iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 13 **Issue:** V **Month of publication:** May 2025

**DOI:** https://doi.org/10.22214/ijraset.2025.71548

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 13 Issue V May 2025- Available at www.ijraset.com*

# Multilevel Conversion of Indian Sign Language from Gesture to Speech

Mr. Shivanshu Shahi[1], Manjeet Kumar[2], SurabhiVerma[3], Talib Usmani[4], Ganesh Ji Patel[5]
*Buddha Institute of Technology, Gida, Gorakhpur*

***Abstract:*** ***Indian Sign Language (ISL) serves as a primary mode of communication for Deaf and hard-of-hearing communities in India. However, despite its societal importance, ISL remains largely unsupported by mainstream technological platforms, limiting inclusive communication. This research introduces a real-time ISL recognition andtranslationsystem thatconvert shandgesturesintocorresponding text and speech outputs, enabling phrase-level communication rather thanisolated characterinterpretation. The architecture uses a modular pipeline approach, with a Convolutional Neural Network (CNN) for accurate gesture classification, a phrase-mapping module to translate gestures into meaningful expressions, a MediaPipe for accurate hand landmark detection, and a text-to-speech (TTS) system to turn the generated text into audible speech output. Unlike previous systems restricted to static signs, our approach supports semantically rich, multi-word phrases, enhancingnatural communicationflow. Aspecially constructed dataset of ten frequently used Indian Sign Language (ISL) phrases was used to train the model. To improve generalization,150samplesfromeachclasswereretakeninvarious lighting and background conditions. The final system achieved 95% classification accuracy, operated at 60 frames per second, and maintainedlatency below100milliseconds. Usabilitytestingwith multiple users confirmed the system's robustness, responsiveness, and accessibility. The findings demonstrate the viability of deploying deep learning-based ISL recognition systems in authentic environments, including public areas, healthcare facilities, and educational institutions.***

## I. INTRODUCTION

Communication is fundamental to social, educational, and professional inclusion. However, individuals in the Deaf and hard-of-hearing communities face persistent challenges in day-to-day interactions, particularlyinasociety wherespoken languagedominates.IndianSign Language (ISL) plays a crucial role in bridging this gap, yet technological support for ISL remains limited. Most existing ISLrecognition systems are constrained to recognizing static alphabets or individual words. These systems often require manual transitions and do not support the fluid, continuous expression characteristic of natural language. Furthermore, they lack real-time responsiveness, making them unsuitable for spontaneous interactions in practical settings.

Thisresearch aimstodevelopareal-timeISLrecognitionsystem that translates dynamic hand gestures into complete phrases and delivers outputs in both text and speech formats. The system integrates computer vision, deep learning, and human-computer interaction techniques to enhance inclusivity and accessibility.

## II. RELATED WORK

Several researchers have contributed to ISL recognition using computer vision and deep learning techniques. Table 1 summarizes key studies:

| Author | Year | Dataset | ModelUsed | Accuracy | Limitations |
|---|---|---|---|---|---|
| Chaudhary& Dey | 2023 | ISL alphabets(A–Z) | CNN | 94% | Onlystaticsigns |
| Rani&Kaur | 2022 | 15ISLsigns | RandomForest | 89% | Not real-time |
| Patil&Kale | 2022 | 25dynamicsigns | Deep CNN | 91% | Highresourcerequirement |
| Srivastava etal. | 2024 | Continuousgesturedataset | MediaPipe + LSTM | 92.3% | LacksTTS,complexsetup |

| | | | | | |
|---|---|---|---|---|---|
| Thiswork(proposed) | 2025 | 10 phrases,150 samples | MediaPipe + CNN+TTS | 95% | Limiteddataset ,staticphrases |

While previous work demonstrated progress in gesture recognition, most models focus on isolated characters or constrained datasets. Our proposed system expands these capabilities by supporting phrase-level interpretation and speech output, offering a more natural user experience.

## III. PROBLEMSTATEMENT

While significant progress has been made in isolated sign recognition, relatively fewer studies have addressed fluid, phrase-leveltranslation.Forinstance,Sharmaetal.(2021)used SVM classifiers for digit recognition but lacked real-time deployment. Similarly, Bose and Narayan (2020) explored dynamic gesture tracking but excluded audio output. These gaps affirm the need for an integrated solution that combines gesture recognition with natural language understanding and speech synthesis.

Current ISL recognition systems fall short in delivering accessible, phrase-level communication. Most are constrained to static gestures and require user intervention for sentence construction.Furthermore,they lack auditory output andoften perform poorly in dynamic or real-time settings.

This project seeks to address these limitations through a comprehensive, phrase-level ISL gesture recognition system that translates sign inputs into coherent sentences and corresponding audio.

## IV. RESEARCH GAP

Despite advances in gesture recognition, several challenges remain unaddressed:

*1)* StaticRecognition:Limitedtoalphabetsordigits.
*2)* NoSpeechOutput:Text-onlysystemsreduce communication effectiveness.
*3)* LatencyIssues:Inadequateperformanceforreal-time interaction.
*4)* LimitedScalability:Difficultyaddingnewphrasesor signers.

Our system bridges this gap by enabling end-to-end, real-time phrase recognition with audio output using lightweight architectures suitable for broader deployment.

Before training, the raw hand landmark data was normalized and standardized to ensure consistency across users. Gesture sequences were manually annotated, a time-intensive process that involved expert signers validating each phrase. To ensure effective model evaluation and generalization, the dataset was divided into subsets to be used for training (80%), validation (10%), and testing (10%). Early stopping and learning rate decay were applied to optimize training performance and prevent overfitting.

## V. METHODOLOGY AND SYSTEM ARCHITECTURE

Themethodologyadoptedforthissystemisdesignedtoensure seamless, real-time conversion of Indian Sign Language (ISL) gesturesintomeaningfultextandspeechoutputs.Thecomplete flow of the system is illustrated in **Figure 1.1**, which outlines each stage from input acquisition to final output delivery.The process begins with the activation of a webcam that captures continuous real-time video of the user's hand gestures. These frames are passed through a hand detection and tracking module, which uses MediaPipe to extract 21 key hand landmarks for gesture analysis.

Following detection, the system checks whether the captured sign corresponds to any gesture in the trained dataset. If the gesture is successfully recognized, it is translated into a predefined natural language phrase, which is then passed to a text-to-speech (TTS) engine for auditory conversion. The speech output is delivered in the user's preferred language, completing the recognition loop. If the gesture is not recognized, the system provides immediate feedback and prompts the user to repeat the sign, ensuring interaction accuracy and user engagement. This structured flow supports both efficiency and accessibility, making the system practical for real-world deployment.

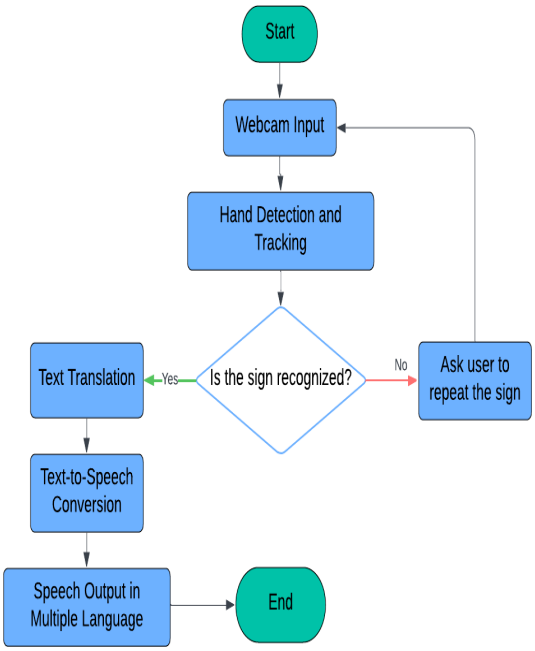


Figure1.1Flowchart

Thesystemcomprisesfivemajorcomponents:

*1) Input Capture*

Video input is acquired using a webcam and preprocessed to normalize frame resolution and lighting.

*2) Handl And Mark Detection*

Inrealtime,MediaPiperecords21keypointsfromthehand, each of which represents the x, y, and z coordinates of a particular joint.The gesture classification process then uses these landmarks as feature inputs.

*3) Gesture Classification*

A CNNistrainedon1,500 gesturesamples(150perphrase class). The network includes:

- 3convolutionallayers(32,64,128filters;kernelsize=3×3)
- Maxpoolinganddropoutlayerstopreventoverfitting
- 2denselayers(256,64units)withReLUactivation
- Softmaxoutputlayerfor10classpredictions

*4) Phrase Mapping and TTS*

Predicted classes are mapped to full phrases, which are then synthesized into speech using a TTS engine (e.g., pyttsx3). Multilingual support is implemented.

*5) User Interface*

UI,developedusingpython,displayspredictedphrasesand provides TTS playback in real time.

Figure1.2UIofPhrasePrediction

## VI. SYSTEM ARCHITECTURE

The architecture of the proposed system follows a sequential pipeline that enables real-time recognition and conversion of Indian Sign Language gestures into text and speech. The process begins with continuous image capture through a camera, followed by hand signal segmentation to isolate relevantgestureregionsfromthebackground.Thesesegmented inputs are then processed using a hand detection and tracking module, which identifies and follows key hand movements usinglandmark-basedanalysis.Oncedetected,thegesturesare classified using a deep learning model, and the recognized outputismappedtoacorrespondingphrase.Finally,thesystem deliverstheresultas textualfeedbackontheuserinterfaceand asspokenoutputviaatext-to-speechengine.Thismodularflow ensures high accuracy, real-time performance, and usability across various environments.

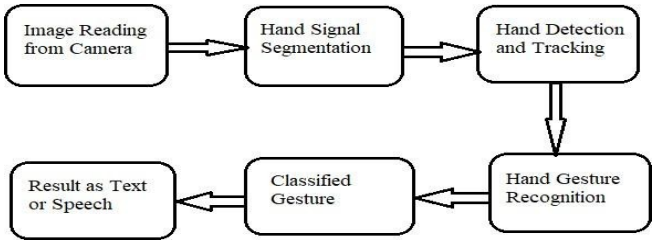



Figure1.3SystemArchitecture

## VII. DATASET AND MODEL TRAINING

10IndianSignLanguage(ISL)phrasesthatarecommonlyused in daily conversations make up the dataset.To improve model generalization,150 labelled samplestakenindifferentlighting and background conditions are used to represent each phrase. Data augmentation methods, such as rotation, scaling, and horizontal flipping, were used to increase robustness even more. TheAdam optimization algorithm was used to train the model over 50 epochs with a batch size of 32.

Raw hand landmark coordinates were standardized and normalized before training to ensure uniformity among users. To guarantee the correctness of every phrase, the gesture sequencesweremeticulouslyannotatedbyprofessionalsigners. The dataset was separated into three sets: testing (10%), validation (10%), and training (80%). Additionally, early stopping and learning rate decay strategies were employed to optimize performance and reduce the risk of overfitting.

## VIII. RESULTS AND EVALUATION

*1) Evaluation Metrics*

- Accuracy:Overallmodel correctness
- FrameRate(FPS):Processingspeed
- Latency:Responsedelay
- User Feedback: Informaluser testing

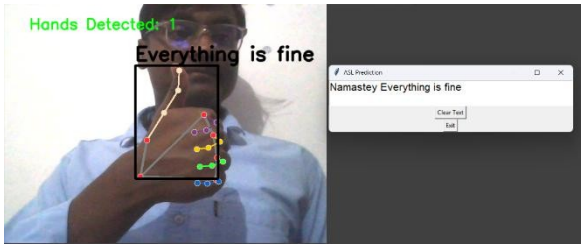
*2) Experimental Results*

- Accuracy:95%acrossallphrases
- FPS:60onmid-rangehardware
- Latency:<100msperframe
- Misclassifications:Observedunderpoorlightingor occluded gestures

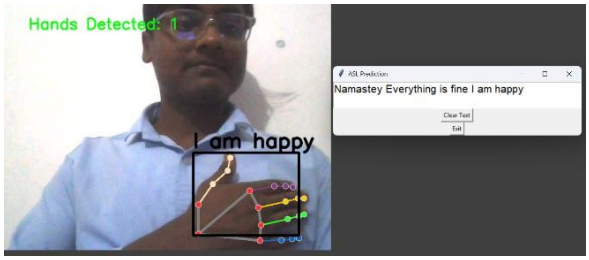
*3) Illustrative Outputs Include*

- “Namastey”

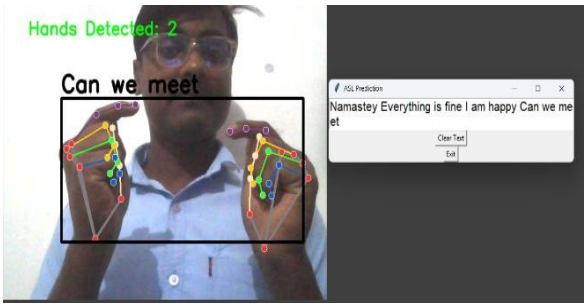
- "Everythingisfine"




- "Iamhappy"




- "Canwemeet?"




## IX. DISCUSSION

With speech output functionality seamlessly integrated, the system demonstrates the ability to recognize phrase-level IndianSignLanguage(ISL)gesturesinreal-time.Unlikeprior systemsthatarerestrictedtostaticalphabetdetectionoroffline usage, our implementation enables continuous, real-time interaction with both visual and auditory outputs.

ThesystemwasimplementedinPythonusingTensorFlowand MediaPipe for gesture processing. The text-to-speech output wasgeneratedusingthepyttsx3engineforofflinefunctionality, making the system usable even in low-connectivity areas. The entire pipeline runs efficiently on a machine with 8GB RAM andnoexternalGPU,confirmingitssuitabilityfordeployment in budget-constrained institutions. In addition, latency tests showed that the total time from gesture detection to audio output remained consistently under 100ms, even with background processes running.

## X. LIMITATIONS

While the results are promising, the system has several limitations:

*1)* Small dataset: Only 10 phrases are currently supported.
*2)* Static phrases only: No support yet for dynamic gesture sequences.
*3)* Limitedsigners:Datasetcollectedfromasmallgroup of individuals.
*4)* No grammar modelling: The system doesn't yet support grammatical construction or contextual understanding.

Addressing these limitations is key to improving generalizability and real-world deployment.

## XI. SECURITY AND PRIVACY CONSIDERATIONS

Given that the system processes real-time video input, privacy andsecurityarecriticalconcerns.Thefollowingmeasureshave been adopted:

*1)* Local Processing: All data remains on the local device; nothing is sent to the cloud.
*2)* No Persistent Storage: The system does not store or log gesture inputs or video frames.
*3)* Transparency: The open-source architecture allows for public inspection and compliance with ethical standards.

Future Enhancements:

- Encryptedlocallogs(ifuseroptsin)
- Securemulti-userprofiles
- Real-timeanomalymonitoringforgesturespoofing

## XII. FUTURE WORK

Theprojectopensupmultipleavenuesforfurtherresearchand enhancement:

*1)* Larger Dataset: Extend to 100+ gestures, including dynamic ISL signs.
*2)* TemporalModels:IncorporateLSTMorTransformer architectures for continuous sign sequences.
*3)* Multilingual Output: Expand TTS to support regional languages like Hindi, Bengali, and Tamil.
*4)* Mobile Integration: Deploy the system on Android/iOS platformsusinglightweightmodels(e.g.,TensorFlowLite).
*5)* Personalization:Integrateactivelearningmodulestoadapt to user-specific signing styles.

## XIII. CONCLUSION

Thisresearchpresentsareal-timeISLgesturerecognitionsystemthat translates signed phrases into text and speech. By leveraging MediaPipe for hand tracking, CNNs for classification, and TTS for output, the system provides an inclusive communication platform for Deaf individuals.

Achieving 95% accuracy and 60 FPS performance, the prototype demonstratesstrongpotentialfordeploymentinaccessibility-focused environments.Futureworkwillfocusonincreasinggesturediversity, improving modeladaptability, and expanding deployment platforms.

TEAM CONTRIBUTIONS

The table below illustrates the distribution of tasks among the team members:

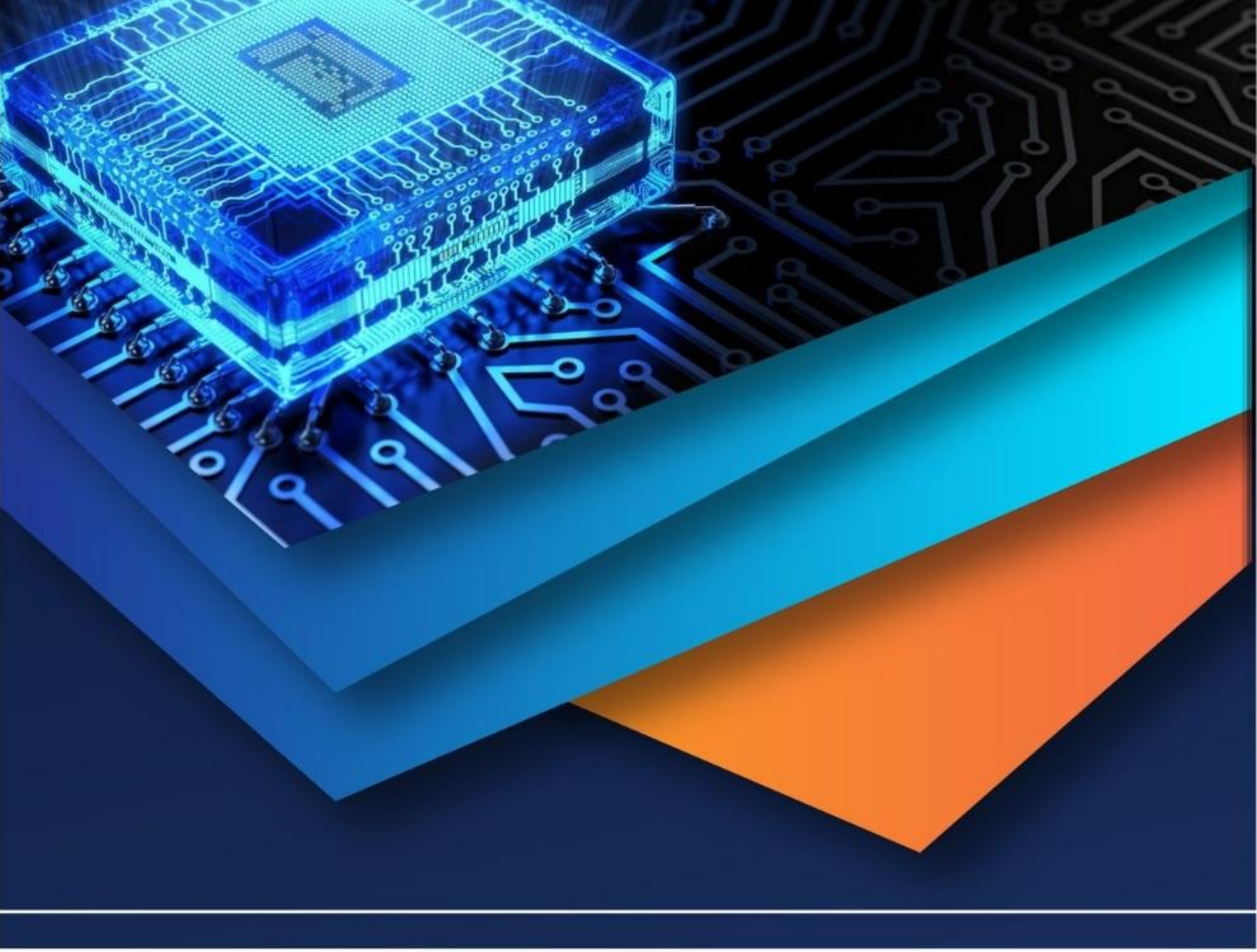
| TEAM MEMBERS | CONTRIBUTION |
|---|---|
| MANJEETKUMAR | CNNtraining,datasetcollection, UI integration |
| SURBHIVERMA | MediaPipeintegration,frontend development |
| TALIBUSMANI | Text-to-Speechmodule,interface testing, evaluation metrics |
| GANESHJI PATEL | Phrasemappinglogic,bugfixing, system documentation |

Witheffectivecollaboration,communication,anddedication,theteam was able to successfully implement allplanned featuresand deliver a fully functional Mint-Verse platform.

## REFERENCES

[1] Chaudhary, R., & Dey, A. (2023). “Real-Time Indian Sign Language Recognition using CNN,” IEEE Access.
[2] Rani, N., & Kaur, H. (2022). “Enhanced ISL Recognition using Machine Learning Algorithms,” International Journal of ComputerApplications.
[3] Patil, S., & Kale, P. (2022). “Real-time ISL Recognition Using Deep Learning Models,” IJIRCCE.
[4] Srivastava, S. et al. (2024). “Continuous Sign Language Recognition Using MediaPipe Holistic,” Wireless Personal Communications.

[5] OpenAIAPIDocumentation–https://platform.openai.com/docs
[6] MediaPipe Hands–https://google.github.io/mediapipe/solutions/hands
[7] Pyttsx3Documentation –https://pyttsx3.readthedocs.io



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY