A Deep Learning-Based Real-Time Sign Language Translation System for Telehealth Applications

Gunarathna L.P.N, Somarathna S.V.A.P.K, Mukunthan T.

Department of Electrical and Electronic Engineering, University of Jaffna

ABSTRACT

Effective communication in healthcare is often limited for individuals with hearing and speech impairments. This research presents a real-time sign language translation system that leverages deep learning and computer vision to enhance accessibility in telehealth environments. Spatial-temporal features were extracted using an Inflated 3D ConvNet (I3D), and three sequence-to-sequence models Encoder–T5–Decoder, Encoder–LSTM–Decoder, and Encoder–Transformer–Decoder were evaluated. The Encoder–Transformer–Decoder achieved the best performance with a BLEU score of 0.0478 , outperforming recurrent models. Real-time interaction is enabled through a socket based communication module for seamless peer-to-peer communication. The prototype demonstrates strong potential to promote inclusive and accessible communication in healthcare and public service applications.

OBJECTIVES

This research aims to develop a real-time sign language translation system for telehealth settings, with the following key objectives,

- Promote inclusivity by enabling smooth interactions between individuals with hearing impairments and the broader community.
- Design an accurate translator that converts sign language gestures into English text in real time to bridge communication gaps.
- Ensure translation accuracy across diverse sign variations to meet the needs of a wide user base.
- Enhance system robustness to perform reliably under challenging conditions such as poor lighting, low-resolution input, and complex backgrounds.
- Achieve real-time performance to support seamless communication in fast-paced environments like hospitals and telehealth consultations.
- Address the gap in medical sign language research by building a specialized system to aid healthcare professionals in communicating with patients who have hearing or speech impairments.

METHODOLOGY

This study develops a real-time sign language translation framework for telehealth applications, designed to facilitate communication for individuals with hearing and speech impairments. The system follows a multi-stage pipeline, as depicted in Figure 1.

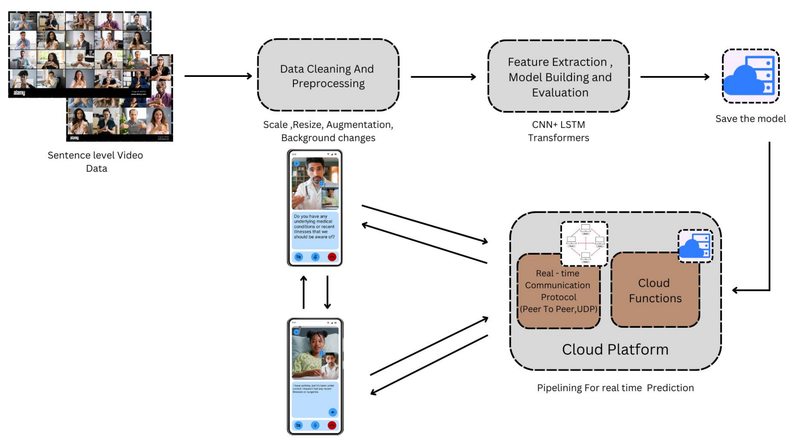


Figure 1 : Overall System Architecture

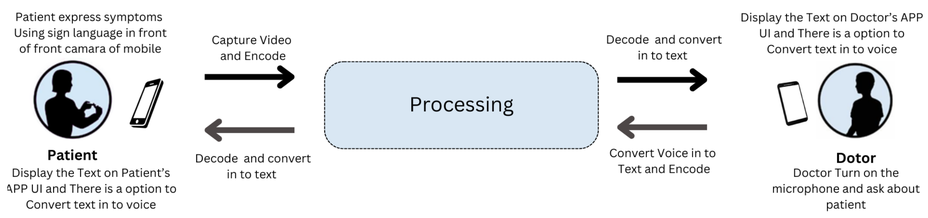


Figure 2 : System Overview

Data Acquisition & Preprocessing: The How2Sign dataset was selected for its rich sentence-level ASL content. Preprocessing involved background removal using OpenCV, skeleton key point extraction via MediaPipe Holistic , and motion analysis through optical flow (TV-L1 algorithm).

Feature Extraction: Spatial-temporal features were extracted using the Inflated 3D ConvNet (I3D) model, which inflates 2D CNN filters into 3D to capture gesture dynamics.

Sequence Modeling: Four deep learning architectures were evaluated: GRU, LSTM, Transformer, and a Encoder LSTM + Transformer model. The Encoder-LSTM-Decoder achieved the best performance based on BLEU scores.

Performance Evaluation: The system was assessed using BLEU Score (for translation accuracy) and Word Error Rate (WER) (for sentence correctness).

Real-Time Mobile Deployment: The model is integrated into a user-friendly mobile app with WebRTC-based peer-to-peer communication. The app supports both edge and cloud inference and includes speech-to text/text-to-speech modules to enhance accessibility

RESULTS

Our system was trained using the How2Sign dataset with over 30,000 sentence-level videos. The I3D feature extraction pipeline was validated against the original dataset, confirming its accuracy. Initial experiments using LSTM and Transformer sequence models showed that LSTM better captured temporal dependencies. We then evaluated three encoder-decoder architectures using BLEU scores.

Table 01 – Results Comparison

Model	Train	Validation	Test
Encoder + T5 Decoder	0.0137	0.0133	0.0129
Encoder + LSTM Layer + Decoder	0.0388	0.0299	0.0312
Encoder + Transformer Layer + Decoder	0.0689	0.0459	0.0478

After training all models, we also obtained the training and validation accuracy and loss curves. These plots provided deeper insights into how each model learned over time.

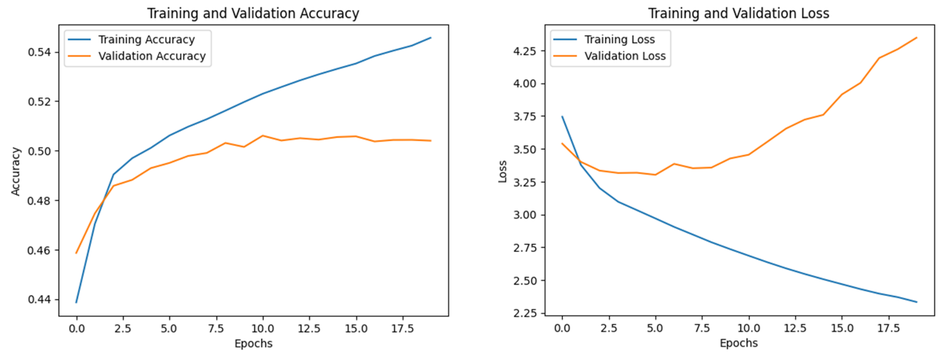


Figure 3 : Training and Validation Accuracy Comparison for LSTM model

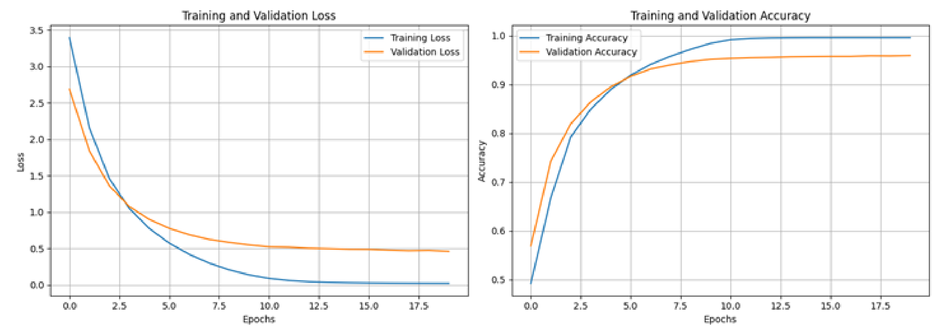
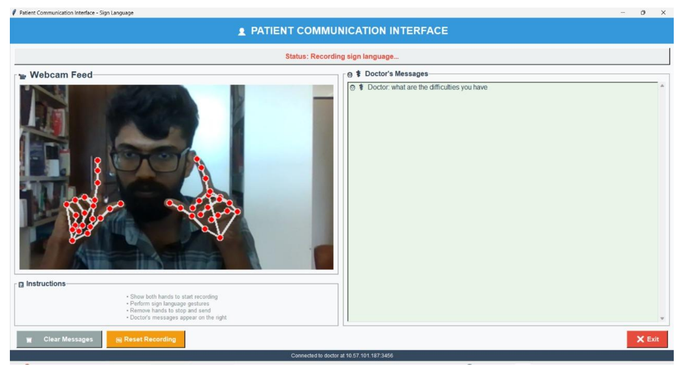


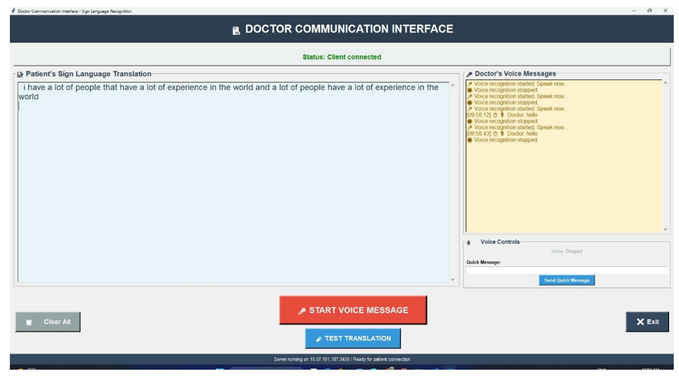
Figure 4 : Training and Validation Accuracy Comparison for Transformer model

Among these, the Encoder + LSTM + Decoder model achieved the highest BLEU scores across all splits To validate real-world applicability, we implemented a real-time two-way communication system using socket programming.

Patient-Side: Captures sign language via webcam and streams it in real time.



Doctor-Side: Receives video, extracts I3D features, processes using the trained LSTM model, and displays translated text.



CONCLUSION

This research presents a real-time sign language translation system combining I3D-based feature extraction with advanced sequence models . Among all, the Encoder + LSTM + Decoder architecture showed the best performance. Though the system achieved promising results, limitations like the absence of a Sri Lankan Sign Language dataset and limited training resources were identified. Future work will focus on developing native datasets, enhancing model training, and deploying the system in real-world healthcare settings to improve accessibility and communication for hearing-impaired individuals.

REFERENCE

1. M. S. Astriani and M. A. R. C. E. L. L. Alvianto, "Telemedicine sign language classification for COVID-19 patients with disability based on LSTM model.," 2023.
2. P. Alvarez , X. Giro , N. Laia and T. Benet , "Sign Language Translation based on Transformers for the How2Sign Dataset.," Image Processing Group Signal Theory and Communications Department Universitat Politècnica de Catalunya. BARCELONATECH, 2022.
3. How2Sign Dataset. (n.d.). Retrieved from https://how2sign.github.io/
4. I. Godage , "Sign Language Recognition for Sentence Level Continuous Signings," Doctoral dissertation, 2021.
5. Y. Zhao , . X. Zhang , R. Hu , J. Xue , X. Li , . L. Che , R. Hu and . L. Schopp , "An automatic captioning system for telemedicine.," In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. vol. 1, pp. pp. I-I, 2006.

Faculty of Engineering, University of Jaffna,
Ariviyar Nagar,
Kilinochchi.