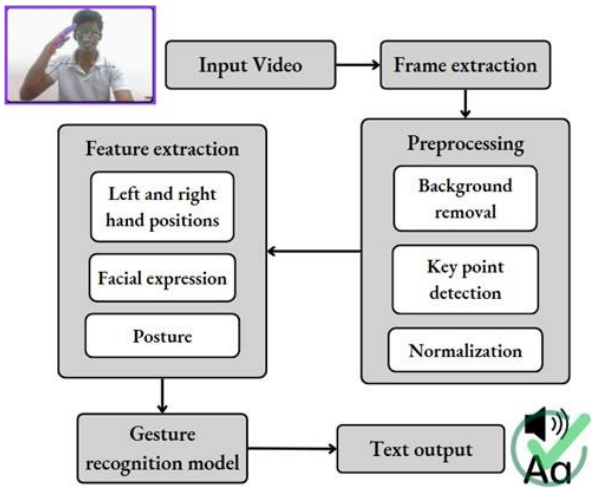

Fig. 4. Sign-to-text conversion

- **Backend**: The backend is implemented in Flask, a lightweight Python web framework, which handles the logic for text processing and gesture recognition.
  - **Text-to-ISL Module:** This module is responsible for processing the text input from the user, applying NLP techniques such as parsing, lemmatization, and stop-word removal, and sending the processed text to the sign mapping module.
  - **Text-to-ISL Module:** This module is responsible for processing the text input from the user, applying NLP techniques such as parsing, lemmatization, and stop-word removal, and sending the processed text to the sign mapping module.
- **Gesture Recognition Pipeline:**
  - The video input is processed to extract frames at regular intervals. Each frame is analyzed using Mediapipe, which identifies keypoints on the hands, face, and body.
  - A sequence of 40 frames is then fed into the LSTM neural network, which predicts the action or sign based on the temporal dependencies in the data. The LSTM output is then classified into a specific ISL gesture.
- **Database:** MongoDB and Firebase are used for storing both the sign language videos and metadata associated with gestures. MongoDB is used for storing larger datasets, including videos and images, while Firebase handles real-time data synchronization for user input and output.
- **Real-Time Processing:** The system ensures real-time performance by minimizing latency in both text processing and gesture recognition. This is achieved by optimizing the LSTM model for fast inference and using efficient data handling techniques in the backend.

- **Security and Authentication:** User authentication is handled through Clerk and OAuth services, ensuring that users can securely access their data and personalize their experience. This is particularly important for ensuring privacy in personal communications.

V. Implementation

A. *Text-to-ISL Conversion Module*

The purpose of this module is to translate English statements into ISL videos. The steps that make up the procedure are as follows:

- Preprocessing Sentences: Entering Tokenization and grammatical reordering of English sentences are done to make them fit the Subject-Object-Verb structure used in ISL.
- Section of Speech Labelling: A natural language processing (NLP) library is used to tag each word in the phrase, making word-level processing and reordering easier.
- Lemmatization and stemming: To precisely map words to their corresponding ISL signs, words are reduced to their most basic forms.
- ISL Video Mapping: Every word in the processed sentence is mapped to a predetermined dictionary of 150 ISL videos. When a term is not found in a dictionary, it is created by fingerspelling it using the signals of its individual letters.
- Video Stitching: To provide a cohesive video representation of the input sentence, separate ISL videos are stitched together sequentially.

B. *ISL Gesture Recognition Module*

This model is used to covert the ISL gestures to English language.

- Gesture Tracking: To extract body landmarks (position, left hand, right hand) from video frames, Mediapipe Holistic is used. To guarantee invariance to distance from the camera, a total of 258 characteristics are retrieved from each frame and normalized depending on torso measurements.
- Feature Encoding: The retrieved landmarks are molded into a 3D input format for the LSTM model and arranged into 40-frame sequences.
- Gesture Classification: To categorize sequences into one of the gesture classes, an LSTM-based model is developed. For every class, probabilities are produced.

VI. Results and Analysis

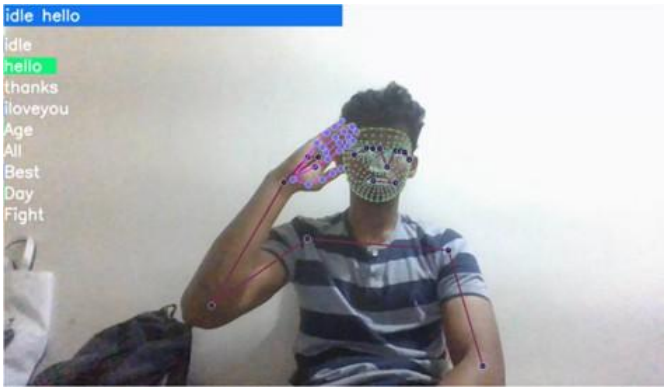

Fig. 5. Live feed output for ISL to text model

A. *Text to ISL conversion module:*

The Text-to-ISL module successfully translates English text into ISL videos using a predefined dictionary of 150 ISL signs. For words that do not have a direct ISL equivalent, the system employs fingerspelling to represent them using individual letter signs. This approach ensures that the translation remains as accurate and comprehensive as possible, even when encountering unfamiliar terms.

B. *ISL to Text conversion module:*

Developing the ISL-to-text module posed significant challenges, primarily due to the need for a large and diverse dataset, as well as substantial computational resources for model training.

Initially we started off with a very basic model that could only classify between 3 words. We then gradually increased the number of words. As the vocabulary expanded, the dataset had to be scaled accordingly to maintain classification accuracy. The initial model of 3 signs needed 30 samples for each sign.

As we increased the number of words to 5, we realised that the inclusion of facial landmarks negatively impacted model performance. Since the initial vocabulary consisted of signs that did not rely on facial expressions, this additional data introduced unnecessary complexity. To optimize performance, facial landmarks were temporarily excluded, with plans to incorporate them in future iterations as the vocabulary expands to include expression-dependent signs.

Currently, under controlled circumstances the model built can classify between 10 words with 100% Training and testing accuracy. In real-time testing, the model correctly classifies the words nearly 98% of the times. nevertheless, it has slight problems with overlapping or ambiguous gestures.

C. *Error handling:*

Fallback techniques such as fingerspelling for unfamiliar words and ongoing gesture recognition optimization under various circumstances are used to manage errors.

VII. CONCLUSION

This paper presented a dual-module system for translating English text to ISL videos and ISL gestures to text using Mediapipe Holistic and LSTM networks. The proposed system effectively bridges the communication gap for the Indian deaf community by offering a scalable, real-time translation solution.

In future work, we aim to:

- Expand the vocabulary to over 100 ISL gestures.
- Integrate 3D hand tracking for finer sign language recognition.
- Deploy the system as a mobile application for accessibility.

By addressing these challenges, our system can further improve inclusivity and real-world usability for ISL users.

VIII. FUTURE SCOPE

- Inclusion of facial expressions, as the vocabulary of the model increases.
- Create a mobile application for this web application to increase utility. Developing a mobile app for the sign language recognition system will make it more accessible. Users can perform real-time sign language translation on the go, receive voice feedback, and use it offline, making the system more practical for daily communication.
- Add more regional Indian languages. To make the system more inclusive, integrate support for regional Indian sign languages like ISL, MSL, TSL, etc. This would allow users from different regions to communicate effectively, improving the system's reach and accessibility for diverse communities.

IX. REFERENCES

[1] Vardan Gupta, Saumya Sinha, Puneet Bhushan, Minal Shettigar, "English Text to Indian Sign Language Translator" *Artificial Intelligence Journal*, 2023.

[2] Qinran, Y. and Weiqun, C. (2021), "Video-Driven 2D Character Animation" *Chinese J. Electron., 30: 1038-1048.*, 2021.

[3] Ms. Satvika Reddy Satti, Mrs.P.Pavithra, "AUDIO OR TEXT TO SIGN LANGUAGE CONVERTER" *Journal of Emerging Technologies and Innovative Research (JETIR)*

[4] Shyam Krishna*, Vijay Vignesh P*, Dinesh Babu J, "SignPose: Sign Language Animation Through 3D Pose Lifting" *International Journal of Automation and Computing*

[5] Das Chakladar, Debashis, Pradeep Kumar, Shubham Mandal, Partha Pratim Roy, Masakazu Iwamura, and Byung-Gyu Kim. 2021. "3D Avatar Approach for Continuous Sign Movement Using Speech/Text" Applied Sciences 11, no. 8: 3439.

[6] Lan Thao Nguyen, Aeneas Stankowski, and Eleftherios Avramidis, "Automatic generation of a 3D sign language avatar on AR glasses given 2D videos of human signers" *Proceedings of the 18th Biennial Machine Translation Summit, Virtual USA, August 16 - 20, 2021*.

[7] Alumona T. L., Okorogu V.N, Nnabude E. F, "Sign Language Recognition System using Flex Sensor Network" *n International Journal of Research and Innovation in Applied Science · October 2023*.

[8] Abey Abraham , Rohini V, "Real time conversion of sign language to speech and prediction of gestures using Artificial Neural Network" *8th International Conference on Advances in Computing and Communication (ICACC-2018).*