# Neural 3D Avatar-Based Bidirectional Communication System For Deaf And Hearing Users

**R.Sowmiya[1], M.Saniya[2], B.Shafeek Ahamed[3], J.Nikile Eines Dhoni[4], S.Mohamed Shath[5]**
[1]Assist prof, Dept of Artificial Intelligence and Data Science
[2, 3, 4, 5] Dept of Artificial Intelligence and Data Science
[1, 2, 3, 4, 5] Mohamed Sathak Engineering College, Kilakarai, Tamil Nadu, India

***Abstract-*** *Deaf and mutes use sign language to communicate, yet communication between signers and non-signers is challenging as there is a lack of awareness. This lays obstacles in such spheres as education, medical care, and government services. The current systems have shortcomings such as limited vocabulary, inadequate real time operation, support of the Indian Sign Language and no. effective two-way communication. This project sug gests an AI-based communication to deal with these challenges. system that facilitates a smooth interaction between sign-language and nonsigners. The model is based on Transformer to identify continuous. facial expressions, body movements, and gestures of live video input, providing proper and real-time interpretation of the sign language. It also incorporates speech recognition and text to speech modules in order to translate. translates voice text and voice even in a noisy setting. Another component of a neural avatar is a neural sign animation to. visually represent communication. The system is adopted as web-based application, which is accessible. and ease of use. Overall, the proposed system enhances inclusive communication and breaks the barrier between deaf and the hearing communities.*

***Keywords:*** AI, Sign Language Recognition, ISL, Trans former, Real-Time System, Gesture Recognition, TTS, Speech Recognition.

## I. INTRODUCTION

The problem of communication among deaf and mute people and non-signers has been one of the greatest because people are not aware of the sign language. This is a communication gap that impacts key areas like education, healthcare, work places and public services. Conventional means such as interpreters or written communication are frequently ineffective, expensive and not applicable in real-time communication.

The current progress in the Artificial Intelligence (AI) and Deep Learning has brought about the possibility to create smart communication systems. Nevertheless, the current systems continue to be limited because of a limited number of vocabulary, inability to operate in real time, poor support of Indian Sign Language (ISL), and the lack of a bidirectional communication.

To mitigate these challenges, this project will present an AI-powered two-directional communication system that will facilitate smooth communication between deaf and hearing people. The system is powered by a Transformer-based model to identify in real-time video input continuous gestures, facial expressions, and body movements. It also incorporates speech recognition and text-to-speech modules to convert voice-totext and audio. There is also a Neural Avatar that creates sign language animations, enhancing the visual communication. The implementation of the system is a web-based application as it is more accessible and usable.

## II. LITERATURE REVIEW

### *A. Existing Sign Language Recognition Systems*

Sign language recognition is a well-known and used assistive technology in recent years in order to overcome the communication barrier between the deaf and hearing people. Different methods grounded on computer vision, machine learning, deep learning, and sensor-based systems have been suggested to enhance recognition accuracy and real-time performance.

Malay Kumar et al. (2025) suggested a better translation system between American Sign Language (ASL) and Indian Sign Language (ISL) through hybrid deep learning models and Large Language Models (LLMs). The system combines the Convolutional neural networks (CNN) and random forest classifier (RFC) to recognize gestures and then uses the LLM to refine the text and generate gestures with RIFE-Net. The framework obtained an accuracy of 93.0

Ozcan et al. (2024) proposed a Zero-Shot Sign Language Recognition (ZSSLR) system that fuses hand and body landmarks and embeds text. This model makes use of selfattention to enhance the representation of features and is able to recognize unseen gestures without labeled data. The

approach minimizes reliance on datasets, but, nonetheless, its performance is very much dependent on the quality of extracted features and textual descriptions.

The study by Zhang et al. (2023) came up with a deep learning-based sign language detection model, which is aimed at differentiating between the correct and incorrect gestures. The system utilizes spatial-temporal feature extraction, optical flow-based hand detection, and attention in an encoder-decoder architecture. Training was done on a new dataset (SLCD). Although the model is better than other models in terms of accuracy in education, it can demand considerable computational resources and does not make real-time systems.

Luqman (2022) suggested a high-performance two-stream deep learning model which is Dynamic Motion Network (DMN), Accumulative Motion Network (AMN), and Sign Recognition Network (SRN). The model is based on accumulative motion representation to effectively represent spatial and temporal features and works well on various datasets.

Nonetheless, it is more specialized in recognizing individual signs, and cannot be applied to continuous gesture recognition. Bala and Shrenika (2021) implemented a sign language recognition system based on the image processing method and template matching. The system utilizes preprocessing, Gaussian filtering, edge detection and Sum of Absolute Difference (SAD) to match gestures. Although the system is basic and affordable, it uses pre-defined datasets and is not scalable and real-time efficient.

Talukder and Jahara (2020) suggested a real-time sign language detector based on the YOLO-based deep learning models. The system interprets gestures and translates them to text and speech and helps in generation of sentences. Its feature extraction methods include SIFT and PCA and it performs well in challenging environment. Nevertheless, it is constrained by the size of the set of data and has difficulties in recognizing continuous gestures.

Dhivyasri et al. (2021) introduced a method to interpret Indian Sign Language (ISL) with machine learning algorithms, e.g., CNN, RNN, and Support Vector Machine (SVM). SURF method is applied to extract features and K-means and Bag of Visual Words (BoV) are used to perform clustering. The system was also successful at recognizing gestures reliably, but is highly sensitive to the quality of data it is given and is not capable of sophisticated real-time.

Anupama and Usha (2021) created an automated sign language interpreter with data gloves with flex sensors. The system has a K-Nearest Neighbor (KNN) algorithm to classify gestures. Although it is good and easy to implement, it uses hardware and makes the user less comfortable, and it can be scaled.

Malli Mahesh Chandra and Rajkumar (2019) suggested a prototype based on a sensor-driven input and SVM classification by converting sign language gestures into speech. The system was highly accurate on a small range of gestures and is only limited to single hand gesture and needs a hardware support.

Kudrinko et al. (2021) conducted a review of wearable sensor-based sign language recognition systems in a comprehensive way. The study analyzed various sensor configurations, classification techniques, and performance metrics. Although wearable systems are highly accurate, they experience user comfort, standardization and availability of data challenges.

In general, the current systems have achieved a lot in terms of sign language recognition and translation. Nevertheless, the vast majority of methods have shortcomings including limited vocabulary, limited to specific datasets or hardware, no support of the Indian Sign Language, and no effective two-way realtime communication. These issues show that a sophisticated AI-oriented solution is necessary that provides the correct, timely, and convenient communication that is proposed in this work.

## III. RELATED WORK

New developments in sign language recognition have investigated various methods such as vision-based, sensor-based and hybrid system. The main difference between vision-based and sensor-based approaches is that the former are based on deep learning models that are used to retrieve spatial and temporal features of the input video, and the latter are based on wearable devices like data gloves that are more precise in gathering gesture data. Hybrid systems also seek to use a combination of both methods in order to enhance performance.

With these developments, there are still a number of challenges. Most of the systems that are in place can only do gesture recognition in isolated form and not in continuous sign language which limits their practical use. Moreover, models tend to rely on huge labeled datasets, thus, being less flexible to new or out-of-sight gestures. Another crucial aspect is realtime performance, since most of the approaches demand a

lot of computational resources, limiting their efficiency in the real-world.

Moreover, some of the systems are specific to particular sign languages like the ASL or local data, which makes the Indian Sign Language (ISL) less supported. Hardware-based solutions though correct, decrease usability because it requires external devices. Conversely, vision-based systems have difficulties with changing lighting conditions, background noise and dense environments.

The other significant constraint is that there is no efficient two-way communication. Majority of the systems only emphasize on translating the sign language into text or voice without offering a reverse process of translating a spoken language into sign language. This limits total communication between the deaf and the hearing people.

To address these shortcomings, the suggested system aims at continuous recognition of gestures in real-time, enhanced flexibility with the help of advanced AI models, and effective two-way communication. Accuracy and user interaction are further improved as the Transformer-based models are integrated with the neural avatar technology, which makes the system more practical and accessible.

## IV. PROPOSED METHODOLOGY

The suggested system will be an AI-based two-way system of communication allowing a smooth interaction between deaf and hearing people. The methodology incorporates sign language recognition, speech processing and avatar based sign synthesis to maintain real-time and correct communication. The total workflow includes three large modules, i.e., gesture recognition, speech-to-text conversion, and sign animation generation.

### *A. System Overview*

The system is used in two directions: (i) Sign language to text/speech conversion, and (ii) Text/speech to Sign language conversion.

During the first stage, the image is recorded by a camera and analyzed by deep learning models to identify gestures. The second stage involves converting spoken input into text and converting it into animated sign language with the help of a neural avatar. This bi-directional method guarantees total interaction among the users.

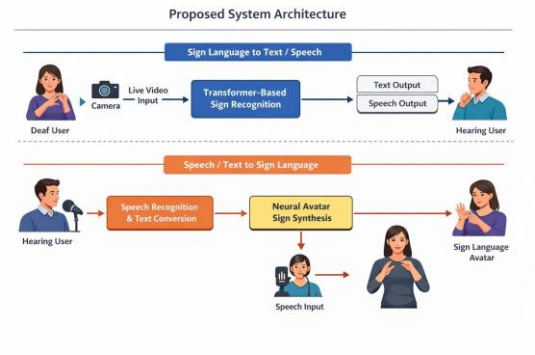


Fig. 1. Proposed system

### *B. Proposed System Architecture*

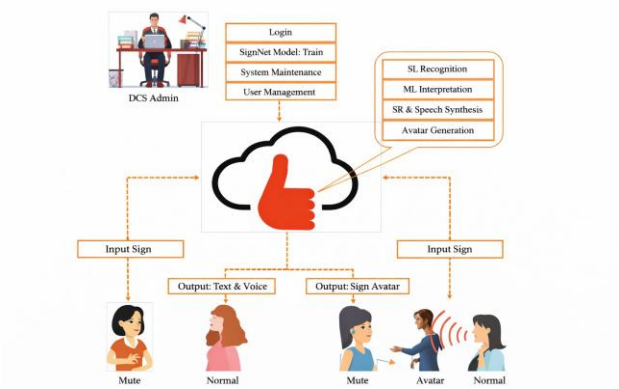


Fig. 2. Proposed system architecture

As illustrated in Fig. 2, the designed system comprises a number of modules such as sign recognition, speech processing and avatar generation.

### *C. Sign Recognition based on transformers.*

Sign recognition module takes live video input and processes continuous gestures such as hand movements, facial expressions and body posture. Sequential data is analyzed with the help of a Transformer-based model that involves a self-attention mechanism that assists in comprehending the temporal connection between gestures.

The Transformer is a successful approach to working with long-range dependencies in the continuous sign sequences, unlike the traditional ones, which enhances recognition accuracy. The output of the module is meaningful text that is relative to the identified Indian Sign Language (ISL) gestures.

### *D.Speech Recognition and Text Conversion*

The speech processing module is the one that converts the spoken language into the text written with the help of the latest deep learning methods. Models such as

Recurrent Neural Network Transducer (RNN-T), Connectionist Temporal Classification (CTC), and Deep Neural Networks (DNN) are used to process speech signals.

These models can process input sequences of varying length and robust to background noise, thus providing correct transcription. The text converted is an intermediate form to be processed further.

*E. Neural Synthesis of Avatars Signs*

The resulting text is sent to a Neural Avatar module that translates it into sign language animations. The avatar also generates realistic gestures such as hand shapes, movements and facial expressions in order to make the communication natural and expressive.

This module improves user comprehension, which is visually depicting the message in sign language which makes the system more interactive and accessible to deaf users.

## V. IMPLEMENTATION DETAILS

*A. Development Environment*

It uses Python to develop the system as a backend, and the Flask framework to support web application functionality. The frontend is created in HTML, CSS, JavaScript, and Bootstrap in order to provide an easy-to-use and responsive interface. The system stores and manages user data, gesture datasets, audio features and avatar mappings in MySQL, which is secure.

Gesture recognition, speech processing, and avatars are merged to facilitate smart processing with libraries like OpenCV, MediaPipe / OpenPose, TensorFlow / PyTorch, Librosa, and NumPy.

*B. Database Design*

A relational schema is used to design the database to effectively handle the system data. It contains user profile, role allocation (Admin, Deaf, Non-Deaf), gesture datasets, audio features, session logs, known text output and avatar animation mapping tables. Appropriate normalization and indexing is used to provide rapid data retrieval, data integrity and secure access between various modules.

*C. Preparation of the dataset and training the model*

Gesture sets of images and video streams are processed with image resizing, grayscale, noise reduction, binarization and background removal. The processed information is then trained on the SignNet model that uses convolutional layers to learn the spatial features and classify gestures correctly.

To process speech, audio signals are transformed into acoustic representation (spectrograms and Mel-frequency cepstral coefficients (MFCCs)). These characteristics are utilized to train models, which are founded on Recurrent Neural Networks (RNN), Connectionist Temporal Classification (CTC) and Deep Neural Networks (DNN), so that they can effectively recognize speech in noisy conditions.

*D. Model Integration and Deployment*

The trained models are incorporated into the web application that is developed using Flask. The input comes in the form of live video which is captured with the help of a camera and processed with the SignNet model and Transformer-based gesture encoder to recognize signs in real-time. Likewise, speech is recognized by the speech recognition module and the text is translated to the sign gestures via the Neural Avatar Synthesis engine. The system allows real-time two-way communication between deaf and hearing users.

*E. Front-End Development*

The front end interface is a responsive web application that has individual dashboards of various user roles. Deaf users are able to input using signs and see system reactions through texts or avatar animation. Other users, who are not deaf, can communicate via speech or text input which is translated to sign language forms. The interface is user-friendly, simple, and easy to use.

*F. Back-End Development*

The backend handles fundamental features like gesture detection, speech, avatars, session control, authentication of the user and database operations. The implementation of RESTful APIs is done to allow communication between the frontend interface, AI models, and the database.

*G. Module Integration*

All the system modules such as SignNet Model, Sign Recognition, Speech Recognition, Neural Avatar Synthesis and User Dashboard are integrated to operate as a single system. The system can accept the live gesture and speech inputs and provide the outputs in real-time, which means that the system is effective in two-way communication.

*H. Quality Assurance and Testing*

The system is measured by accuracy of gesture recognition, speech recognition accuracy, response time and quality of avatars animation. Functional testing, integration testing and stress testing are done in order to make sure that it is reliable. The system is also put to test with various user conditions to test its scalability and real time performance.

## VI. PERFORMANCE METRICS

The effectiveness of the suggested AI-based bidirectional communication system is measured in terms of gesture recognition rates, speech processing rate, real-time response and quality of avatars. These measures are used to make sure that both sign-to-speech and speech-to-sign communication would be effective.

*A. Sign Language Recognition Accuracy*

This is a measure of the accuracy with which the system recognizes hand gestures, facial expressions, and body motions based on live video input.

Accuracy= Total Predictions/ Correct Predictions

Improved accuracy implies improved performance of the Transformer-based gesture recognition model.

*B. Speech-to-Text Accuracy*

This metric is used to evaluate the accuracy of text translation of a speech model (RNN-T, CTC, DNN).

Accuracy= Total Spoken Words/ Correctly Transcribed Words

It guarantees a good communication between deaf and hearing users.

*C. Latency (Real-Time Response)*

Latency is a measure of time spent by the system to respond to input (sign or speech) and produce output.

Latency=Output - TimeInput Time

Real-time interaction and free flow of communication require low latency.

*D. Frame Processing Rate (FPS)*

FPS presents the number of video frames that are handled every second when recognising signs.

High FPS → motion tracking of gestures. Slow FPS (Framework) = slow output.

This measure indicates the real-time system capability.

*E. Avatar Generation Quality*

This would determine the fidelity of the neural avatar to sign language gestures, such as:

Hand movements Facial expressions Gesture clarity

The quality of improvement enhances clarity among deaf users.

*F. Noise Resistance (Speech Processing)*

This is a metric that determines how well the system is able to identify speech in a noisy environment correctly.

Increased robustness guarantees sound working under reallife conditions.

*G. End-to-End System Performance*

This checks the efficiency of the system in whole communication circles:

Sign → Text/Voice Speech → Sign Avatar

It indicates the level of effectiveness of the system in terms of the two-way communication.

## VII. RESULT AND CONCLUSION

The suggested AI-driven bi-directional communication system was tested to assess its efficiency in facilitating a smooth communication between deaf and hearing people. Transformer-based sign recognition module was found to be highly accurate in the recognition of continuous gestures such as movements of the hands, facial expression, and body movements of a live video. The model also managed to capture temporal dependencies well leading to better recognition performance.

The RNN-T, CTC and DNN models-based speech recognition module was able to turn the spoken language into a text with a satisfactory level of accuracy, even in relatively noisy conditions. This provided good communication between the hearing and deaf users. The system also ensured low latency, which facilitated real-time processing and interaction.

The Neural Avatar module created expressive and natural sign language animations out of the text that was processed. The avatars were true to gesture and facial expression, increasing visual comprehension and user experience. The combination of the modules led to the effective end-to-end communication which was able to support both the sign-tospeech and speech-to-sign conversion.

Generally, the findings indicate that the suggested system can be an effective and practical solution to inclusive communication. It addresses several of the main drawbacks of current systems, such as the ability to recognize gestures continuously, to perform in real-time, and to engage in twoway communication. The system may be further improved by offering multiple language support, scaling, and implementing it in the real world like healthcare, education, and government services.

TABLE I
SYSTEM PERFORMANCE COMPARISON

| System | Real-Time | Continuous Gestures | ISL Support | Two-Way Communication | Accuracy |
|---|---|---|---|---|---|
| Template Matching (2021) | No | No | Limited | No | Low |
| Two-Stream DL (2022) | Partial | No | No | No | Medium |
| YOLO-Based (2020) | Yes | Partial | No | Partial | Medium |
| LLM-Based (2025) | Yes | No | Yes | Partial | High |
| Sensor-Based System | Yes | No | Limited | Yes | High |
| Proposed System | Yes | Yes | Yes | Yes | High |

## REFERENCES


[1] T. Reddy Gadekallu, G. Srivastava, and M. Liyanage, "Hand gesture recognition based on a Harris hawks optimized convolution neural network," *Computers & Electrical Engineering*, vol. 100, Art. no. 107836, 2022.

[2] G. T. R. Chiranji Lal Chowdhary and B. D. Parameshachari, *Computer Vision and Recognition Systems: Research Innovations and Trends*. CRC Press, 2022.

[3] M. M. Riaz and Z. Zhang, "Surface EMG real-time Chinese language recognition using artificial neural networks," in *Intelligent Life System Modelling*, Springer, vol. 1467, 2021.

[4] G. Halvardsson, J. Peterson, C. Soto-Valero, and B. Baudry, "Interpretation of Swedish sign language using convolutional neural networks and transfer learning," *SN Computer Science*, vol. 2, no. 3, pp. 1–15, 2021.

[5] P. Likhar, N. K. Bhagat, and R. G. N, "Deep learning methods for Indian sign language recognition," in *Proc. IEEE ICCE-Berlin*, 2020, pp. 1–6.

[6] F. Li *et al.*, "Deep transfer learning for time series data based on sensor modality classification," *Sensors*, vol. 20, no. 15, p. 4271, 2020.

[7] J. J. Bird, A. Ekart, and D. R. Faria, "British sign language recognition´ via late fusion of computer vision and leap motion," *Sensors*, vol. 20, no. 18, p. 5151, 2020.

[8] S. Sharma, R. Gupta, and A. Kumar, "TRBaggBoost: An ensemblebased transfer learning method applied to Indian sign language recognition," *J. Ambient Intell. Human Comput.*, 2020.

[9] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: A review," *J. Imaging*, vol. 6, no. 73, 2020.

[10] Z. M. Shakeel *et al.*, "MAST: Myo armband sign-language translator," in *Proc. IEEE ICTC*, 2020, pp. 494–499.

[11] M. Zakariya and R. Jindal, "Arabic sign language recognition system on smartphone," in *Proc. ICCCNT*, 2019.

[12] E. Abraham, A. Nayak, and A. Iqbal, "Real-time translation of Indian sign language using LSTM," in *Proc. GCAT*, 2019.

[13] O. Koller *et al.*, "Weakly supervised learning with multi-stream CNNLSTM-HMMs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2306–2320, 2019.]

[14] A. A. Hosain *et al.*, "Sign language recognition analysis using multimodal data," 2019.

[15] J. Huang *et al.*, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018.

[16] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Sign language recognition based on skeletal data," in *3DTV-CON*, 2018.

[17] C. Motoche and M. E. Benalcazar, "Real-time hand gesture recognition´ using EMG signals," in *ICANN*, Springer, 2018.

[18] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network for continuous sign language recognition," in *IJCAI*, 2018.

[19] R. Cui, H. Liu, and C. Zhang, "Recurrent CNN for continuous sign language recognition," in *CVPR*, 2017.

[20] S. Y. Kim *et al.*, "A hand gesture recognition sensor using reflected impulses," *IEEE Sensors Journal*, vol. 17, no. 10, pp. 2975–2976, 2017