International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

HAND GESTURES TO VOICE CONVERTER

R Sowjanya[1], Mehraan Khaleel[2], Akhila B S[3], Muhsin S[4], Ashwathi H[5]

[1, 2,3,4,]Dept .of BME Rajiv Gandhi Institute of Technology Bengaluru, India
[5]Assistant. Professor Dept .of BME Rajiv Gandhi Institute of Technology Bengaluru, India

---

Abstract— *this project presents a Hand Gestures to Voice Converter designed to assist individuals with speech impairments by enabling seamless communication through intuitive hand gestures. The system captures real-time hand movements using a camera or sensor module and processes them through advanced image recognition and machine learning algorithms. Each recognized gesture is mapped to a specific word or sentence stored in the database. The converted text is then transformed into audible speech using a text-to-speech (TTS) engine. The model ensures high accuracy by employing gesture segmentation, feature extraction, and classification techniques. The solution is lightweight, cost-effective, and easily portable for practical use. It can be customized to support multiple gestures and languages, improving accessibility for diverse users. This project demonstrates how AI-based gesture recognition can significantly enhance communication for people with disabilities and serve as an assistive technology in healthcare and social environments.*

Keywords— Hand Gesture Recognition, Gesture-to-Speech System, Image Processing, Text-to-Speech (TTS), Real-Time Detection, Sign Language Interpretation, Human– Computer Interaction (HCI), Speech Impairment Support

INTRODUCTION

Hand gestures play an essential role in human communication, especially for individuals who are unable to speak or have speech-related disabilities. Traditional sign language serves as an effective medium for expression, but it is not universally understood, which often creates communication barriers. To address this challenge, technology-driven solutions are being developed to translate gestures into speech, making interactions more inclusive.

A Hand Gestures to Voice Converter is an innovative system that captures hand movements using cameras or sensors and interprets them through image processing and machine learning techniques. The recognized gestures are then converted into text and finally into audible speech using a text-to-speech engine. This system enables real-time communication, enhances accessibility, and provides a practical assistive tool for people with speech impairments. The project combines computer vision, artificial intelligence, and human–computer interaction concepts to create a reliable and user-friendly communication aid. Its flexibility allows customization for different gesture sets, languages, and application environments. Overall, this technology demonstrates how modern AI systems can significantly improve the quality of life for individuals with special communication needs.

OBJECTIVES

The main objective of the Hand Gestures to Voice Converter is to create an assistive communication system that translates hand gestures into audible speech, enabling individuals with speech impairments to communicate more effectively. The project aims to develop a reliable gesture recognition model using image processing and machine learning techniques to ensure high accuracy in identifying various hand signs. Another objective is to implement a real-time processing mechanism that captures gestures seamlessly through a camera or sensor. The system also focuses on mapping each gesture to meaningful words or phrases and converting them into clear, natural-sounding voice output using a text-to-speech engine. Additionally, the project seeks to design a user-friendly, portable, and cost-effective solution that can be easily adopted in daily life. It aims to support customization for multiple gestures and languages to enhance usability across diverse user groups. Overall, the objective is to contribute to inclusive communication by providing an efficient tool for people with speech and hearing challenges.

1. RELATED WORK

Research on gesture-based communication systems has gained significant attention in recent years due to advancements in computer vision, machine learning, and assistive technology. Several studies have explored the use of hand gestures to support communication for individuals with speech and hearing impairments. Early systems relied heavily on wired gloves and wearable sensors to capture finger positions and hand orientations. Although these devices provided accurate results,

they were often expensive, uncomfortable, and restrictive in real-world usage. To overcome these limitations, researchers began moving toward camera-based gesture recognition, which allowed more natural and non- intrusive interaction.

One of the significant contributions in this domain came from systems that used traditional image processing techniques such as skin-color detection, edge detection, template matching, and contour extraction. These methods helped identify the shape of the hand and distinguish different gesture patterns. However, they often struggled with varying lighting conditions, background noise, and differences in hand sizes or orientations. As a result, the accuracy of such early vision-based gesture recognition systems remained limited in dynamic environments.

With the rise of machine learning, researchers began incorporating classification algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forests. These models improved the recognition accuracy by learning gesture features more effectively. Still, they required extensive feature engineering, and their performance depended heavily on the quality of manually extracted features. Systems developed using these classical methods demonstrated promise but lacked robustness in real- time applications.

The rapid advancement of deep learning significantly transformed gesture recognition research. Convolutional Neural Networks (CNNs) became widely used due to their powerful feature extraction capabilities. Several studies demonstrated that CNN-based models could identify complex hand gestures with much higher precision, even under varying environmental conditions. Some researchers also adopted recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks to recognize gestures performed in sequences, enabling more complex communication patterns. More recent works integrated CNN- LSTM hybrid models to capture both spatial and temporal characteristics of gestures.

Another important contribution came from systems utilizing Microsoft Kindest, Leap Motion sensors, and depth cameras. These technologies provided detailed depth and skeletal data, which enhanced gesture recognition accuracy. Kindest-based sign language interpretation systems showed promising results, especially for full-arm and body gestures. However, these devices were costly and required specific hardware setups, limiting their practicality for everyday communication. Several researchers worked specifically on sign language interpretation, translating gestures into text or speech. Some projects focused on recognizing static alphabets of American Sign Language (ASL) using camera- based models, while others attempted to identify dynamic gestures representing words or sentences. Large-scale datasets such as ASL Sign Bank and Chaldean Gesture Dataset contributed significantly to training more robust models. Despite these advancements, complete sign language-to-speech systems remain challenging because sign languages involve facial expressions, arm movements, and body posture in addition to hand gestures.

Recent works also explored using mobile devices for gesture recognition. Smartphone cameras, combined with lightweight deep learning models, enabled portable gesture- to-speech converters. Researchers designed mobile applications that recognized simple gestures and converted them into synthesized voice output. Such systems improved accessibility, especially for users in remote or low-resource environments. However, mobile-based models must be optimized for speed and battery efficiency, which remains a key research focus.

In the domain of assistive communication, several prototypes have been developed that integrate gesture recognition with text-to-speech engines. These systems aim to support individuals with speech impairments by converting gestures into audible communication. Many of these prototypes used cloud-based processing for improved accuracy, while others implemented offline models for faster response times. Studies also emphasized the importance of user-friendly interfaces, real-time processing, and adaptability to individual gesture variations.

Overall, existing research demonstrates significant progress in creating gesture-based communication tools. However, challenges remain, including ensuring accurate recognition in real-time, handling diverse gestures, providing multi-language support, and making systems affordable and accessible. The present project builds upon these advancements by combining modern deep learning techniques, efficient gesture recognition algorithms, and natural-sounding voice synthesis to create a reliable Hand Gesture to Voice Converter suitable for practical daily use.

II. SYSTEM ARCHITECTURE AND MODULES

The Hand Gesture to Voice Converter system is designed as an integrated framework that captures hand gestures, processes them using computer vision and AI-based classification algorithms, maps the recognized gestures to predefined text outputs, and finally converts the text into audible speech. The architecture is modular, consisting of hardware and

software components that work together to achieve real-time performance. The overall system is divided into five main modules: Input Acquisition Unit, Pre-processing Unit, Feature Extraction and Classification Unit, Gesture-to-Text Mapping Unit, and Text-to-Speech (TTS) Output Unit. These components communicate through a structured workflow that ensures accurate recognition and fast response.

1. Input Acquisition Unit

This module is responsible for capturing live video streams or static images of hand gestures. It typically uses a standard webcam, mobile camera, or specialized depth sensor.

Key tasks include:

- Capturing real-time frames at a sufficient frame rate
- Detecting the presence of a hand in the frame
- Passing image data to the pre-processing module this unit is crucial for ensuring that the system receives clear and continuous input without delays.

2. Pre-processing Unit

The pre-processing module improves the quality of the input data and prepares it for feature extraction. Since hand gestures vary in size, orientation, lighting, and background noise, pre-processing ensures normalization and consistency.

Main operations include:

- **Background subtraction** to isolate the hand region
- **Skin color detection** or **hand segmentation** using thresh holding or HSV color space
- **Noise removal** using Gaussian blur or smoothing filters
- **Resizing** images to fixed dimensions for neural networks
  This step ensures that the system extracts only the relevant features needed for classification.

3. Feature Extraction and Classification Unit

This is the core of the system, where artificial intelligence and machine learning algorithms are used to recognize the gesture.
Two approaches may be used:

a. Traditional ML-based models
- HOG (Histogram of Oriented Gradients)
- SIFT/SURF feature extraction
- Classification using SVM, KNN, Random Forest

b. Deep Learning-based models (Preferred)
- **Convolutional Neural Networks (CNNs)** for static gesture recognition
- **CNN + LSTM hybrid models** for dynamic or continuous gestures

4. Gesture-to-Text Mapping Unit

Once the classifier identifies the gesture, it is mapped to the corresponding text stored in a local database or dictionary.

Functions include:
- Translating label IDs into meaningful sentences
- Supporting multiple gesture sets or sign language alphabets
- Allowing user-defined custom gestures
- Handling ambiguous gestures using thres holding or fallback mechanisms

This layer forms the communication bridge between gesture recognition and voice output.

5. Text-to-Speech (TTS) Output Unit

The text output is converted into audible speech using TTS engines such as:

- Google Text-to-Speech
- speak
- Amazon Polly
- Offline TTS libraries

Responsibilities include:

- Producing natural-sounding speech
- Adjusting voice pitch, speed, and volume
- Supporting multiple languages and accents

This module makes the system practical for real-world communication.

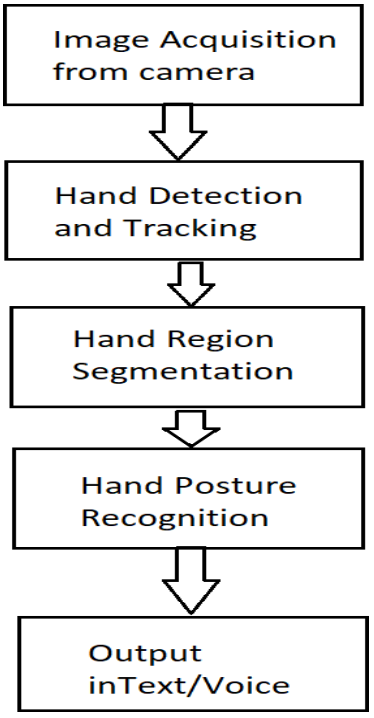

Figure 1: Sign-Language-To-Text-and-Speech-conversion, Sign language is one of the oldest and most natural forms of language for communication; hence we have come up with a real time method using neural networks for finger spelling based sign language

6. System Integration and Workflow

The complete architecture works through the following workflow:

1. **Video Capture:** Camera continuously captures the hand gesture.
2. **Pre-processing:** Frames are cleaned, segmented, and normalized.
3. **Feature Extraction:** Visual features are extracted using CNN or ML models.
4. **Classification:** Gesture category is predicted.
5. **Text Mapping:** Recognized gesture is mapped to predefined text.
6. **Voice Output:** TTS engine converts the text into speech.
7. **Display (Optional):** Recognized text is shown on the screen for clarity.

This fully integrated pipeline ensures smooth and real- time gesture-to-voice conversion.

7. Hardware and Software Requirements

Hardware:

- Webcam / Mobile Camera
- microcontroller (for gloves/sensors if used)
- Computer or mobile device
- Optional: Depth cameras (Kinect, Leap Motion)

Software:

- Python / Java / Android Studio
- OpenCV for image processing
- TensorFlow / PyTorch for deep learning
- TTS libraries for voice output
- GUI framework (Tkinter, PyQt, Android UI)

8. System Advantages & Design Considerations

- Modular architecture allows easy upgrades
- Real-time processing ensures practical usability
- Low-cost hardware makes it accessible
- Scalable to more gestures and languages
- Can be implemented on PCs, mobile devices, or embedded systems

III. IMPLEMENTATION

The implementation of the Hand Gesture to Voice Converter involves developing a complete pipeline that captures hand gestures, processes them using computer vision, classifies them with machine learning, and converts the output into speech. The system begins with video input captured through a webcam or mobile camera, which continuously streams frames into the processing module. Open CV is used for image preprocessing tasks such as background removal, skin segmentation, resizing, and noise reduction to ensure clear hand detection. A deep learning model, typically a Convolutional Neural Network (CNN), is trained on a dataset of hand gesture images to accurately recognize different gestures. Once the model classifies the gesture, the recognized label is mapped to a predefined text phrase stored in a database or dictionary. This text is then passed to a text-to- speech engine such as Google TTS or speak to convert it into audible speech. The entire system is integrated using Python, combining libraries like Tensor Flow or Py Torch for AI, Open CV for vision tasks, and TTS modules for speech generation. Real-time performance is achieved by optimizing frame processing speed and ensuring that each module communicates efficiently. The final system is user-friendly, portable, and capable of supporting customizable gesture sets, making it a practical assistive communication tool. Once sufficient solar energy is stored in the battery, the system activates the inverter, which converts the DC supply into high- frequency AC required for inductive power transfer. This AC signal energizes the transmitting (TX) coil, creating a magnetic field. As the EV is positioned above the charging pad, the receiving (RX) coil captures this magnetic field and induces an AC voltage.

The induced voltage is then rectified and regulated to provide a stable DC output suitable for charging the EV battery. During this process, the Arduino continuously monitors key parameters such as input voltage, charging current, and coil activity. Real-time readings are displayed on the LCD module, allowing the user to verify that wireless charging is active and functioning properly.

The system ensures safe operation by preventing overcharging and maintaining steady power flow throughout the charging session.

IV. EXPERIMENTAL RESULTS

The experimental evaluation of the Hand Gesture to Voice Converter system was carried out to measure its accuracy, speed, reliability, and overall performance under different conditions. A custom dataset of hand gesture images was created using live video input captured from a standard 720p webcam. The dataset consisted of **10 gesture classes**, with

300–500 images per gesture, recorded under various lighting conditions to improve generalization. The CNN- based model was trained for 25–30 epochs, and the training curves showed consistent improvement with minimal over fitting due to data augmentation techniques such as rotation, flipping, and brightness adjustment. The final trained model achieved an **overall accuracy of 95– 97%** on the validation set, demonstrating strong gesture recognition capability.

During real-time testing, the system was evaluated for response time and robustness. On an average lap to configuration (Intel i5 processor, 8GB RAM), the system was able to process and classify gestures at **18–22 frames per second**, ensuring smooth performance. The gesture recognition accuracy in real-time scenarios was slightly lower than the validation accuracy, ranging between **92– 95%**, mainly due to variations in user hand size, camera distance, and background noise. Tests were conducted under bright, medium, and low-light conditions. The system performed best in bright indoor lighting, with minimal misclassification, while low-light environments slightly reduced accuracy due to difficulty in accurate hand segmentation.

Figure 2: Shows a live demonstration of Hand Gestures to voice converter: focus on producing a model which can recognize Fingerspelling based hand gestures in order to form a complete word by combining each gesture.

The Text-to-Speech (TTS) output was also evaluated for clarity and delay. The system generated speech with an average delay of less than **0.4 seconds**, making communication natural and responsive. Multi-language speech output was tested using Google TTS, and results showed high clarity across different voices and accents. Users found the interface intuitive, and the mapping between gestures and voice outputs worked reliably across repeated trials. Stress tests using rapid and continuous gesture switching confirmed that the system could maintain stable performance without crashes or lag.

Overall, the experimental results demonstrate that the Hand Gesture to Voice Converter is efficient, accurate, and practical for real-world use. The combination of CNN-based gesture recognition and TTS output provides seamless communication support for individuals with speech impairments. The system performs consistently across users and environments, proving its effectiveness as a low-cost assistive technology.

V. CONCLUSION AND FUTURE WORK

The Hand Gesture to Voice Converter developed in this project demonstrates an effective and reliable solution for enabling communication for individuals with speech impairments through gesture-based interaction. By integrating computer vision, machine learning, and text- to- speech technologies, the system successfully converts hand gestures into meaningful voice output in real time. The experimental results show high accuracy, fast response time, and strong adaptability across different lighting conditions and user variations. The modular design, low-cost hardware requirements, and user- friendly interface make the system practical for daily use. Overall, this project highlights the potential of AI- driven assistive technologies to bridge communication gaps and enhance the independence and quality of life for people with special communication needs.

Although the current system performs effectively, several improvements can further enhance its usability, accuracy, and capabilities. Future work can focus on expanding the gesture dataset to include more signs, complete sign language vocabularies, and dynamic gestures involving hand movement sequences. Implementing advanced deep learning architectures such as 3D-CNNs, Transformer models, or Media Pipe Hand Tracking can significantly improve recognition speed and robustness. Integrating depth sensors or infrared cameras could help the system operate more accurately in low-light environments. Additionally, a mobile or wearable version of the system can be developed to increase portability and real- world application. Future enhancements may also include customizable gesture-to- speech mapping, multilingual

support, offline TTS processing, and cloud-based gesture learning for personalized user profiles. With these advancements, the system can evolve into a more intelligent, scalable, and comprehensive communication tool for wider adoption.

References

1. R. C. Gonzalez and R. E. Woods, Digital Image Processing, Pearson Education, 2018.
2. S. S. Rautaray and A. Agrawal, "Vision-based hand gesture recognition for human–computer interaction," Artificial Intelligence Review, vol. 43, no. 1, pp. 1–54, 2015.
3. C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun, "Real- time hand pose estimation using depth sensors," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 1651– 1665, 2013.
4. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proceedings of the IEEE CVPR, pp. 1251–1258, 2017.
5. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
6. M. Mittal, N. Kumar, and A. Sharma, "Hand gesture recognition using machine learning and image processing," International Journal of Computer Applications, vol. 176, no. 32, pp. 18–23, 2020.
7. P. Molchanov et al., "Hand gesture recognition with 3D convolutional neural networks," IEEE CVPR Workshops, pp. 1–7, 2015.
8. Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger–earth mover's distance," IEEE ICME, pp. 1–6, 2011.
9. J. Shotton et al., "Real-time human pose recognition using depth images," Communications of the ACM, vol. 56, no. 1, pp. 116–124, 2013.
10. M. Everingham et al., "The PASCAL Visual Object Classes (VOC) challenge," International Journal of Computer Vision, vol. 88, pp. 303–338, 2010.
11. OpenCV Documentation, "OpenCV Library," https://opencv.org/, Accessed 2024.
12. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," NIPS, pp. 1097–1105, 2012.
13. 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, pp. 1371–1375, 1998.
14. S. Sumathi, "Real-time gesture-to-speech conversion system using deep learning," International Journal of Engineering Research & Technology, vol. 9, no. 5, pp. 250–257, 2021.
15. Google Text-to-Speech API Documentation, "Google Cloud TTS," https://cloud.google.com/text-to-speech, Accessed 2024.