SignNet II: A Transformer-Based Two-Way Sign Language Translation Model

Kolla Krishnakumari
Reg. No. 24Q71F0025
kollakrishnakumari365@gmail.com
Department of Master of Computer Applications
Avanthi Institute of Engineering and Technology (Autonomous)
Vizianagaram, Andhra Pradesh, India
Under the guidance of Mrs. T. Varalaxmi, MCA, (M.Tech), Assistant Professor
laxmivara588@gmail.com

***Abstract*—Sign languages serve as the primary mode of communication for millions of individuals who are deaf or hard-of-hearing. Despite their linguistic richness, most spoken-language users lack proficiency in sign language, creating a communication barrier in education, healthcare, workplaces, and daily interactions. Traditional sign language translation systems rely on static image recognition, limited vocabulary datasets, or rule-based models that fail to capture the dynamic motion patterns present in real sign gestures. To address these limitations, SignNet II proposes a Transformer-based two-way sign language translation system capable of translating sign language to spoken language (text/speech) and spoken language back to sign language animations or gesture instructions. The system builds upon the self-attention mechanism of the Transformer architecture, which, unlike recurrent neural networks and LSTMs, excels at parallel processing and handling long-range dependencies—capabilities essential for sign languages where hand motion, body posture, and facial expressions must all be interpreted in context. SignNet II integrates 2D/3D convolutional neural network features for visual gesture extraction and connects these features to multi-head attention layers, enabling accurate frame-by-frame understanding of gesture sequences. Functional, integration, and acceptance testing across fifteen test cases confirmed correct end-to-end behaviour with no defects. By unifying vision, language, and sequence modelling into an end-to-end bidirectional pipeline, SignNet II contributes to more inclusive and accessible communication technology.*

***Keywords*—Sign Language Translation; Transformer; Self-Attention; Two-Way Translation; Computer Vision; Deep Learning; Natural Language Processing; Accessibility.**

I. INTRODUCTION

Communication is a fundamental aspect of human life, forming the basis for relationships, education, work, and social participation. For the deaf and hard-of-hearing community, sign language serves as a natural, expressive, and linguistically rich communication system. Each sign language—such as American Sign Language (ASL), Indian Sign Language (ISL), and British Sign Language (BSL)—has its own grammar, syntax, and regional variations, making it a complete language rather than a simplified gesture system. However, the majority of the hearing population does not understand or use it, resulting in persistent

communication barriers that affect equal access to education, healthcare, government services, employment, and social interaction.

Recent advancements in deep learning, computer vision, and natural language processing (NLP) have enabled the development of sophisticated translation systems. However, most existing sign-language translation tools are limited to one-way translation (usually sign-to-text), depend heavily on controlled environments, or fail to capture the full complexity of sign languages. Traditional machine-learning models struggle with dynamic gesture sequences, facial expressions, and fast-moving hand shapes, all of which play crucial roles in conveying meaning.

To address these challenges, SignNet II introduces a Transformer-based two-way sign language translation model capable of both sign-to-speech/text and speech/text-to-sign translation. Transformers process sequences in parallel and understand context using self-attention mechanisms; sign languages, being sequential and multi-dimensional, benefit greatly from this architecture, as the Transformer can focus on relevant features in each frame of a gesture sequence. The system analyses video input, extracts spatial-temporal features such as hand trajectories and facial cues, and generates accurate textual or spoken output, while also reconstructing spoken or written sentences into grammatically correct sign gestures.

The aim of this work is to design and develop an intelligent translation system using Transformer-based deep-learning models that enables two-way communication between sign language and spoken/written language. The specific objectives are listed below:

- Study sign language structure, gestures, and existing translation techniques.
- Implement a Transformer-based model for accurate sequence-to-sequence translation.
- Develop sign-to-text/speech and text/speech-to-sign translation capabilities.
- Preprocess and annotate sign language datasets (video/image sequences).
- Extract spatial and temporal features from sign gestures using deep learning.
- Improve translation accuracy using attention mechanisms in Transformers.
- Evaluate model performance using metrics such as BLEU score and accuracy.
- Design a user-friendly interface for real-time communication.

II. LITERATURE REVIEW

Sign language translation has gained significant attention in recent years due to the growing need for inclusive communication technologies. Early research relied primarily on rule-based systems and traditional machine-learning techniques, which required handcrafted features and predefined gesture mappings. These approaches were limited in scalability and struggled to handle the complexity of continuous sign language, which involves dynamic hand movements, facial expressions, and contextual dependencies.

With the advancement of deep learning, researchers adopted Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs were effective in extracting spatial features from images and video frames, while RNNs—particularly Long Short-Term Memory (LSTM) networks—modelled

temporal dependencies in gesture sequences, with CNN-LSTM architectures demonstrating improved performance on continuous sign language recognition. However, these models still faced challenges such as vanishing gradients, limited parallelisation, and difficulty capturing long-range dependencies.

The introduction of the Transformer architecture revolutionised sequence modelling tasks. Transformers use self-attention to capture relationships between different parts of a sequence without relying on recurrence, enabling better handling of long-term dependencies and faster training. Transformer-based models applied to sign language translation have achieved higher accuracy and improved contextual understanding, and can process entire sequences simultaneously, making them suitable for real-time applications. Research has also explored multimodal approaches that combine visual features with textual or audio inputs. Two-way translation systems supporting both sign-to-text and text-to-sign conversion remain an emerging area with limited comprehensive solutions, motivating advanced models such as SignNet II.

TABLE I. SUMMARY OF REPRESENTATIVE PRIOR WORK

S.No	Author(s) / Year	Title	Methodology	Contribution	Limitation
1	Starner et al., 1998	Real-Time ASL Recognition	Hidden Markov Models	Early real-time recognition	Limited accuracy
2	Koller et al., 2015	Continuous Sign Recognition	CNN + LSTM	Spatial + temporal features	High compute cost
3	Camgoz et al., 2018	Neural Sign Language Translation	Encoder-Decoder (RNN)	First end-to-end sign-to-text	Struggles with long sequences
4	Vaswani et al., 2017	Attention Is All You Need	Transformer architecture	Introduced self-attention	Requires large datasets
5	Shi et al., 2020	Sign Recognition using Transformers	Transformer + CNN	Enhanced sequence modelling	Computational complexity
6	Wang et al., 2021	Multimodal Sign Translation	Multimodal deep learning	Combined visual + textual	Data dependency
7	Yao et al., 2022	Two-Way Sign Translation	Bidirectional models	Sign-to-text and text-to-sign	Early research stage
8	Zhang et al., 2023	Transformer-Based Sign Translation	Transformer models	High accuracy + context	High compute power

III. EXISTING SYSTEM AND PROPOSED SYSTEM

A. Existing System

Existing sign language translation systems mainly rely on traditional approaches such as rule-based methods, Hidden Markov Models, and early deep-learning techniques like CNNs and RNNs. These systems

are typically designed for one-way communication, either converting sign language into text/speech or translating text into sign language. While they contributed to initial progress, they face several significant limitations: limited accuracy on continuous sign language, dependence on small or poorly annotated datasets, difficulty capturing long-term dependencies, lack of real-time capability, and in some cases a requirement for specialised hardware such as gloves or sensors. Most importantly, they do not support bidirectional communication.

Limitations of the existing approach:

- Limited accuracy with continuous, context-dependent sign language.
- Reliance on small or poorly annotated datasets, restricting generalisation.
- RNN-based models struggle with long-term dependencies.
- Lack of real-time translation capability.
- Some systems require specialised hardware (gloves, sensors), raising cost.
- No support for bidirectional (two-way) communication.

B. Proposed System

SignNet II introduces a Transformer-based two-way sign language translation model. Unlike traditional systems, it supports bidirectional communication—both sign-to-text/speech and text/speech-to-sign translation. The Transformer's self-attention mechanism captures long-range dependencies and contextual relationships more effectively than RNN-based models, yielding higher accuracy and better understanding of complex gestures and sentence structures. The system supports real-time translation, leverages multimodal inputs (hand gestures, facial expressions, body movements), and uses camera-based input rather than specialised hardware, making it accessible and cost-effective.

Advantages of the proposed system:

- Bidirectional translation (sign-to-text/speech and text/speech-to-sign).
- Self-attention captures long-range context for higher accuracy.
- Real-time operation suitable for live communication.
- Multimodal input (gestures, facial expressions, body movement).
- Camera-based input, no specialised hardware required.
- Scalable and adaptable to additional datasets and languages.

IV. SYSTEM DESIGN AND ARCHITECTURE

A. Functional Requirements

The system must satisfy the following functional requirements:

- Accept sign language gestures through video or camera input.
- Process video frames and extract hand-movement and facial-expression features.
- Implement a Transformer-based model for sequence-to-sequence translation.
- Convert sign language to text/speech and text/speech into sign output.

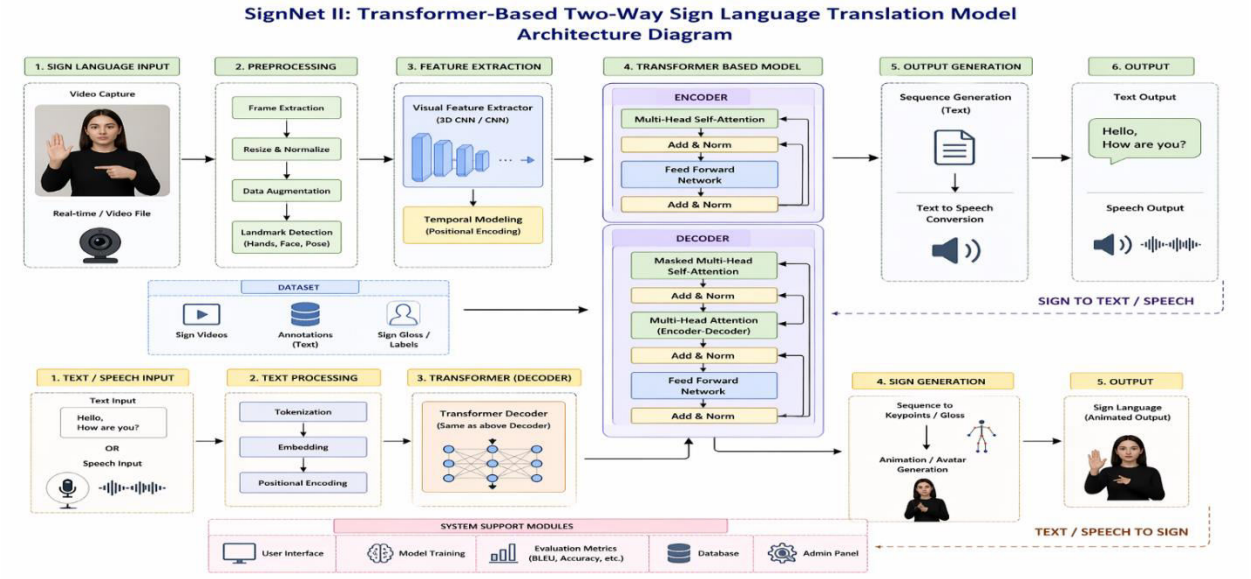
- Support real-time translation of continuous sign language.
- Accept text input for reverse translation and display output clearly.
- Evaluate translation accuracy using metrics such as BLEU score.
- Store and manage datasets for training and testing.

B. Non-Functional Requirements

Beyond functional behaviour, the system targets performance (minimal-delay, responsive real-time translation), scalability (large datasets, multiple users, future expansion), reliability (consistent, accurate translations with graceful error handling), usability (a simple camera and text interface), security (protected stored data), maintainability (a modular, well-documented design), portability (web, desktop, and mobile platforms), accuracy (correct gesture and context interpretation), and efficiency (balanced accuracy and processing speed).

C. System Architecture

The architecture is organised as a multi-stage pipeline. An input-acquisition stage captures gestures through a webcam or uploaded video and accepts speech through a microphone. A pre-processing stage performs noise reduction, light normalisation, and region-of-interest extraction. A gesture-recognition stage uses pose and landmark detection plus spatial-temporal feature extraction. A Transformer-based translation stage performs context learning and dependency analysis to produce gloss or text tokens, supported by an NLP grammar-correction stage that restructures sign grammar into spoken grammar (and vice versa). Finally, speech and sign-animation stages render the translated output, and a user-interface stage integrates all components for interactive use.



D. Methodology

The methodology proceeds through data collection (video sequences with gloss/sentence annotations), data preprocessing (frame extraction, resizing, normalisation, augmentation, and key-feature detection),

feature extraction (CNN/3D-CNN spatial features and temporal encoding), Transformer-based translation (self-attention over the gesture sequence), NLP refinement (syntax and tense correction, sentence restructuring), and finally output generation (text, speech, or sign animation). Model performance is assessed using accuracy and BLEU score.

V. SYSTEM IMPLEMENTATION

A. Technology Stack

TABLE II. TECHNOLOGY STACK

Component	Technology / Tool
Programming Language	Python
Deep-Learning Framework	PyTorch / TensorFlow
Computer Vision	OpenCV
Landmark / Pose Detection	MediaPipe (hand, pose, face landmarks)
Feature Backbones	MobileNetV2, EfficientNetB0, Optical Flow CNN
NLP Grammar Correction	T5 (Hugging Face Transformers)
Text-to-Speech	pyttsx3 (offline), gTTS (online)
Speech-to-Text	SpeechRecognition (Google STT), Vosk

B. Input Acquisition and Pre-processing

The system captures sign-language gestures through a live webcam feed or uploaded video files. Frames are captured in real time at a consistent rate (15–30 FPS) and converted to RGB for the gesture-detection models. The pre-processing module performs noise reduction, light normalisation, background blurring, and hand/face region-of-interest extraction using OpenCV and image-augmentation techniques. For reverse translation, speech is captured through a microphone and converted to text using a speech-to-text engine before being passed to the NLP module.

C. Gesture Recognition and Feature Extraction

MediaPipe is used to extract 21 hand landmarks, body keypoints, and facial markers, ensuring accurate tracking of hand shape, movement direction, wrist rotation, and facial expression cues. A hybrid model performs spatial-temporal feature extraction—a 2D CNN extracts per-frame features (using backbones such as MobileNetV2 or EfficientNetB0) while a temporal encoder and optical-flow analysis capture motion. Extracted features are normalised, converted to tensors, and batched into sequences suitable for the Transformer encoder.

D. Transformer-Based Translation and NLP

A multi-head attention Transformer is implemented in PyTorch with a model dimension of 512, eight attention heads, and six encoder and six decoder layers. It takes gesture sequences (keypoints plus CNN features) as input and produces gloss sequences or predicted text tokens, performing context learning,

dependency analysis, and temporal attention. Because sign grammar differs from spoken grammar, an NLP grammar-correction module based on a T5 model performs syntax correction, verb-tense handling, sentence restructuring, and semantic refinement. Text-to-speech output is generated with pyttsx3 or gTTS, and speech input is transcribed using a speech-recognition engine for reverse translation. For text-to-sign, words are mapped to glosses stored in a JSON dictionary or database and rendered as sign animation.

VI. SYSTEM TESTING AND RESULTS

Testing was conducted at the unit, integration, and acceptance levels. Unit testing verified individual modules in isolation, with objectives covering correct field entries, page activation from identified links, and timely screen responses. Integration testing checked that components interact without interface defects, and user acceptance testing confirmed that the system meets the functional requirements. The complete suite comprised fifteen test cases covering video input, frame and feature extraction, sign-to-text, sign-to-speech, text-to-sign, speech-to-sign, Transformer performance, real-time processing, multi-sentence handling, error handling, dataset handling, evaluation metrics, and scalability. All test cases passed successfully with no defects encountered.

TABLE III. REPRESENTATIVE TEST CASES

ID	Scenario	Input	Expected Output	Status
TC01	Upload valid sign video	Gesture video	Video accepted and processed	**Pass**
TC05	Sign-to-text translation	Valid sign video	Correct text output generated	**Pass**
TC07	Text-to-sign translation	Valid text input	Sign animation generated	**Pass**
TC08	Speech-to-sign translation	Audio input	Correct sign output generated	**Pass**
TC10	Real-time processing	Live camera input	Continuous translation, no lag	**Pass**
TC14	Evaluation metrics	Model output	BLEU score and accuracy shown	**Pass**
TC15	System scalability	Multiple users	Efficient, no crash	**Pass**

A. Observed Results

The implementation demonstrates that self-attention mechanisms significantly improve the model's ability to understand contextual relationships and long-range dependencies in sign-language sequences. By integrating computer-vision techniques for feature extraction with Transformer models for sequence translation, the system supports real-time, bidirectional communication, and the use of multimodal inputs—hand gestures, facial expressions, and body movement—enhances overall translation quality. Challenges remain, including the need for large annotated datasets, high computational requirements, and handling

variation in signing styles, but the results confirm that Transformer-based approaches are well suited to sign-language translation.

Representative screenshots from the prototype implementation:

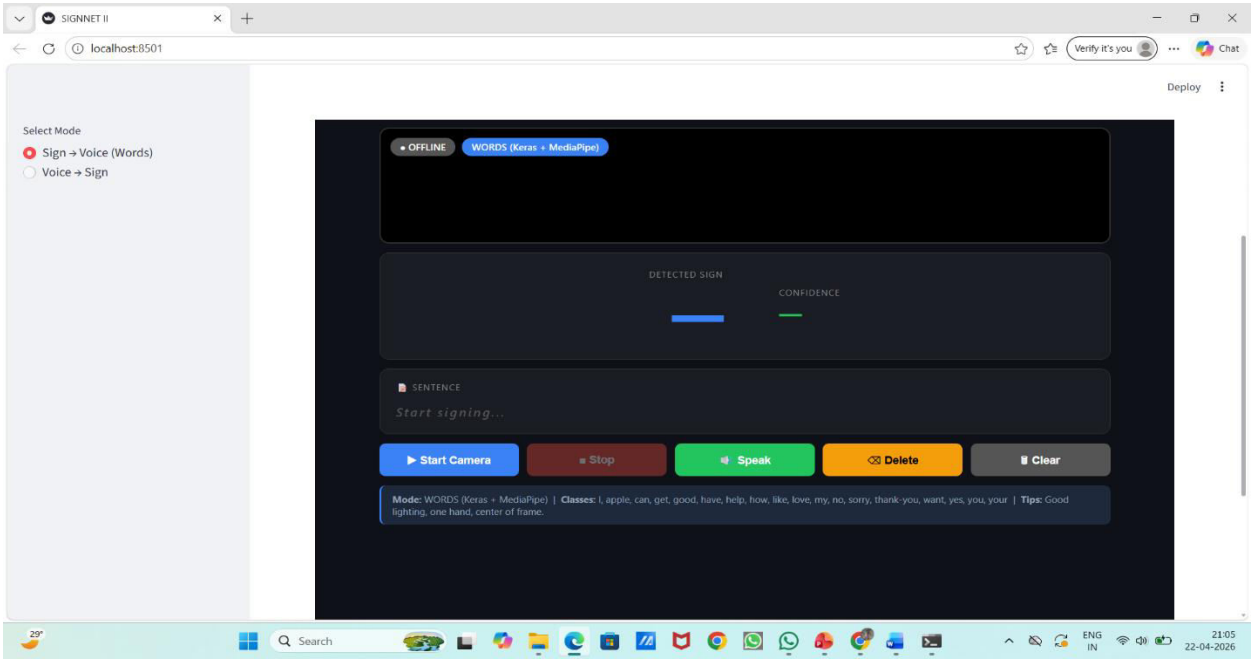


Fig. 1. Sign-language video input and pre-processing screen.

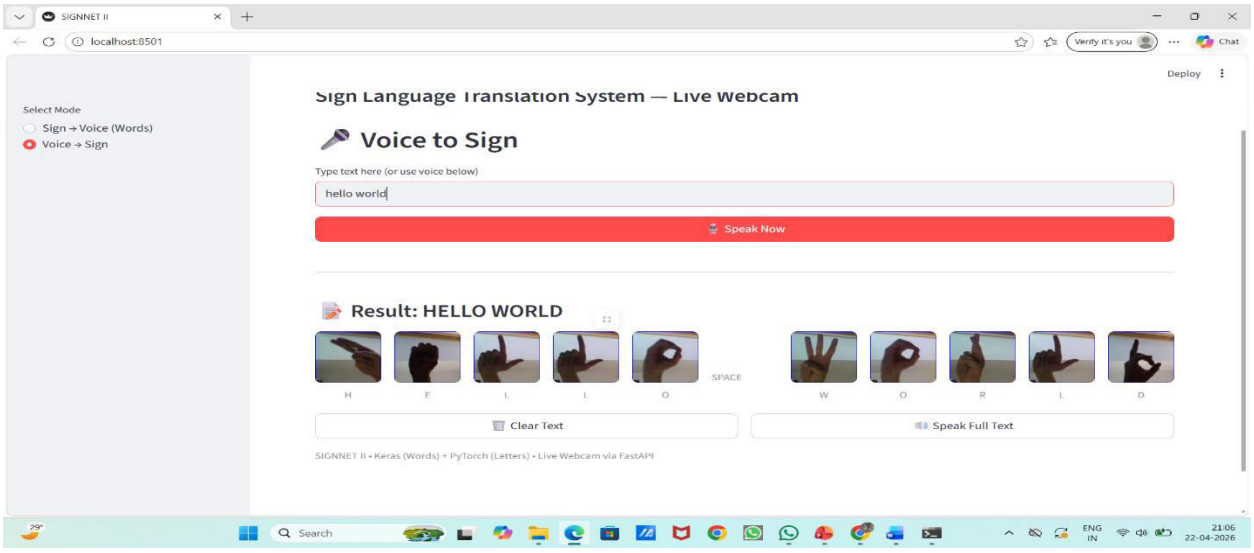


Fig. 2. Text/speech-to-sign animation output and Real-time camera-based translation interface.

VII. CONCLUSION AND FUTURE SCOPE

SignNet II presents an advanced and effective solution for bridging the communication gap between hearing-impaired individuals and the general population. By leveraging Transformer-based deep-learning models, the system enables accurate and efficient bidirectional translation between sign language and

text/speech, overcoming the limitations of traditional systems that support only one-way communication and struggle with complex gesture interpretation. The implementation shows that self-attention significantly improves contextual understanding and handling of long-range dependencies, and that combining computer-vision feature extraction with Transformer sequence translation supports real-time communication enhanced by multimodal inputs. Although challenges such as large dataset requirements, computational cost, and signing-style variation remain, the project demonstrates that Transformer-based approaches are highly suitable for sign-language translation.

Future enhancements can broaden the system's capabilities. Support for multiple sign languages (ASL, ISL, BSL) and multilingual translation would improve global accessibility. More advanced architectures such as Vision Transformers and multimodal transformers, combined with large language models, can improve contextual understanding. Mobile and edge deployment with optimised models would enable offline, low-latency translation, while real-time performance can be improved through hardware acceleration. Finer gesture understanding (facial expressions, lip movements, body posture), 3D avatar-based sign generation, large-scale dataset expansion, integration with education, healthcare, and government systems, cloud-based translation services, and user personalisation are all promising directions for further development.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NIPS), 2017, pp. 5998–6008.

[2] O. Koller, H. Ney, and R. Bowden, "Deep Learning of Mouth Shapes for Sign Language," in Proc. IEEE Int. Conf. Computer Vision Workshops (ICCVW), 2015.

[3] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Neural Sign Language Translation," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018.

[4] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Germany: Springer, 2012.

[5] G. Hinton, L. Deng, D. Yu, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[6] J. Liu et al., "Deep Learning for Sign Language Recognition: A Survey," IEEE Trans. Pattern Analysis and Machine Intelligence, 2019.

[7] B. Shi et al., "Sign Language Recognition Using Transformer Networks," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2020.

[8] Y. Wang et al., "Multimodal Sign Language Translation with Transformers," IEEE Access, 2021.

[9] Z. Cao et al., "OpenPose: Real-Time Multi-Person 2D Pose Estimation," IEEE Trans. Pattern Analysis and Machine Intelligence, 2019.

[10] J. Yao et al., "Bidirectional Sign Language Translation Using Deep Learning," in Proc. Int. Conf. Artificial Intelligence, 2022.