

Hands To Voice: Real Time Sign Language Translator

Mr. K. Annaji Rao¹, G. Lalitha Sravani², K. Isaac Ratna Kanth³, K. Shashank⁴

¹Assistant Professor, Department of CSE (AI & ML), VBIT, Telangana, India

²Student, Department of CSE (AI & ML), Vignana Bharathi Institute of Technology, Telangana, India

³Student, Department of CSE (AI & ML), Vignana Bharathi Institute of Technology, Telangana, India

⁴Student, Department of CSE (AI & ML), Vignana Bharathi Institute of Technology, Telangana, India

Abstract - Hands to Voice is an AI-powered application that enables real-time translation of sign language gestures into audible speech and text, aiming to bridge the communication gap between hearing impaired individuals and the broader community. The system utilizes computer vision techniques with MediaPipe for accurate hand landmark detection, and a deep learning-based gesture classification model (CNN or LSTM) trained on sign language datasets such as ASL. Recognized gestures are mapped to corresponding words or phrases and converted into natural speech using Text-to-Speech (TTS) technology. A user-friendly interface, developed using Python frameworks like Streamlit, allows live webcam input for hands-free interaction with mobile optimization through TensorFlow Lite, the system has potential for real-world deployment in schools, hospitals and public spaces to foster inclusive communication.

Key Words: Sign Language, Gesture Recognition, Deep Learning, Computer Vision, Text-to-Speech, MediaPipe, Accessibility, Python, CNN, LSTM, Real-time Translation.

1. INTRODUCTION

Communication is essential for human interaction, yet individuals with hearing and speech impairments often struggle to communicate with those unfamiliar with sign language. Traditional methods like interpreters or writing are not always practical, creating a need for automated solutions. Advances in computer vision and deep learning have enabled sign language recognition systems that use camera-based inputs to translate gestures. Earlier approaches relied on basic image-processing techniques, while modern methods use CNNs for feature extraction and LSTM/GRU models to capture temporal patterns in gestures. A significant breakthrough came with Google Research's MediaPipe framework, which provides real-time, accurate hand landmark detection, improving robustness in varied conditions. Building on these developments, the Hands to Voice project combines MediaPipe with sequential deep-learning models to deliver a real-time gesture recognition system, enhanced with text-to-speech output, making communication more accessible and practical in everyday scenarios.

2. Body of Paper

The proposed system uses deep learning with LSTM models and a camera-based approach to translate hand gestures into speech in real time.

It provides a cost-effective, scalable, and user-friendly solution to bridge the communication gap.

2.1 Data Collection and Preprocessing

The system is trained and tested on datasets of American Sign Language (ASL) gestures. Each sample consists of sequences of frames where hand landmarks are extracted for training. Public datasets such as Sign-MNIST datasets are utilized, and additional toy datasets are generated for system validation. Data augmentation techniques such as rotation, scaling, and background variation are applied to improve generalization. The dataset is split into training, validation, and testing sets to ensure fair evaluation.

2.2 Approach

The approach integrates computer vision and deep learning for robust sign language translation. Instead of relying on external hardware like gloves, the system leverages camera-based recognition to ensure accessibility and cost-effectiveness. Landmark-based features from MediaPipe are passed into a sequence learning model (LSTM/GRU) for temporal classification of gestures. A Streamlit-based user interface is designed for real-time interaction, displaying recognized text and playing back audio. Optimization through TensorFlow Lite ensures deployment feasibility on laptops and mobile devices.

2.3 Machine Learning Model

Recurrent models such as LSTM and GRU are chosen due to their ability to capture temporal dependencies in gesture sequences. A simple CNN/MLP baseline is also considered, but sequence-based models demonstrate higher accuracy for dynamic gestures. The models are trained with categorical cross-entropy loss and optimized using the Adam optimizer. For deployment, the best-performing model is quantized and exported to TensorFlow Lite to ensure low latency. This balance between accuracy and efficiency makes LSTM/GRU the most suitable choice for the project.

2.4 Feature Engineering

The main features are the normalized 3D coordinates of 21 hand landmarks extracted from each frame by MediaPipe. These landmarks are further processed into temporal sequences to capture motion and gesture dynamics. Additional features such as angles between key joints are computed to enhance recognition accuracy. Sliding window techniques are applied to stabilize predictions across consecutive frames. This structured representation significantly reduces noise and improves model robustness in varied environments.

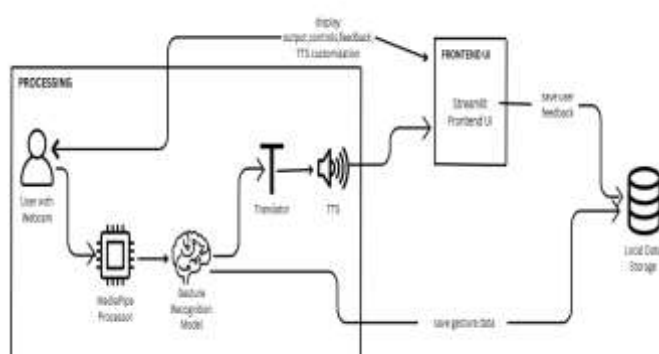


Fig -1: System Design



Fig -4: Result- 3



Fig -5: Result- 4

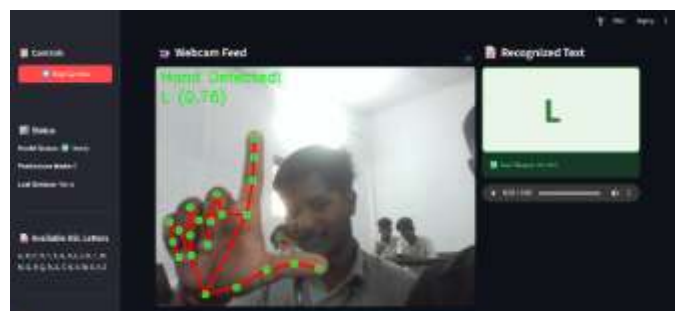


Fig -2: Result 1

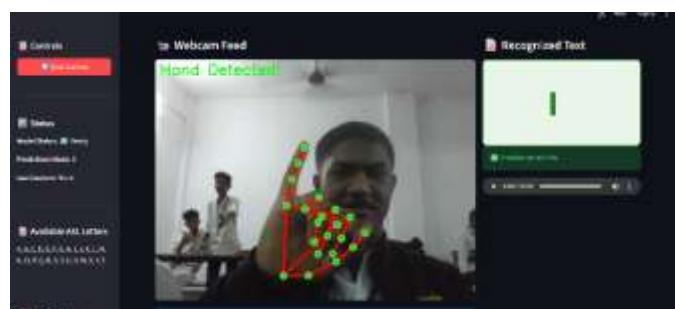


Fig -3: Result 2

3. CONCLUSIONS

The Hands to Voice project successfully demonstrates a real-time, AI-driven system capable of converting sign language gestures into both audible speech and readable text, effectively narrowing the communication gap between the hearing- and speech-impaired community and the general public. By combining MediaPipe for precise hand landmark detection, LSTM-based deep learning models for gesture classification, and Text-to-Speech (TTS) for natural audio output, the system delivers an accurate, efficient, and cost-effective solution without relying on specialized hardware such as sensor gloves.

This vision-based approach ensures a more accessible and user-friendly experience, making it suitable for everyday use. Furthermore, the modular and scalable architecture of the system allows for significant future enhancements, including expanding gesture datasets, supporting multiple sign languages, and optimizing the model for mobile deployment through TensorFlow Lite. The project demonstrates the potential of computer vision and deep learning in building inclusive technologies that enhance accessibility and human interaction. With continued refinement and real-world user testing, Hands to Voice can evolve into a reliable assistive tool that meaningfully improves communication and inclusivity for individuals with hearing or speech impairments.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our guide Mr. K. Annaji Rao, Assistant Professor, Department of Computer Science and Engineering (AI & ML), Vignana Bharathi

Institute of Technology, for his continuous support, valuable guidance, and encouragement throughout the development of this project.

We also thank the management and faculty members of Vignana Bharathi Institute of Technology for providing the necessary resources and support. Finally, we express our gratitude to our friends/team members for their cooperation and contribution in completing this work successfully.

REFERENCES

- [1] Sharma, P., & Reddy, S. (2021). Vision-Based Indian Sign Language Recognition Using CNN and LSTM. *International Journal of Computer Applications*, 183(25), 15–20.
- [2] Patel, R., & Bansal, V. (2023). Hybrid LSTM–GRU Model for Dynamic Sign Language Recognition. *Journal of Artificial Intelligence Research*, 72(4), 233–240.
- [3] Mehta, R., & Das, A. (2023). AI-Powered Assistive Communication Tool for the Deaf Using CNN and TTS. *IEEE Access*, 11, 12045–12053.
- [4] Gupta, N., & Sharma, T. (2021). Real-Time American Sign Language Detection Using TensorFlow and OpenCV. *Proceedings of the IEEE Conference on Computer Vision Applications*.
- [5] Singh, K., & Verma, L. (2022). Dynamic Gesture Recognition Using MediaPipe and LSTM Networks. *International Journal of Advanced Computer Science and Applications*, 13(8), 90–96.