Volume 10, Issue 6, June – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun285

Sign Language to Text and Speech Conversion

(Empowering Communication Beyond Gestures)

W. Sweta[1]; J. Kartiki[2]; K. Prerana[2]; M. Aarya[2]; P. Rutuja[2]

[1]Project Guide, [2]Student

[1,2] Department of Artificial Intelligence and Data Science Engineering
Ajeenkya D.Y Patil School of Engineering Pune, India

Publication Date: 2025/06/12

Abstract: This report presents the design and development of a Sign Language to Text and Speech Conversion System. The main goal of this project is to improve communication for people who are deaf or hard of hearing by translating sign language gestures into text and spoken words in real time. This helps bridge the communication gap between sign language users and people who don't know sign language, making daily conversations easier and more inclusive. Our system uses a gesture recognition model based on Convolutional Neural Networks (CNNs) to accurately detect hand gestures that represent different signs. One of the major challenges in this process is handling changing lighting conditions, various backgrounds, and differences in hand shapes or gestures. To overcome these issues, we use vision-based techniques and landmark detection with the help of the MediaPipe library, which enhances the accuracy and performance of the system. After recognizing a gesture, the system converts it into text and uses Text-to-Speech (TTS) technology to generate clear spoken output. This allows people with hearing disabilities to communicate more smoothly with those unfamiliar with sign language, making interactions quicker and more effective. The report also discusses the positive impact this technology can have in places like schools, offices, and public service areas. It emphasizes the importance of ongoing improvements in machine learning and computer vision to make such systems even more reliable and user-friendly. Overall, this project highlights how modern technology can promote a more inclusive, accessible world for everyone.

Keywords: *Sign Language to Text and Speech Conversion, Gesture Recognition, Convolutional Neural Networks (CNN), Text-to-Speech, Accessibility, Inclusivity, Computer Vision, MediaPipe, Real-Time Communication.*

How to Site: W. Sweta, J. Kartiki; K. Prerana; M. Aarya; P. Rutuja; (2025); *Sign Language to Text and Speech Conversion*. *International Journal of Innovative Science and Research Technology*, 10(6), 141-147. https://doi.org/10.38124/ijisrt/25jun285

I. INTRODUCTION

Communication is a fundamental human activity, but for over 466 million people with disabling hearing loss (WHO, 2023), spoken language can be a barrier. Sign language provides a natural and expressive way to communicate, yet in mainstream settings like schools, workplaces, and public services, most people do not understand it, making interaction difficult for deaf individuals.

To address this, our project proposes a Sign Language to Text and Speech Conversion System. It aims to improve communication for deaf and hard-of-hearing individuals by translating sign language gestures into text and audible speech in real time. The system uses computer vision, deep learning (CNNs), and text-to-speech (pyttsx3) technology. MediaPipe is integrated for accurate hand landmark detection under varied conditions, enhancing gesture recognition stability. The identified gesture is transcribed and then transformed into clear speech output. This solution is especially valuable in places requiring instant verbal interaction and contributes to creating a more inclusive society. Future improvements could involve recognizing phrases and enabling two-way communication by converting spoken words into sign language animations.

In addition to facilitating personal conversations, this system has potential applications in public services such as hospitals, banks, transportation hubs, and educational institutions where quick, effective, and accessible communication is essential. By providing an intuitive and efficient communication bridge, it can help reduce social isolation and ensure that deaf and hard-of-hearing individuals can engage confidently in everyday interactions. Moreover, this project highlights how advancements in artificial intelligence and computer vision can be meaningfully applied to create assistive technologies that directly address real-world challenges and promote equal opportunities for all.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun285

II. PROBLEM STATEMENT

Daily interactions present significant challenges for deaf community members, particularly in environments where sign language literacy is limited among the general population. To address this, we propose developing a mobile application that uses machine learning, built with TensorFlow and Keras, to recognize and translate sign language gestures from live video feeds into text and speech. The app aims to provide real-time, accurate translations to support natural conversations in various environments.

Key challenges include ensuring reliable gesture detection across different lighting conditions and backgrounds, maintaining fast processing for smooth interaction, and offering a simple, accessible interface for users with diverse technical skills. The system will feature offline functionality for areas with poor internet connectivity and continuous learning capabilities to improve recognition accuracy over time. Additionally, it will support multiple spoken and sign languages for global accessibility.

This solution aspires to empower deaf individuals, promoting independence and inclusion in everyday social, professional, and emergency situations

III. METHODOLOGY

- *Literature Survey and Research:*

We conducted a comprehensive analysis of contemporary sign language recognition technologies, examining their technical approaches, performance limitations, and successful implementation strategies. Examine recent advancements in computer vision and deep learning technologies to guide the selection of appropriate methods for gesture detection and translation.

- *Data Collection and Preparation:*

Compile a comprehensive and diverse dataset of sign language gestures through video recordings, ensuring it includes variations in signers, dialects, and environmental contexts. Apply preprocessing techniques like normalization, resizing, and data augmentation to improve dataset quality and diversity.

- *Model Selection and Development:*

Identify and select suitable deep learning models, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), optimized for processing image and video data. Implement these models using Tensor Flow and Keras, focusing on achieving high recognition accuracy and system reliability.

- *Model Training and Validation:*

Divide the dataset into training, confirmation, and testing subsets for methodical model development and evaluation. Use techniques like cross-validation, hyperparameter tuning, and regularization to enhance model performance while preventing overfitting.

- *Real-Time System Implementation:*

Design and develop a real-time processing system that captures video input, detects gestures, and converts them into text and speech outputs. Prioritize optimizing system performance for low latency and efficient resource management to enable smooth, uninterrupted conversations.

- *System Evaluation and Performance Metrics:*

Indicate evaluation criteria, like delicacy, perfection, recall, and F1 score, to measure the model's recognition capabilities. Conduct user experience studies to assess the system's overall effectiveness, accessibility, and impact on improving everyday communication for deaf users.

IV. SYSTEM ARCHITECTURE

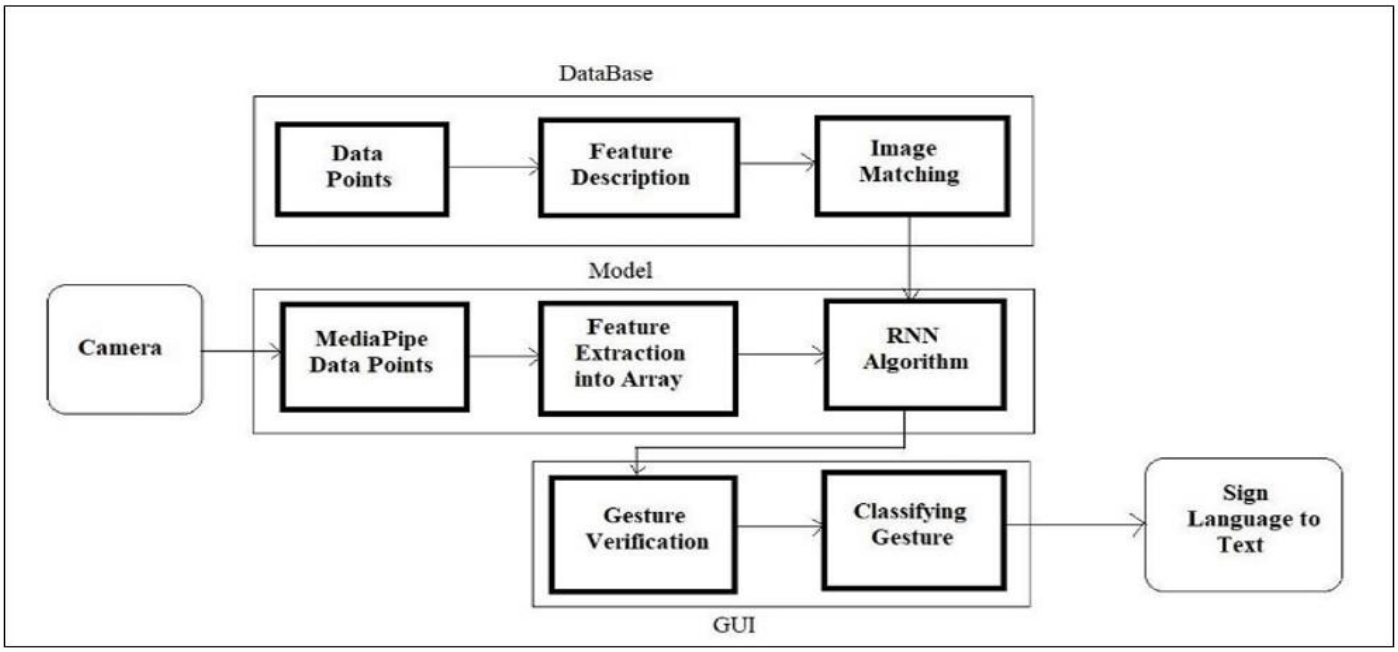

Fig 1 System Architecture

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun285

The architecture of the Sign Language to Text and Speech Conversion System is designed to facilitate seamless, real-time recognition and translation of hand gestures into text. This complex system integrates several tightly coupled components, primarily divided into three major modules: the Database, the Model, and the Graphical User Interface (GUI). Each of these modules plays a crucial role in ensuring accurate gesture detection, processing, and user interaction.

At the front end of the system lies the camera, which continuously captures video streams of the user's hand gestures. This live video feed acts as the raw input, feeding dynamic visual data into the processing pipeline. The choice of camera and its frame rate significantly affect the system's responsiveness and accuracy. Higher frame rates enable smoother tracking of gestures, while the resolution ensures sufficient detail for detecting subtle finger movements. The video frames captured by the camera are then processed by the subsequent modules in real-time.

The Database serves as a comprehensive repository that stores reference information about various sign language gestures. It maintains a rich collection of data points representing critical hand landmarks that define each gesture. These data points typically include spatial coordinates of key locations such as fingertips, joints, and palm center, extracted from a diverse dataset during the training phase. To transform these raw data points into a format conducive for recognition, the database module includes a feature description component. This component extracts meaningful features—such as relative positions, angles between fingers, and normalized distances—that characterize each gesture in a compact and discriminative manner. Such feature descriptors reduce the dimensionality of the data while preserving distinctive attributes necessary for differentiating between similar signs. The database also supports an image matching process wherein the features extracted from the live input are compared against stored gesture representations. This matching involves calculating similarity measures or distances between the incoming gesture's feature vector and the reference vectors in the database. By narrowing down potential candidates through this matching step, the system efficiently guides the classification process towards accurate recognition.

Central to the system's processing is the Model module, responsible for interpreting the continuous video stream and identifying the performed gestures. Leveraging the capabilities of the Google-developed MediaPipe framework, the model begins by extracting precise, real-time hand landmarks from each video frame. MediaPipe detects multiple key points on the hand—tracking the three-dimensional coordinates of joints and fingertips with robustness even under varying lighting conditions and complex backgrounds. These extracted landmarks are then transformed into structured numerical arrays that encapsulate the spatial configuration of the hand at each point in time. This transformation includes normalization steps to standardize the data across different users and conditions, as well as the computation of derived metrics like joint angles or finger curvature, enhancing the feature richness. Since sign language gestures inherently consist of temporal sequences rather than static poses, the system employs a Recurrent Neural Network (RNN) algorithm tailored to process sequential data. The RNN analyzes the series of feature arrays extracted from consecutive video frames, capturing temporal dynamics and motion patterns critical for distinguishing gestures that may be visually similar when viewed in isolation. By modeling the sequential dependencies and transitions between hand positions, the RNN can infer the intent behind the movement, substantially improving recognition accuracy.

The Graphical User Interface (GUI) module represents the final stage of the system, focusing on user interaction and output presentation. It serves as a medium through which recognized gestures are verified, classified, and displayed to the user in textual form. Prior to finalizing the classification, the GUI incorporates a gesture verification step to ensure the completeness and validity of the detected sign. This filtering process reduces misclassifications by eliminating incomplete or ambiguous gestures that may result from transient hand movements or occlusions. Following verification, the GUI uses the classification results generated by the RNN and cross-references them with database matches to confirm the gesture identity. Once the gesture is confidently recognized, the GUI translates the classification into corresponding text, which is then presented on the screen for the user to read. This text output not only assists hearing-impaired individuals in communicating but also serves as the input for any subsequent text-to-speech conversion module that may be integrated to provide audible translations.

The system architecture reflects a sophisticated integration of data storage, advanced machine learning, and user-centered design. The continuous video stream from the camera flows seamlessly through the MediaPipe-powered feature extraction and RNN-based temporal analysis, enhanced by robust matching with a well-curated gesture database, and finally delivered to the user via a responsive graphical interface that verifies and presents the results. This layered approach enables accurate, real-time sign language recognition while ensuring ease of use and practical utility in real-world communication scenarios.

V. IMPLEMENTATION

The implementation of sign language to text and speech conversion systems leverages advanced technologies such as Convolutional Neural Networks (CNNs) and MediaPipe to enhance real-time communication for individuals with hearing and speech impairments. By utilizing deep learning techniques, these systems employ computer vision to accurately recognize and interpret sign language gestures captured through video input. CNNs play a crucial role in feature extraction, allowing the model to learn complex patterns from the visual data, which significantly improves the accuracy of sign language translation. MediaPipe provides a robust framework for hand tracking and gesture recognition, facilitating seamless integration into real-time applications.

Once gestures are recognized, the system converts them into text format, which is then synthesized into speech using Text-to-Speech (TTS) technologies, such as Google Text-to-Speech (gTTS). This process not only enhances accessibility

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun285

for sign language users but also incorporates Natural Language Processing (NLP) to ensure that the generated text is contextually appropriate and coherent. Overall, the combination of these technologies fosters effective communication, bridging the gap between sign language users and the broader population.

VI. RESULTS

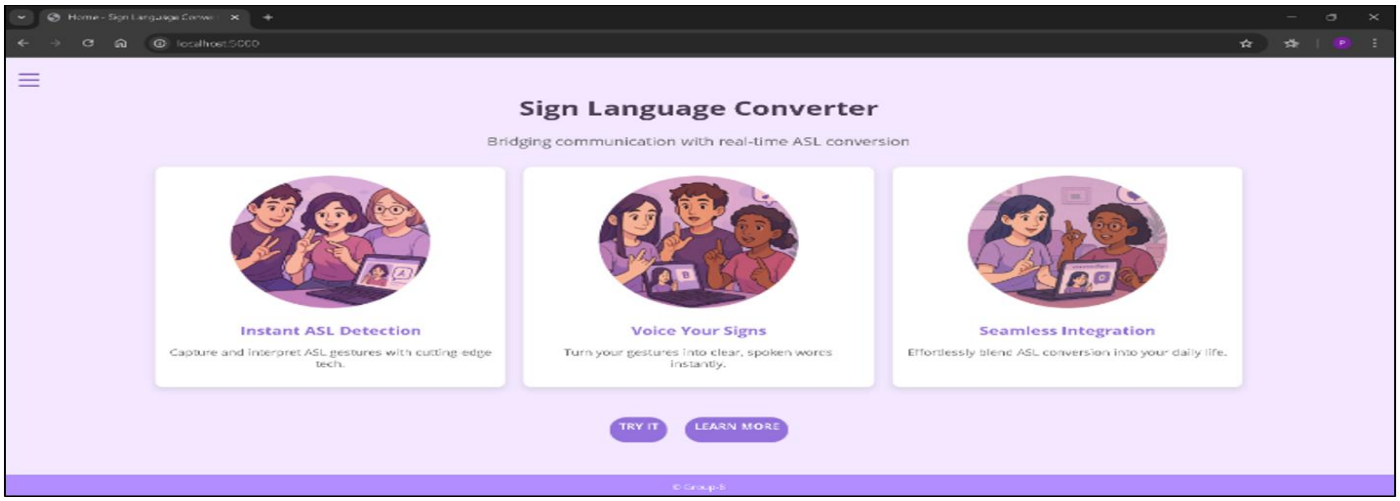

Fig 2 Homepage

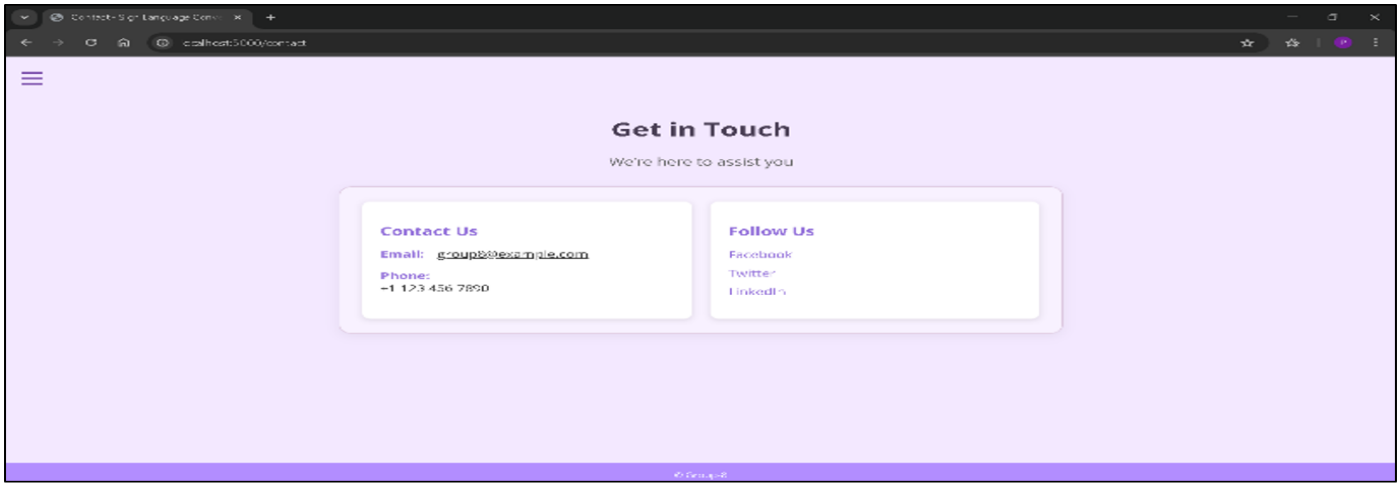

Fig 3 Contact Us Page

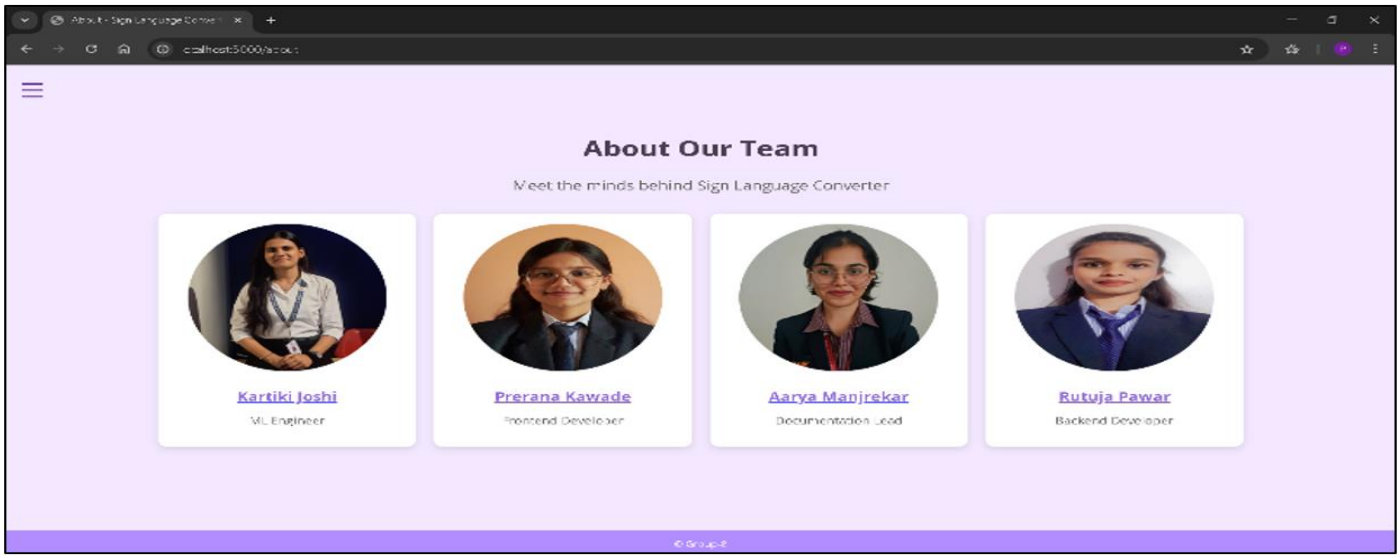

Fig 4 About Us Page

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun285

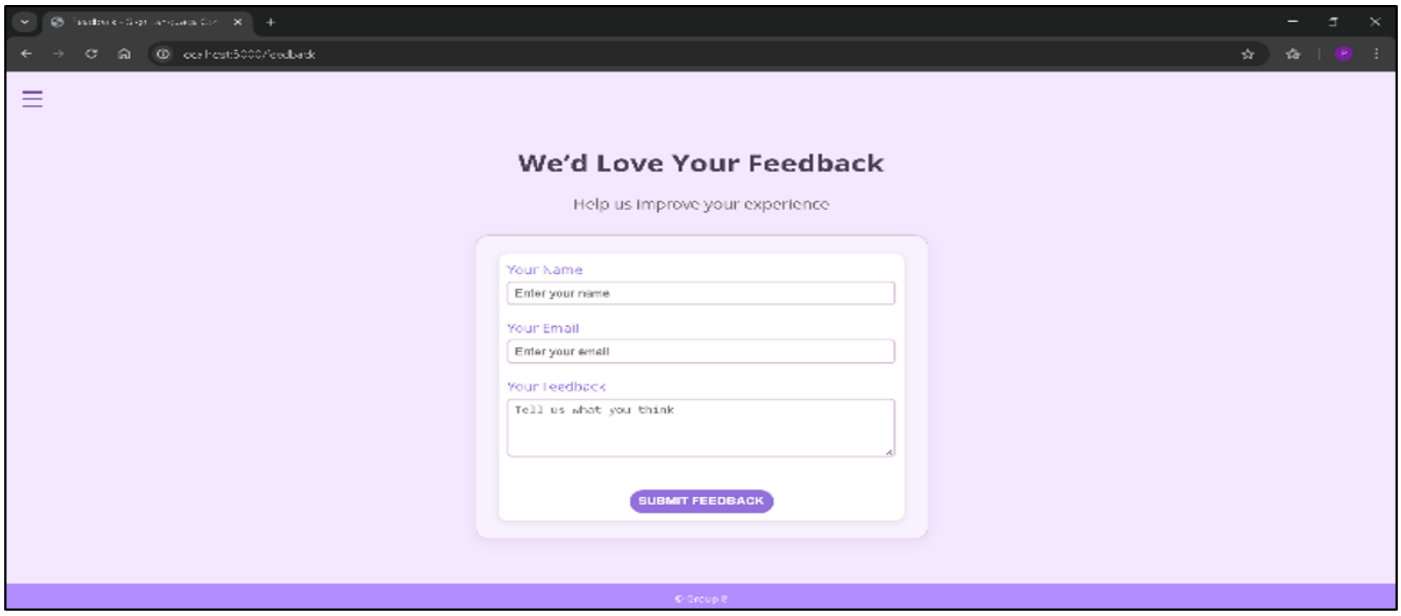

Fig 5 Feedback Page

VIII. CONCLUSION

The development of a camera-based Sign Language to Text and Speech Conversion System marks an important step toward making everyday communication more accessible for people who are deaf or hard of hearing. Communication is a basic human need, and this system makes it easier for people who rely on sign language to interact with others who may not know how to sign. By combining the power of Convolutional Neural Networks (CNN) for accurate hand gesture recognition and Text-to-Speech (TTS) technology for generating clear voice output, this system successfully enables smooth, real-time conversations between sign language users and non-signers Throughout this project, modern technologies such as computer vision, machine learning, and human-computer interaction principles have been brought together to solve a real-world problem. One of the key highlights of this system is its ability to accurately recognize static and dynamic hand gestures in various conditions, like changing lighting, different hand sizes, and diverse backgrounds, using vision-based techniques with MediaPipe for landmark detection. The extracted hand landmark data is processed by a CNN model, which classifies the gestures and converts them into readable text. This text is then converted into audible speech through the TTS module, making conversations more inclusive and efficient.

This project not only demonstrates how Artificial Intelligence (AI) can be practically applied to help society but also highlights how emerging technologies can be used to promote equality and inclusivity in communication. It acts as a starting point for building more advanced and sophisticated assistive tools in the future. The successful implementation of this system lays a solid foundation for future enhancements, such as adding more gestures, supporting different sign languages from around the world, and integrating this technology into mobile applications or wearable devices.

Furthermore, this system proves that with continuous advancements in machine learning algorithms, deep learning frameworks, and real-time image processing tools, it is possible to develop efficient, reliable, and user-friendly solutions that help bridge communication gaps in social, educational, healthcare, and professional environments. In conclusion, this project represents a meaningful contribution toward creating a more inclusive digital society where everyone, regardless of their abilities, can communicate freely and comfortably

FUTURE SCOPE

While the current version of the Sign Language to Text and Speech Conversion System achieves its main objectives, there is a lot of room for future development and improvements. The system has the potential to become even more powerful, versatile, and helpful with several enhancements. Some of the promising areas where this system can grow are discussed below:

- *Expanded Gesture Library:*

At present, the system recognizes a limited set of sign language gestures, primarily focusing on basic alphabets or simple words. In the future, the gesture library can be expanded by integrating a much larger and more diverse dataset. This would allow the system to recognize a broader range of gestures, including dynamic signs, phrases, and complex hand movements. Additionally, incorporating region-specific signs and gestures can make the system adaptable for different communities that use unique local sign languages.

- *Multilingual Support:*

Currently, the system is designed for American Sign Language (ASL). However, there are many different sign languages used around the world, such as British Sign Language (BSL), Indian Sign Language (ISL), and Japanese

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun285

Sign Language (JSL). Future upgrades could involve training the system to recognize gestures from multiple sign languages. This multilingual capability would make the system useful for diverse communities across various countries, thus increasing its global reach and social impact.

- *Mobile Application Development:*

To make the system more portable and easily accessible, a lightweight mobile application can be developed. This mobile app would enable users to access gesture recognition services directly from their smartphones or tablets without needing a computer setup. Features like offline mode, voice output, and gesture learning modules could be added to improve the app's functionality and convenience, especially for on-the-go communication.

- *Bidirectional Communication:*

An exciting future possibility is to convert this one-way translation system into a two-way communication platform. Currently, the system only translates sign language into text and speech. In the future, it could also recognize spoken language and convert it into animated sign language visuals on the screen. This would enable non-signers to respond to deaf individuals, creating a complete, real-time interactive conversation between both parties without needing an interpreter.

- *Integration with Assistive Technologies:*

Combining this system with modern assistive devices, such as smart glasses, Augmented Reality (AR) headsets, or wearable cameras, could enable on-the-go real-time sign recognition and communication assistance in crowded or public places. For example, AR smart glasses could detect hand gestures through the glasses' built-in camera and display translated text or speech output directly in the user's view

- *Cloud-based Processing*

As gesture recognition models become larger and more complex with expanded datasets, it may become difficult to run them efficiently on small devices like smartphones or wearables. Future versions of the system can be developed using cloud-based servers for handling heavy computation tasks like gesture processing, model updating, and data storage. Cloud infrastructure would allow the system to maintain high performance while reducing the load on the local device, ensuring smooth and fast operation without affecting battery life or device performance.

By adopting these enhancements, the system can become a smarter, more flexible, and widely applicable tool for breaking communication barriers and creating an inclusive environment for people with hearing impairments. Continuous research and collaboration between AI developers, healthcare professionals, and the deaf community will be essential to make these future improvements a reality

ACKNOWLEDGMENT

We would like to take this opportunity to thank our Internal Guide Prof. Sweta Wankhade for giving us all the help and guidance we needed. We are grateful to her for her kind support. Her valuable suggestions were very helpful. We are also grateful to Dr. Bhagyashree Dhakulkar, Head of Artificial Intelligence and Data Science Department, ADYPSOE, Lohegaon, Pune for the indispensable support, suggestions, and motivation during the entire course of the project.

We would also be grateful to our Principal Dr. F.B. Sayyad who encouraged us and created a healthy environment for all of us to learn in the best possible way. We also thank all the staff members of our college and technicians for their help in making this project a successful one.

REFERENCES

[1]. A. S. K. Raj and P. S. Babu, "A survey on sign language recognition system for Indian sign language using CNN," Journal of Computational and Theoretical Nanoscience, vol. 16, pp. 3983–3991, 2019.
[2]. A. Z. Choudhury and S. P. Ghosh, "Sign language recognition using convolutional neural networks and MediaPipe for real-time applications," in Proc. IEEE Int. Conf. on Advanced Networks and Telecommunications Systems (ANTS), 2021.
[3]. P. S. Patil, P. S. M. Sharma, and R. K. Chatterjee, "Real-time hand gesture recognition for sign language translation," in Proc. Int. Conf. on Artificial Intelligence and Computer Science (AICS), pp. 213–217, 2020.
[4]. Google Inc., "Google Text-to-Speech (gTTS) Documentation," [Online]. Available: https://pypi.org/project/gTTS/. [Accessed: Feb. 15, 2025].
[5]. H. J. Nguyen and M. B. Y. Chang, "Real-time sign language recognition using MediaPipe framework and deep learning," Int. J. of Computer Vision, vol. 31, no. 6, pp. 524–538, 2020.
[6]. T. M. Soong, "Deep learning for gesture recognition in sign language communication," Journal of Artificial Intelligence in Engineering, vol. 28, pp. 215–225, 2018.
[7]. A. Shalal, "Survey of modern techniques for real-time gesture recognition systems," IEEE Access, vol. 8, pp. 55871–55881, 2020.
[8]. K. L. R. R. Reddy, S. S. Srinivas, and K. C. S. Prasad, "Sign language recognition using CNNs and deep learning," IEEE Access, vol. 8, pp. 117148 117160, 2020.
[9]. A. Z. Choudhury and S. P. Ghosh, "Sign language recognition using convolutional neural networks and MediaPipe for real-time applications," IEEE Int. Conf. on Advanced Networks and Telecommunications Systems (ANTS), 2021. ADYPSOE, Department of Artificial Intelligence and Data Science 2024-25 69
[10]. A. C. R. D. S. A. Rajasekaran, "Sign language recognition using hand gestures with MediaPipe and CNN," Int. J. of Computer Science and Network Security, vol. 21, pp. 241–245, 2021.
[11]. P. M. B. C. E. J. Doe, "Exploring CNN-based approaches to real-time sign language recognition," Int. J. of Computer Vision, vol. 41, no. 7, pp. 891 904, 2019.
[12]. F. H. P. Zhang and J. W. Li, "Convolutional neural networks for sign language recognition with real-time

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun285

processing," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, pp. 35–47, 2021.
[13]. S. A. P. R. U. B. A. Kumar, "Real-time hand gesture recognition using MediaPipe framework and CNNs," Journal of Artificial Intelligence and Computer Vision, vol. 32, pp. 108–120, 2021.
[14]. M. M. A. Singh and J. G. Ghosh, "Deep learning-based sign language recognition system for communication assistance," Int. J. of Applied Artificial Intelligence, vol. 30, pp. 1254–1265, 2020.
[15]. R. Sharma and J. P. W. Wang, "Combining CNNs with MediaPipe for real-time gesture recognition in sign language," in Proc. Int. Conf. on Image Processing and Computer Vision, pp. 225–231, 2020.
[16]. R. H. K. W. Wang, "Real-time translation of sign language to text and speech using machine learning," IEEE Transactions on Human-Machine Systems, vol. 50, no. 8, pp. 743–750, 2020.