IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

AI-Based Sign Language to Speech Converter

C.A Rasne, C.M Bhagat, S.A Dalvi, Ravindra Lawande
Electronics and Telecommunication Engineering
Dr. Vithalrao Vikhe Patil College of Engineering, Ahilyanagar, India

Abstract: *Sign language is a primary means of communication for individuals with speech and hearing impairments; however, communication becomes challenging when the listener does not understand sign language. This paper presents an AI-based sign language to speech converter that translates hand gestures into audible speech in real time. The proposed system uses computer vision techniques to capture and recognize hand gestures, which are then processed to generate corresponding text and speech output. The system is designed to be low-cost, portable, and efficient by utilizing a Raspberry Pi and external processing support. Experimental results demonstrate that the system provides accurate gesture recognition and smooth speech output, making it suitable for assistive communication applications.*

Keywords: Sign Language Recognition, Gesture-to-Speech Conversion, Computer Vision, Artificial Intelligence, Text-to-Speech, Assistive Communication, Raspberry Pi, Indian Sign Language

I. INTRODUCTION

The process of communicating with speech-impaired people and the general population at large becomes difficult due to the lack of understanding of sign language among non-signers. Sign language can be an effective tool in communicating with people, but this can only be done by people who understand sign language. Recent developments in computer vision and AI technologies have led to the development of machines that can effectively read and understand sign language through camera-based vision [1], [2].

The vision-based gesture recognition systems can effectively avoid the use of wearable devices, creating a more natural interaction environment. Deep learning techniques using Convolutional Neural Networks (CNN) have significantly improved the accuracy of vision-based gesture recognition systems, as they can effectively learn meaningful spatial features from images [2], [4]. These vision-based gesture recognition systems can effectively transform the recognized sign language into speech using text-to-speech systems and can be used to effectively communicate with people who know sign language and non-signers alike [3]. The main objective of this work is to develop an AI-based sign language to speech converter that can effectively recognize hand gestures and transform them into speech.

II. LITERATURE SURVEY

In the past, research in sign language recognition was carried out using sensor-based approaches that involved the use of data gloves with accelerometers and flex sensors to track hand
movements. Although this approach was accurate in terms of gesture recognition, it was disadvantageous due to its high cost and discomfort in continuous use [4]. In order to solve this problem, vision-based approaches have been proposed to recognize hand gestures using cameras.

Support Vector Machine (SVM) and k-Nearest Neighbour (KNN) classifiers were the first machine learning algorithms to be used in sign language recognition to recognize static hand gestures [5]. However, with the introduction of deep learning, Convolutional Neural Networks (CNN) have gained popularity in recognizing hand gestures due to their ability to extract spatial features from images, improving the accuracy of sign language recognition systems [2], [7]. In the case of dynamic hand gestures, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks have been used to recognize hand movements.

DOI: 10.48175/IJARSCT-34075

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Recently, the implementation of gesture recognition systems using embedded devices such as Raspberry Pi has been researched to create portable devices that are cost-effective and assistive [6]. The major challenge, however, lies in the maintenance of high accuracy with real-time processing capabilities. The surveys and recent research emphasize the need to create efficient and effective vision-based sign language recognition systems to be used in real-life applications [8, 9].

III. PROPOSED METHODOLOGY:

The proposed AI-Based Sign Language to Speech Converter can translate the hand gestures into speech through computer vision and AI. It follows a modular design, including image acquisition, preprocessing, gesture recognition, text generation, sentence formation, multi-language conversion, and speech synthesis. The overall system workflow can be represented through the block diagram shown in Fig. 1.

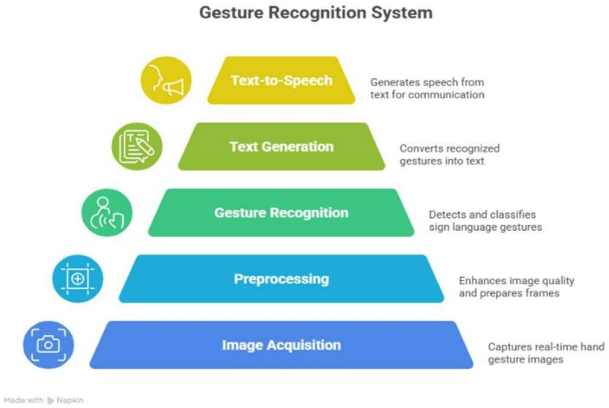

Figure 1.System modules

A. Image Acquisition: In the first step, real-time video frames are captured through the camera module connected with the system. The camera is used to capture hand gestures made by the user. These captured video frames are used as input data for the gesture recognition process. Vision-based systems are used instead of sensor-based systems because they are convenient for users, eliminating the need for wearable systems [1], [8].

B. Preprocessing: In the second step, the captured video frames are processed to enhance the quality of the captured images. Noise reduction is also done during the preprocessing step. Resizing the captured video frames, normalization of the background, and noise reduction are the preprocessing techniques used in this step. These techniques are used to ensure proper identification of the hand region from the background.

C. Gesture Recognition: The core component of the proposed system is the gesture recognition module. At this stage, the artificial intelligence models process the images and recognize the hand gesture being made. Feature extraction models are used to recognize the hand and the pattern of movement. Convolutional Neural Networks (CNNs) have been found to perform better in recognizing hand gestures and improving the accuracy of the model [2], [7].

D. Text Generation: Once the hand gesture is recognized, the system converts the recognized gesture into the corresponding text. Each sign language gesture has a corresponding word or phrase, which is stored in the database. This allows the system to translate the gestures into meaningful text.

E. Sentence Formation Module: While producing output based on the recognized gestures, the generated text might include isolated words or incomplete phrases that do not adhere to correct grammar rules. To solve this problem, a sentence formation module will be included which will help transform text into grammatically sound sentences.

The sentence formation module employs some of the simple NLP techniques such as rule-based grammar correction, word order adjustments, and contextual mapping of words. The purpose of the module is to ensure that the text generated makes sense to the recipient. For instance, if there is an output text consisting of the words “YOU GO MARKET,” it can be converted into “You are going to the market.”

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

F. Multi-language Conversion Module: In order to make the proposed system more user-friendly, it will have a feature of multi-language conversion module. Using the multi-language conversion module, the sentence generated can be translated into other languages such as Hindi, Marathi, or English.

For implementing the multi-language module, the use of API services for translating texts can be considered. By using this feature, the user can choose the regional language of his/her choice while interacting with the system.

G. Text-to-Speech Conversion: The generated text is then sent to the text-to-speech converter. The text-to-speech converter converts the generated text into sound. This allows people who are not familiar with sign language to understand the message being conveyed by the user.

H. Audio Output: Finally, the synthesized speech is played through a speaker connected to the system. This completes the gesture-to-speech conversion process, enabling effective real-time communication between sign language users and non-signers. The modular design of the system allows it to be implemented efficiently on embedded platforms such as Raspberry Pi for portable assistive applications [6].

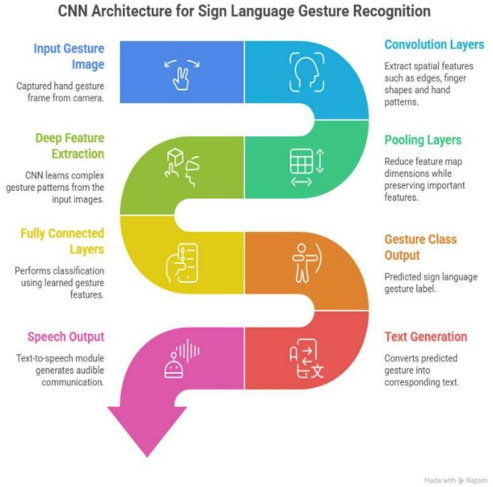

Figure 2. CNN Architecture for Gesture Recognition

A) Convolution Operation:

The spatial feature extraction is performed by the convolution layer on the given gesture image.

Let the image be represented as I(x, y) and the convolution kernel be represented as K(i, j).

Convolution operation is defined as:

$$(I * K)(x,y) = \sum_{i=-m}^{m} \sum_{j=-n}^{n} I(x+i, y+j)\, K(i,j)$$

where:

I(x, y) = pixel intensity of the input image K(i, j) = filter, or convolution kernel m, n = size of the filter

The process enables the network to recognize edges, shapes, and hand gesture patterns from the image.

B) Activation Function (ReLU):

Non-linear activation is achieved by using the Rectified Linear Unit (ReLU) function:

$$ReLU(x) = \max(0, x)$$

The ReLU function helps in enhancing the learning potential by introducing non-linearity and avoiding the vanishing gradients during training.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

C) Pooling Operation:

The Pooling layers reduce the spatial dimensions of the feature maps, and the important features are preserved. Max Pooling is often utilized.

$$P(x,y) = \max_{(i,j)\in\Omega} I(x+i, y+j)$$

Where:

Ω represents the region

The maximum value in the region is chosen

Pooling Operation improves computational efficiency and invariance to translation.

D) Feature Flattening:

The feature maps are then flattened into a one-dimensional vector after the convolution and pooling layers.

$$z = flatten(F)$$

where:

FFF = extracted feature maps.

zzz = flattened feature vector.

This vector is used as input to the fully connected layer.

E) Softmax Classification:

Once the convolution and pooling operations are performed, the feature maps are flattened into a vector that is one-dimensional in nature, given by:

Where,

$$P(y=k|z) = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}}$$

F = feature maps that are extracted ZK = feature vector that is flattened.

IV. SYSTEM DESIGN AND ARCHITECTURE

Circuit Diagram Description

1. Raspberry Pi Processing Unit:

The Raspberry Pi is used as the processing unit in the system. It receives input from the camera module and processes the captured gesture images using Python-based computer vision algorithms. The Raspberry Pi is used for image acquisition.

2. Camera Interface (CSI Port):

The system makes use of a camera module for capturing real-time hand gesture images using the Raspberry Pi's CSI port. The captured images are used as input for the sign language recognition system using the CNN-based gesture recognition model.

3. Output Display and Audio Module:

The Raspberry Pi is connected to a monitor using the HDMI interface for display and processing environment. The recognized gesture is converted to text and further converted to speech using a text-to-speech module. The generated speech output is transmitted through speakers using the audio jack for real-time speech communication.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

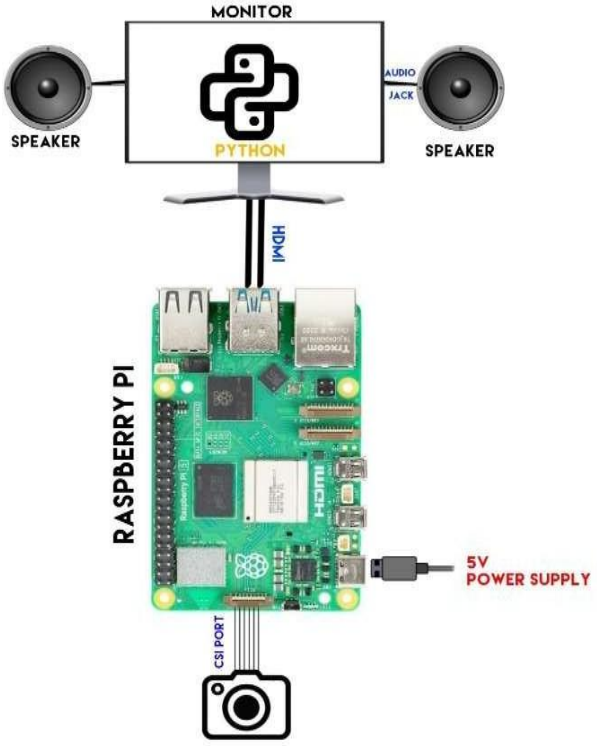

Figure 3. Circuit diagram

V. RESULT AND ANALYSIS

The proposed AI-Based Sign Language to Speech Converter was implemented and evaluated in real-time with the help of a webcam interface for the input module. The CNN-based gesture recognition module was found to be quite accurate for a predefined set of hand gestures. However, with an increase in the number of gesture classes, a slight decrease in the accuracy level was noticed due to the increased similarity between the gestures and the complexity of the classification process.

The proposed system was found to have low latency in converting the gestures into speech, thereby facilitating a smooth interface between the user and the system. The interface was successfully able to display the gesture label along with the corresponding text output, which was converted into speech with the help of the text-to-speech module. The experimental results clearly indicate that the proposed system is an efficient and cost-effective solution for facilitating communication between the sign language users and non-signers.

VI. CONCLUSION AND FUTURE WORK

In this paper, the authors proposed an AI-Based Sign Language to Speech Converter that can be of great use to individuals with speech and hearing impairments to enable real-time communication. The proposed system uses computer vision and the CNN gesture recognition model to recognize hand gestures and convert them into corresponding text and speech output. The proposed system has been observed to have reliable performance, low latency, and ease of use, making it suitable for practical use by individuals with speech and hearing impairments.

However, the proposed system has some limitations, including the reduced accuracy of the proposed system in handling large numbers of gestures, as well as the effect of lighting conditions and background on the proposed system. Future work on the proposed system can be directed towards improving the accuracy of the proposed system by using

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

deep learning techniques, increasing the number of gestures, and enabling the proposed system to recognize sign language continuously. Furthermore, the proposed system can be implemented on mobile devices to increase its usability and accessibility in real-world scenarios

REFERENCES

[1] P. Kumar and R. K. Aggarwal, "Vision-Based Sign Language Recognition Using Deep Learning," IEEE Access, vol. 8, pp. 198734–198745, 2020.
[2] S. Masood, M. A. Khan, and A. H. Mirza, "Hand Gesture Recognition Using Convolutional Neural Networks," IEEE Transactions on Human-Machine Systems, vol. 49, no. 1, pp. 1–12, Feb. 2019.
[3] A. Mittal, P. Kumar, and S. Roy, "Real-Time Sign Language Recognition System Using Computer Vision," Proc. IEEE Int. Conf. Signal Processing and Integrated Networks (SPIN), pp. 721–726, 2018.
[4] R. Sharma and T. K. Bhowmik, "A Vision-Based Sign Language Recognition System Using Artificial Neural Networks," Proc. IEEE Int. Conf. Advances in Computing, Communications and Informatics (ICACCI), pp. 1965–1970, 2017.
[5] A. G. Nandy, J. Saha, and S. C. Bagchi, "Indian Sign Language Recognition Using Deep Learning," Proc. IEEE Int. Conf. Intelligent Systems and Control (ISCO), pp. 1–6, 2021.
[6] S. Jain, V. Kanhangad, and S. D. Joshi, "Real-Time Embedded Sign Language Recognition System Using Raspberry Pi," Proc. IEEE Int. Conf. Communication and Signal Processing (ICCSP), pp. 1243–1248, 2019.
[7] M. H. Siddiqui and S. A. Khan, "Continuous Sign Language Recognition Using CNN-LSTM Framework," IEEE Access, vol. 9, pp. 115243–115254, 2021.
[8] T. Starner and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.
[9] H. Cooper, B. Holt, and R. Bowden, "Sign Language Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 71–80, Nov. 2012.
[10] J. Shotton et al., "Real-Time Human Pose Recognition in Parts from Single Depth Images," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1297–1304, 2011.