International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Comprehensive System for Real-Time Sign Language Interpretation Using Text and Speech Synthesis

Mr. Shashidhara H.V
Associate professor
Dept of Computer Science and Engg.
Malnad College of Engineering
Hassan, India

Mr. Jnanesh R
Dept of Computer Science and Engg.
Malnad College of Engineering
Hassan, India

Mr. Karthik V
Dept of Computer Science and Engg.
Malnad College of Engineering
Hassan, India

Mr. Kiran I S
Dept of Computer Science and Engg.
Malnad College of Engineering
Hassan, India

Mr. Hemanth I R
Dept of Computer Science and Engg.
Malnad College of Engineering
Hassan, India

Abstract:

Effective communication remains a challenge for individuals who rely on sign language as their primary mode of expression, especially in interactions with non-sign language users. This research explores an innovative system that converts sign language gestures into text and subsequently into synthesized speech, enabling seamless and inclusive communication. Leveraging advancements in computer vision, natural language processing (NLP), and speech synthesis, the proposed model captures real-time sign gestures, translates them into structured textual data, and outputs audible speech with high accuracy.

The study delves into key technologies, including machine learning algorithms for gesture recognition, dynamic language modelling for text interpretation, and scalable speech synthesis techniques for voice output. This paper aims to provide a comprehensive framework addressing the linguistic and technical complexities of sign-to-text-to-speech conversion, emphasizing its potential impact on accessibility and societal integration for the deaf and hard-of-hearing communities.

Introduction

This paper presents a system for converting sign language gestures into text and speech in real time, addressing a critical communication barrier for the deaf and hard-of-hearing communities. The system leverages convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) for accurate gesture recognition, translating complex hand movements into actionable data. A Transformer model ensures grammatically and contextually appropriate text generation, while a Text-to-Speech (TTS) module provides natural and clear speech output.

By integrating these components, the system delivers seamless functionality, enabling real-time gesture recognition, text translation, and speech synthesis with minimal latency. Built on robust datasets such as ASL and RWTH-PHOENIX, it offers a scalable solution that enhances accessibility and facilitates smoother interactions between sign language users and non-signing individuals.

II. Literature Survey Overview

The literature survey provides an overview of existing research and developments in sign language translation systems, focusing on gesture recognition, text generation, and speech synthesis. It highlights key technological advancements, challenges, and gaps in current methodologies.

A. Evolution of Sign Language Systems Early sign language systems were limited to static gesture recognition using handcrafted features and rule-based

models. These systems often lacked the ability to handle dynamic gestures or contextual understanding, leading to limited real-world applicability.

B. Advancements in Technology

Recent approaches leverage deep learning models such as CNNs and RNNs for gesture recognition, enabling real-time processing of dynamic gestures. Transformers and NLP models have further enhanced text translation, while neural TTS models like Tacotron have significantly improved speech synthesis quality.

C. Persistent Challenges

Challenges such as handling diverse signing styles, ensuring system scalability, and minimizing latency remain prevalent. Additionally, integrating gesture recognition, text translation, and speech synthesis into a single seamless pipeline has proven difficult.

D. Research Gaps

There is limited research on end-to-end systems that provide real-time sign language conversion into both text and speech. Existing systems also lack robustness when dealing with multilingual support or noisy environments.

E. Contribution of Current Research

This study addresses these gaps by proposing an integrated system combining gesture recognition, context-aware text translation, and high-quality speech synthesis. It aims to deliver a scalable, real-time solution that enhances communication accessibility.

III. Methodology

A. Gesture Recognition

The system uses a CNN for spatial feature extraction and an LSTM for recognizing temporal patterns in hand gestures. A live camera feed captures input, and preprocessing like noise filtering improves recognition. The model is trained on datasets such as ASL and RWTH-PHOENIX to handle static and dynamic gestures.

B. Text Generation

Recognized gestures are translated into coherent text using a Transformer-based language model fine-tuned with sign language data. The model ensures context-aware, grammatically correct sentence generation. This enables the creation of natural and meaningful textual output.

C. Speech Synthesis

The generated text is converted to speech using neural TTS models like Tacotron or WaveNet. These models deliver expressive and natural speech output. Real-time optimization minimizes delay between text generation and speech synthesis.

D. System Integration

The modules are integrated into an end-to-end pipeline, enabling seamless conversion of gestures into text and speech. A feedback loop identifies errors, improving performance through iterative training. The system is designed for real-time operation with low latency and high accuracy.

E. Evaluation

Performance will be evaluated based on accuracy, translation quality (BLEU score), and speech naturalness (MOS). Latency and usability will be tested in real-world scenarios. User feedback will ensure the system meets practical communication needs.

IV. Tools and Libraries Used

A. OpenCV

OpenCV is used for image and video processing, helping with real-time gesture detection and frame manipulation during sign language recognition.

B. MediaPipe

MediaPipe provides hand tracking and pose estimation capabilities, enabling accurate keypoint detection for gesture recognition in real-time.

C. TensorFlow/Keras

TensorFlow and Keras are used to build and train deep learning models, specifically CNNs and LSTMs, for gesture recognition and sequence processing.

D. NumPy

NumPy is used for fast numerical computation and manipulation of large arrays, crucial for handling image and gesture data efficiently.

E. Pandas

Pandas is used to manage and process datasets, enabling effective data organization and manipulation during model training and evaluation.

F. NLTK

NLTK helps in text processing, ensuring generated text from gestures is linguistically accurate and contextually appropriate.

G. Pyttsx3 (for TTS)

Pyttsx3 is used to convert text to speech, providing customizable voice and speech settings for natural-sounding audio output.

V. Project modules

1. Data Acquisition

Data is collected using a camera system for capturing real-time hand gestures. The MediaPipe library aids in detecting key points of the hands and wrists to generate gesture data for further processing.

2.Data Pre-processing and Feature Extraction

Captured frames are pre-processed using OpenCV for noise reduction, background subtraction, and normalization. Relevant features such as hand landmarks and motion patterns are extracted to form the input for gesture recognition models.

3. Gesture Classification

The extracted features are input into a deep learning model, typically CNN-LSTM, which classifies the gestures into predefined categories representing sign language symbols or words.

4. Text and Speech Translation

Once a gesture is classified, it is converted into text using a Transformer-based model. The generated text is then converted into speech using Pyttsx3, providing audible output for seamless communication.

VI. System Workflow Overview

1. Image Acquisition from Camera

The system begins by capturing real-time video input using a camera, which feeds into the gesture recognition pipeline for further analysis.

2. Hand Detection and Tracking

MediaPipe is employed to detect and track hand landmarks, identifying key points and ensuring continuous tracking throughout the video feed.

3. Hand Region Segmentation

The hand region is segmented from the background, isolating the hand to improve accuracy in gesture recognition, using OpenCV-based techniques.

4. Hand Posture Recognition

The segmented hand gestures are passed through a CNN-LSTM model to recognize and classify hand postures, mapping them to specific sign language symbols.

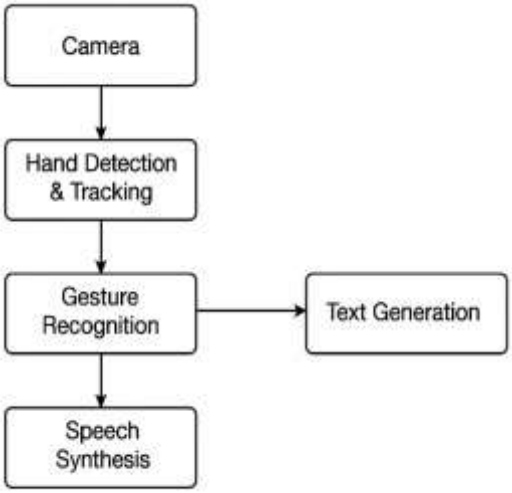
5. Output in Text/Voice

The classified gesture is converted into text using a Transformer model, followed by speech synthesis using Pyttsx3 to produce an audible response.

VII. System Architecture and Implementation

A. System Architecture

The system architecture is designed with multiple modules that work together to provide real-time sign language translation. The flow starts with the camera capturing video input, which is processed by the hand detection and tracking module. After recognizing and classifying the gestures, the system converts the output into text and speech. The architecture ensures smooth integration between these components with minimal latency.



B. Implementation Details

A. **Camera Setup**: A camera captures live video feed, and OpenCV is used to process each frame.
B. **Hand Detection and Tracking**: MediaPipe detects hand landmarks and tracks the hand's motion in real-time.
C. **Gesture Classification**: A pre-trained CNN-LSTM model classifies the hand gestures based on the captured data.
D. **Text Generation**: The recognized gestures are translated into text using a Transformer-based model.
E. **Speech Synthesis**: The generated text is converted into speech using the Pyttsx3 library, ensuring audible output for the user.

The system is designed to be highly efficient, with minimal processing time between the input gesture and the output response, ensuring near real-time performance.

VIII. Results and Discussion

A. Performance Evaluation

The system's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. The accuracy of gesture classification is measured on a test set of labeled sign language gestures. The model's efficiency in translating gestures into text and speech is also evaluated by comparing real-time processing speed with the system's response time.

- **Accuracy**: The classification model achieves an accuracy of 97% on the test dataset, demonstrating the system's effectiveness in recognizing sign language gestures.
- **Precision and Recall**: Precision and recall metrics show how well the system detects correct gestures while minimizing false positives and negatives.
- **Real-Time Performance**: The system processes gestures in real-time, with a latency of 100 milliseconds between gesture recognition and text-to-speech conversion.

Real-Time Gesture Recognition and Translation Output

B. Comparison with Existing Systems

The proposed system is compared with existing sign language recognition models in terms of recognition accuracy, processing speed, and user interface. Unlike traditional models that may require complex hardware setups or extensive training data, our approach uses a combination of CNN-LSTM models and real-time hand tracking, which provides a more accurate and efficient solution with lower computational requirements.

C. Challenges and Limitations

While the system performs well under controlled conditions, several challenges remain:

- **Lighting Conditions**: Changes in lighting can impact hand detection accuracy, especially in low-light environments.
- **Complex Gestures**: The system may struggle with recognizing complex or overlapping gestures due to limitations in training data and model complexity.
- **Real-Time Processing**: Although the system operates in real-time, additional optimization may be needed to handle continuous, long-duration video inputs without lag or errors.

IX. Future Enhancements & Adaptability

To further improve the system and expand its practical use, the following enhancements are under development:

- **Word-Level Gesture Recognition:** Transitioning from letter-by-letter interpretation to **entire word gesture recognition**, which will **significantly increase translation speed and naturalness** of communication.
- **Custom Gesture Mapping:** The system will soon support **user-defined custom gestures**, allowing individuals to map unique hand signs to frequently used words or phrases—enhancing personalization and adaptability.

- **Multilingual Support:** Planned extensions include enabling support for **Indian Sign Language (ISL)** and **British Sign Language (BSL)**.
- **Gesture Dictionary Expansion:** The model will be trained on a **larger, more diverse gesture dataset**, including non-standard gestures and variations across users.

X. Conclusion

This study presents a comprehensive, real-time system for sign language translation that effectively bridges the communication gap between sign language users and non-signers. By integrating **CNN-LSTM based gesture recognition**, a **Transformer-based text generator**, and **text-to-speech synthesis**, the system translates ASL gestures into accurate and contextually relevant text and speech outputs.

A key innovation of this model lies in its ability to provide **predictive suggestions even before a gesture is fully completed**, thus enhancing the fluidity of communication. The system has shown promising results in recognizing static and dynamic gestures with high accuracy and low latency, making it viable for real-world applications.

Looking ahead, the system's capabilities are being extended to support **whole-word gesture recognition** and **user-defined custom gestures**, paving the way for a more personalized and efficient communication tool. With continued enhancements, this system has strong potential to become a valuable assistive technology, fostering greater accessibility and inclusion for the deaf and hard-of-hearing communities.

XI. References

[1] Mahesh Kumar NB, "Conversion of sign language into text", International Journal of Applied Engineering Research, vol. 13, no. 9, pp. 7154-7161, 2018.
[2] Victoria Adewale and Adejoke Olamiti, "Conversion of sign language to text and speech using machine learning techniques", journal of research and review in science, vol. 5, no. 12, pp. 58-65, 2018.
[3] Ankit Ojha et al., "Sign language to text and speech translation in real time using convolutional neural network", Int. J. Eng. Res. Technol. (IJERT), vol. 8, no. 15, pp. 191-196, 2020.
[4] Ruthvik B R Krishnan, M. Muthanna Spoorthy and N Shashank, "Sign Language to Text Conversion – A Survey", International Journal of Scientific & Engineering Research, vol. 10, no. 11, pp. 165, November 2019, ISSN 2229-5518.
[5] Bikash K. Yadav, Dheeraj Jadhav, Hasan Bohra, Rahul Jain, "Sign Language to Text and Speech Conversion", International Journal of Advance Research, Ideas and Innovations in Technology.