Grenze International Journal of Engineering and Technology, June Issue

Hybrid Sensor and Vision Assisted Gesture-to-Voice Conversion System for Real-Time Assistive Communication

Sudha K.L[1], Navya Holla K[2], Manasa R[3], Chaitra A[4], Ninu Rachel Philip[5] and Trupthi Rao[6]
[1-5]Dayananda Sagar College of Engineering, K.S Layout, Bengaluru, INDIA
Email: drsudha-ece@dayanandasagar.edu, hollanavya@gmail.com, manasar.87@gmail.com, chaitra.a9@gmail.com, ninurachelphilip-ece@dayanandasagar.edu
[6]Dayananda Sagar University, Bangalore, Karnataka, India
Email: trupthirao-aiml@dsu.edu.in

***Abstract*— Speech-impaired individuals are suffering from communication barriers which are leading to intelligent assistive systems capable of translating sign gestures into understandable speech. This work recommends a Gesture-to-Voice Conversion System which combines flex sensor-based glove recognition with vision-assisted machine learning-based gesture classification suitable for real-time speech synthesis. The system employs flex sensors which changes its resistance with bending, Arduino board for processing, Bluetooth communication, and Android text-to-speech interface. The software subsystem utilizes KNN-SVM assisted gesture recognition to enhance flexibility and scalability. Experimental results show gesture recognition accuracy of 92.8%, and end-to-end response latency below 750 ms. Implementation cost of the system is very low. Novelty in the work is dual-mode sensing, hybrid recognition, and real-time mobile speech integration.**

***Index Terms*— Gesture Recognition, Assistive Technology, Flex Sensors, Arduino, Text-to-Speech, Human Computer Interaction.**

I. INTRODUCTION

Speech-impaired individuals still relay on sign language for communication medium which has limited understanding among the general population and creates communication difficulties. Advances in electronics such as wearable sensing, embedded systems, machine learning and mobile computing can help in converting gestures into text and speech. There are many such systems to assist speech impaired people but are limited by environmental sensitivity, limited gesture vocabulary or computational complexity.

Identifying the research gap, this work proposes a hybrid sensor-vision assistive communication system combining wearable flex sensor glove detection and machine-learning based vision support with mobile speech synthesis. The proposed design provides better portability, affordability, and practical deployment. Key contributions of this paper are dual-mode gesture identification, real-time gesture-to-speech conversion under 1 second latency, low-cost wearable prototype using Arduino and Bluetooth, and scalable design for multilingual and AI-enhanced expansion.

Grenze ID: 01.GIJET.12.2.101_26

II. Related Work and Literature Review

Literature review indicates that the research is done in both vision-based and sensor-based gesture-to-speech systems. Khan et al.[1] suggested CNN-driven gesture-to-voice conversion attaining high classification accuracy but their method requires more computational resources. System developed by Zaidi et al.[2] is a bidirectional sign-speech translator with real-time performance. But system complexity in their work is also high. Gayitri et al.[3] designed glove-based flex sensor gesture recognition with portable implementation but their system has limited gesture vocabulary. MediaPipe and CNNs are used for accurate gesture classification in the paper by Sweta et al.[4]. But performance of the system in their design continued to be sensitive to background conditions. Gholap et al.[5] achieved 94.6% accuracy in their gesture recognition system using sensors and provided solutions for calibration challenges. Some of the papers published in conferences [5-10] provides implementation aspects of gesture to voice conversion and review of various techniques related to it. Literature tells us gaps in hybrid robustness, low latency wearable systems, and integrated low-cost real-time mobile speech conversion. Table I compares some of the papers which are recently published and identify few variations in their parameters.

Table I: Comparative Analysis of Existing Systems

Reference	Method	Accuracy (%)	Latency (ms)	Cost	Portability	Complexity
Khan et al.[1]	CNN Vision	95	1100	High	Medium	High
Zaidi et al.[2]	Bidirectional ML	93	1000	High	Medium	High
Gayitri et al.[3]	Flex Sensor	91	850	Low	High	Medium
Swetha et al[4]	Sensor Fusion	94.6	800	Medium	High	Medium
Gholap et al [5]	Hybrid Sensor-Vision	92.8	750	Low	High	Medium
Kadam et al[6]	Image Processing-based	89–91	Moderate (~1.2 s)	Medium	Camera + PC	Sensitive to illumination and background noise
Mitra et at[7]	Survey of Sensor and Vision Gesture Methods	85–92*	Moderate	Medium-High	Cameras / Sensors	Good conceptual framework, limited real-time deployment
Y. Wu et al[8]	Vision-Based Gesture Recognition Review	80–88	High latency	High	Camera Systems	Early models; poor robustness to occlusion and lighting
Molchanov et al [9]	3D CNN-based Gesture Recognition	96–98	Low (~0.8 s)	Very High	Depth Camera + GPU	Very accurate but computationally expensive
Shotton et al [10]	Depth Image Pose Recognition	94–96	Low (~0.9 s)	High	Depth Sensor (Kinect)	Strong real-time capability but hardware dependent

Existing studies indicate vision systems provide higher vocabulary but depend on environment. Sensor systems provide robust detection but suffer calibration issues. Hybrid models improve reliability but often increase complexity.

The proposed work balances these tradeoffs through lightweight hybrid architecture. The novelty of the proposed system lies in . Dual-Mode Hybrid Gesture Recognition, Low-Latency Embedded-Mobile Integration, cost-Effective Assistive Framework and Adaptive Expandable Architecture. The architecture supports extension toward multilingual output, AI adaptation and IoT integration.

III. System Overview

The system consists of two modules; Sensor-based Hardware Recognition Module shown in Fig.1 and Vision-Based Recognition Module shown in Fig. 2.

The proposed module 1 shown in Fig. 1 integrates a glove embedded with five flex sensors whose resistance variation is converted into analog voltages through voltage divider circuits. Arduino acquires and processes these signals through threshold-based gesture mapping and transmits recognized gestures via HC-05 Bluetooth to a mobile application where text-to-speech synthesis produces voice output.

In parallel, module 2 shown in Fig. 2, a software-assisted vision pipeline captures gestures through camera input followed by preprocessing, segmentation, PCA feature extraction and KNN/SVM classification. High reliability is obtained by combining robust hardware software intelligence.

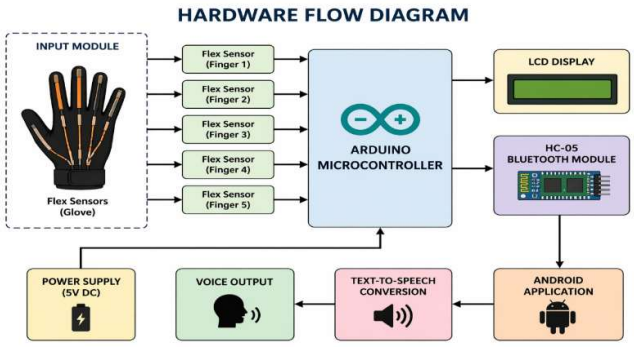

Figure 1. Hardware Module

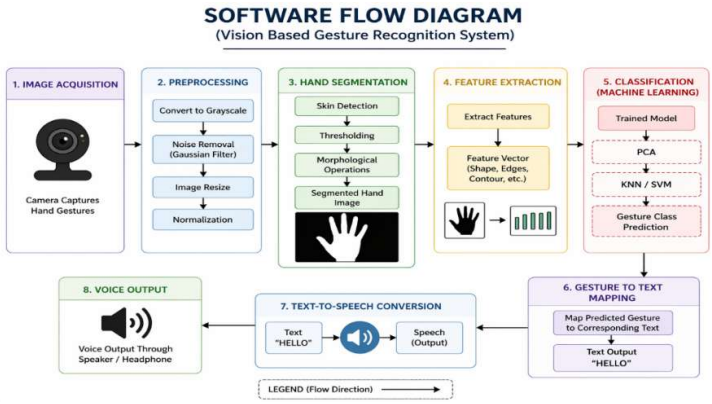

Figure 2. Software module

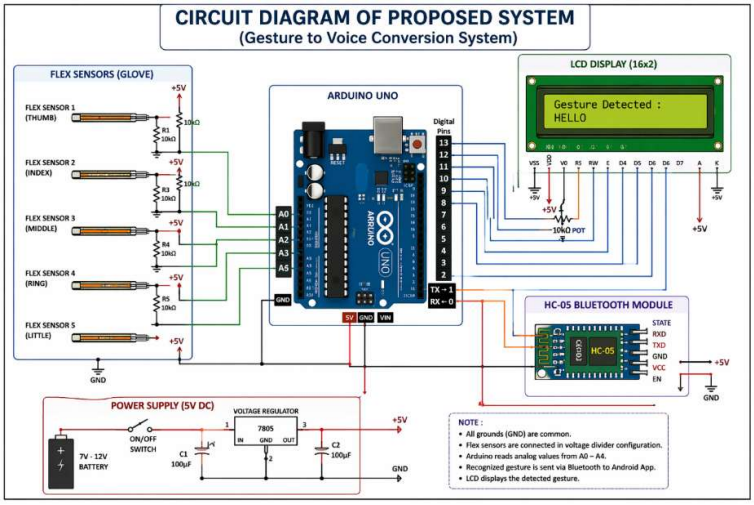

Figure 3. Connection diagram for hardware module

The circuit connection for module 1 is shown in Fig. 3 which comprises Flex Sensor Interface, Each flex sensor forms a voltage divider:

$$V_{out}=V_{cc}\left(\frac{R_{fixed}}{R_{fixed}+R_{flex}}\right) \quad (1)$$

Finger bending changes Rflex producing varying analog voltage for gesture discrimination. Arduino Processing Unit reads analog sensor values, performs ADC conversion, matches thresholds for gesture recognition and

sends output to LCD and Bluetooth module. Bluetooth Interface (HC-05) enables wireless transmission to Android speech application through UART communication at 9600 baud. LCD feedback Provides local gesture validation. Once the system is initialized, flex sensor values are read, preprocessing and normalizing data happens. Values are compared with gesture threshold database to identify the gesture. Bluetooth is used to transmit. At the receiving end mapped text is converted into speech.

IV. Results and Discussion

The hardware implementation is shown in Fig. 4 with flex sensors connected to glouse which intern is connected to ESP 32. Gesture recognition results with flex sensors as well as vision based CNN/SVM based recognition is shown in Fig. 4 and Fig. 6.

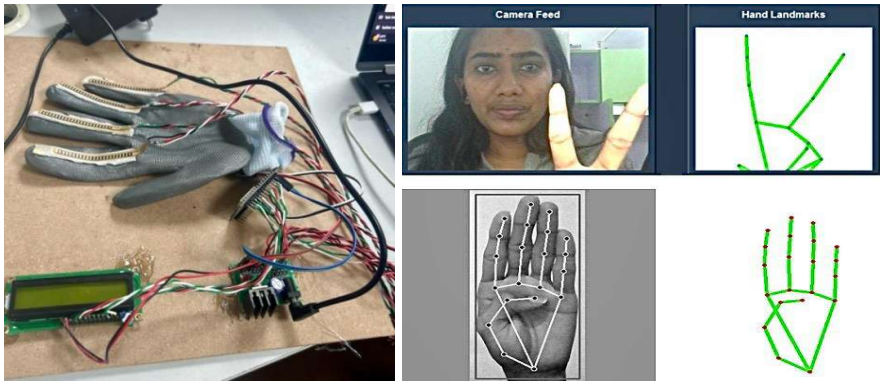

Figure. 4 Hardware connection and gesture recognition

Experimental Parameters considered are listed below in Table II and Performance Comparison is shown in Table III.

Table II: Experimental Parameters

Parameter	Value
Number of Gestures	5
Test Trials	50/gesture
Avg Accuracy	92.8%
Avg Latency	750 ms
Bluetooth Range	10 m

Table III: Performance Comparison

Metric	Proposed
Accuracy	92.80%
Precision	91.80%
Recall	92.80%
Avg Latency	750 ms
Throughput	High

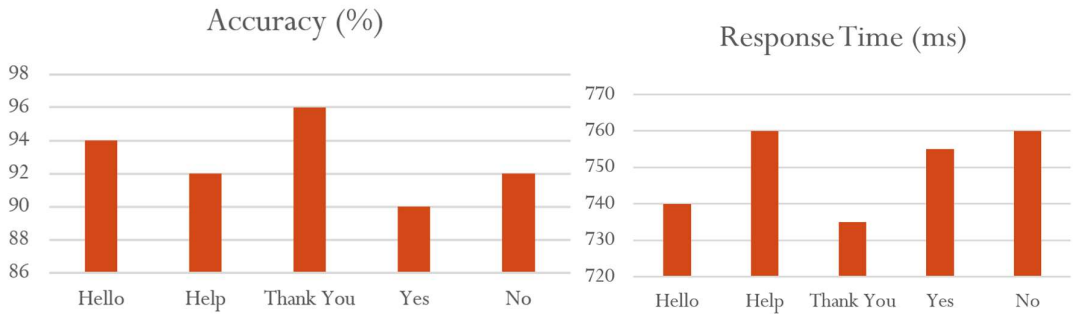

Figure 5. Performance statistics a. accuracy in recognition and b. Response time

Accuracies ranging from 90% to 96% is guaranteed with this system as the recognition accuracy graph in Fig. 5a confirms consistent classification performance across all tested gestures. The gesture Thank You exhibits the highest recognition accuracy due to distinct finger-bending patterns.

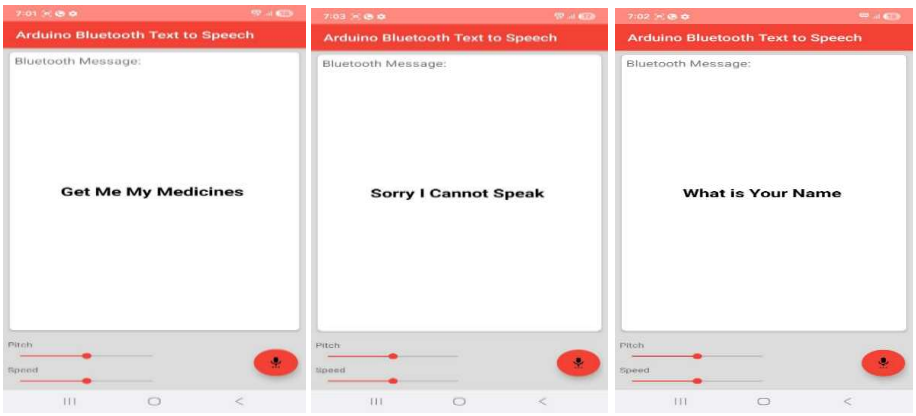

Figure 6. Gesture Recognition outputs-Text and Voice

Slightly lower performance for Yes is because of overlap in flex sensor symbols with neighboring gesture classes, causing sometimes ambiguity. Mean system recognition rate is 92.8% which confirms consistent operation of the hybrid sensor-assisted design. The relatively small variation between highest and lowest accuracy values indicates stable calibration and sensor responses. The proposed approach shows improved performance compared with literature-reported sensor-based systems. Its reduced complexity and lower implementation cost are the major advantages. The response time graph in Fig. 5b tells that total end-to-end latency is below 750 ms, and satisfy the real-time communication necessities. Text-to-speech synthesis requires the largest delay, which is more than half of overall latency. Sensor acquisition and embedded processing operations takes comparatively minimal time, confirming efficiency of the Arduino-based implementation.
Fig(7)shows latency profile indicates that communication and speech synthesis contributes maximum delay compared to sensing or classification. This is significant and need to be addressed by considering reduced algorithmic complexity in future versions without affecting response performance. The low response time confirms that the proposed system can support real time conversations without time lag.

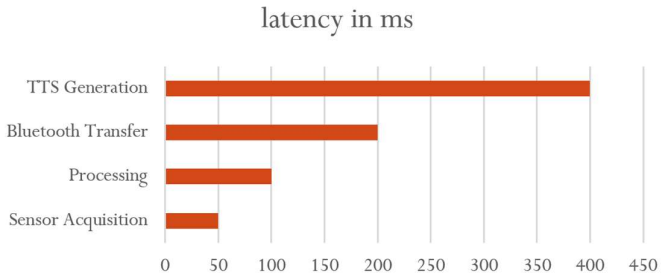

Figure 7. Analysis of Latency Performance

It is evident from the overall results, that integrating wearable sensing with lightweight intelligent processing delivers platform suitable for embedded implementation where in edge-AI extensions are possible. Flex sensor hardware guarantees stable gesture identification, free of lighting conditions while vision support deals scalability. Hybrid system improves robustness compared to standalone constructions. Performance results show satisfactory balance between cost, latency and accuracy. Minor errors are seen due to gesture overlap and calibration drift, where one can bring in the concepts of AI-based adaptive calibration.

V. CONCLUSION

The paper presents a hybrid Gesture-to-Voice Conversion System for hearing impaired to communicate effectively with other people. The proposed system combines flex sensing glouse with vision-based machine

learning. Bluetooth communication and mobile speech generation are added to deliver practical real-time gesture-to-speech translation. Experimental results confirms 92.8% average accuracy with below one second latency and trustworthy communication. The suggested method improves sturdiness and can be constructed with low cost. The work helps differently abled individuals to communicate effectively with others. Deep learning models such as CNN/LSTM can be used in future to go for Glove-free gesture recognition. Work can be extended for continuous sign sentence recognition, multilingual speech output etc.

ACKNOWLEDGEMENT

We thank the students ABHIGNYA MELINN, DWITI R N, JEEVIKA, MUKTHA S R for getting involved in this project and helping us to construct the circuit.

REFERENCES

[1] A. A. Khan, H. Raza Zaidi, and M. Usman, "Hand Gesture to Voice Conversion Using Artificial Intelligence," Spectrum of Engineering Sciences, vol. 3, no. 2, pp. 45–52, 2025.

[2] H. R. Zaidi, S. Ahmed, and M. Tariq, "Real-Time Bidirectional Translation Between Sign Language and Voice," International Journal of Advanced Engineering Research and Science, vol. 10, no. 1, pp. 112–120, 2025.

[3] G. H. M., S. R. Kumar, and P. S. Rao, "Hand Gesture Recognition and Voice Conversion for Speech Impaired," Mathematical Statistician and Engineering Applications, vol. 71, no. 4, pp. 567–575, 2022.

[4] S. Sweta, R. Singh, and A. Gupta, "Sign Language to Text and Speech Conversion Using CNN and MediaPipe," International Journal of Innovative Science and Research Technology, vol. 10, no. 5, pp. 123–130, 2025.

[5] S. Gholap, P. Patil, and R. Deshmukh, "Sensor-Based Sign Language to Speech Conversion System," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 13, no. 3, pp. 210–218, 2025.

[6] R. Kadam and V. S. Pawar, "Hand Gesture Recognition System Using Image Processing," IEEE International Conference on Computing, Communication and Automation, pp. 123–127, 2018.

[7] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," IEEE Transactions on Systems, Man, and Cybernetics, vol. 37, no. 3, pp. 311–324, 2007.

[8] Y. Wu and T. Huang, "Vision-Based Gesture Recognition: A Review," IEEE International Gesture Workshop, pp. 103–115, 1999.

[9] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand Gesture Recognition Using 3D Convolutional Neural Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–7, 2015.

[10] J. Shotton et al., "Real-Time Human Pose Recognition in Parts from Single Depth Images," IEEE CVPR, pp. 1297–1304, 2011.