WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Vaani-Setu: A Hybrid Bidirectional Model using Media-Pipe and Bi-LSTM

Adarsh Kumar*, Aman Kumar, Aman Pratap, Ajeet Soni, Deepsikha Sharma

Department of Computer Science and engineering Oriental Institute of Science and Technology Bhopal, Madhya Pradesh, 462022
*Corresponding Author

ABSTRACT—

Individuals with sensory and speech impairments often face communication challenges because they rely on different expressive modalities, including sign language, text, and voice. The lack of compatibility among these modes creates significant barriers during daily interactions in educational, workplace, medical, and public environments. This paper intro- duces Vaani-Setu, an assistive communication platform designed to bridge these gaps by integrating sign-language recognition, speech-to-text, text-to-speech, and text-to-sign rendering into a unified, real-time translation system. The platform enables seamless communication among deaf, blind, and speech-impaired individuals by converting one modality of communication into another without loss of meaning. Vaani-Setu leverages deep- learning–based gesture recognition, speech processing pipelines, natural-language interpretation, and a modular system architecture to ensure high adaptability across various use cases. The implementation builds upon the publicly available a11y-bridge repository. We demonstrate that our Tree Structure Skeleton Image (TSSI) based approach combined with a hybrid CNN- Bi-LSTM architecture achieves a Top-1 sign accuracy of 83.14% and a Speech WER of 9.3%, offering a computationally efficient alternative to complex 3D-CNN models.

***Index Terms*—**Assistive Technology, Sign Language Recognition (SLR), Bi-LSTM, Media-Pipe, Human Action Recognition (HAR), Multimodal Communication.

Introduction

In recent years, rapid progress in assistive communication technologies has significantly improved human–computer interaction for individuals with sensory and speech impairments. Among these technologies, Sign Language Recognition (SLR), Speech Processing, and Text-Based Communication have played a crucial role in enhancing digital accessibility in education, healthcare, public services, and remote communication [1].

Despite these advancements, communication between deaf, blind, and speech-impaired individuals remains a major challenge due to their reliance on different and often incompatible communication modalities. Deaf individuals typically communicate using sign language, while blind individuals depend primarily on voice-based communication, and speech-impaired individuals frequently use text or gesture-based expressions [10]. This lack of a shared communication modality creates barriers, resulting in reduced social inclusion and limited access to essential resources.

Therefore, developing a unified communication framework that supports real-time conversion among sign, speech, and text is essential to ensure equitable participation for all users. Recent research in deep learning (DL) and Human Action Recognition (HAR) has shown promising results in sign language recognition, especially with the use of 3D-CNNs, Transformer models, and Graph Neural Networks (GNNs) for gesture, body-pose, and skeletal feature extraction [3]. However, such architectures often require large computational resources, making them unsuitable for real-time deployment on edge devices where privacy-sensitive communication can- not rely on cloud processing.

To address these limitations, lightweight approaches using skeletal pose extraction have been proposed, demonstrating robustness to variations in background, illumination, and signer appearance. Modern frameworks such as Media-Pipe facilitate fast estimation of facial, hand, and body key-points, enabling efficient multimodal communication systems [2].

In this work, we present **Vaani-Setu**, a unified assistive communication platform that enables seamless interaction between blind, deaf, and speech-impaired individuals. The system integrates sign recognition, speech transcription, and text/voice rendering in real time. Unlike conventional single-modality assistive tools, Vaani Setu functions as a bidirectional communication bridge, allowing users to convert:

- Sign → Text / Speech
- Speech → Text / Sign
- Text → Speech / Sign
-

The primary contributions of this work can be summarized as follows:

1. We propose a unified assistive communication frame- work that supports real-time multimodal translation among sign, speech, and text for inclusive interaction [1].

2. We integrate Media-Pipe-based skeletal extraction with lightweight neural models (CNN-Bi-LSTM) to enable on-device sign interpretation without high computational overhead or cloud dependencies [4].
3. We demonstrate the applicability of Vaani-Setu in class- rooms, hospitals, and workplaces, highlighting its potential to reduce communication barriers in everyday interactions.

The implementation code is available at: https://github.com/mishraadarsh27/a11y-bridge

Related Work

Existing solutions in this domain are primarily categorized into three areas: skeleton-based SLR using deep neural net- works, speech-based accessibility systems, and multimodal translation frameworks.

A. Skeleton-Based Sign Language Recognition

Early skeleton-based works in Human Action Recognition (HAR) demonstrated that pose landmarks extracted from videos can be represented as structured images and processed with convolutional neural networks (CNNs). In [3], one of the first CNN-based models, the joints of a human body were divided into semantic groups (e.g., left arm, right arm, trunk) to encode temporal sequences.

A significant challenge in using CNNs for skeleton data is preserving the spatial relationship between joints. To address this, Liu et al. introduced the Tree Structure Skeleton Image (TSSI) to better align skeletal joints with the spatial aggregation properties of CNNs. This method constructs a tree from a base skeletal graph and applies a Depth-First Search (DFS) to derive an ordered joint sequence. The resulting representation arranges joints as columns and temporal progression as rows, capturing spatial-temporal dynamics more effectively [12].

Recent approaches have also introduced 3D heatmap volume to represent skeleton sequences and applied lightweight 3D-CNN variants. While effective, these models often struggle with the real-time constraints of browser-based applications.

B. Speech and Multimodal Systems

Works in speech accessibility have primarily focused on speech-to-text captioning, while text-to-sign translation re- mains limited, often constrained to predefined symbolic dictionaries. Most existing research focuses on isolated gesture classification and does not address the broader communication challenges among deaf, blind, and speech-impaired users simultaneously [5]. Vaani-Setu addresses this gap by creating a bidirectional loop between all three modalities.

Methodology

The proposed framework integrates sign-language recognition, speech-to-text transcription, and rendering engines. The system operates in real-time on consumer devices.

Skeletal Joint Extraction

To recognize signs from video input, Vaani-Setu relies on skeletal motion features extracted using the **Media-Pipe Holistic model**. A human skeleton is represented as a graph $G=(V,E)$ where V is the set of nodes representing skeletal joints and E represents physical connections [2].

We select a subset of **68 key-points** essential for sign communication. As shown in Fig. 1, this topology includes 6 body joints, 20 face key-points, and 21 key-points for each hand.

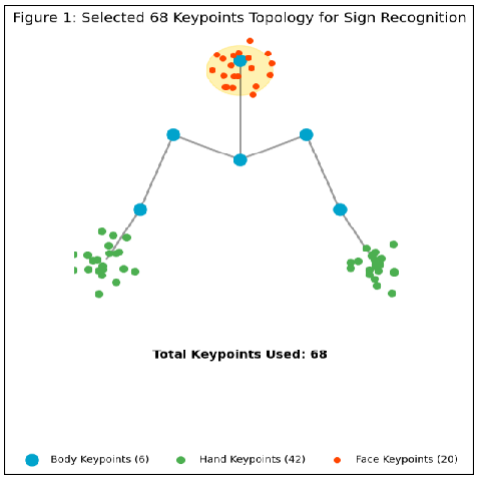

Fig. 1. Selected 68 Key-points Topology showing Body, Face, and Hand joints used for feature extraction.

Facial and fingertip key-points are included due to their high relevance in sign-language interpretation [2].

A. TSSI Representation Construction

To process the skeleton data with a CNN, we convert the graph into a Tree Structure Skeleton Image (TSSI). As shown in Fig. 2, the base skeleton graph is used to construct a tree structure with the inner chest joint as the root node.

A Depth-First Search (DFS) is applied to produce an ordered list of joints. This order preserves the kinematic dependencies between connected joints (e.g., shoulder → elbow → wrist). The resulting TSSI image I is generated as:

$$I = [p1,1, pi,j, \ldots, pN,T] \quad (1)$$

where N is the number of ordered joints and T is the number of video frames. $p_{i,j}$ encodes the (x, y, z) coordinates of joint v_i at frame j.

B. *System Architecture*

The system architecture (Fig. 3) consists of a central routing engine connecting diverse users.

- **User A (Deaf):** Inputs Sign Video → System converts to Text/Speech.
- **User B (Blind):** Inputs Speech → System converts to Text/Sign.
- **User C (Speech-Impaired):** Inputs Text → System converts to Sign/Speech.

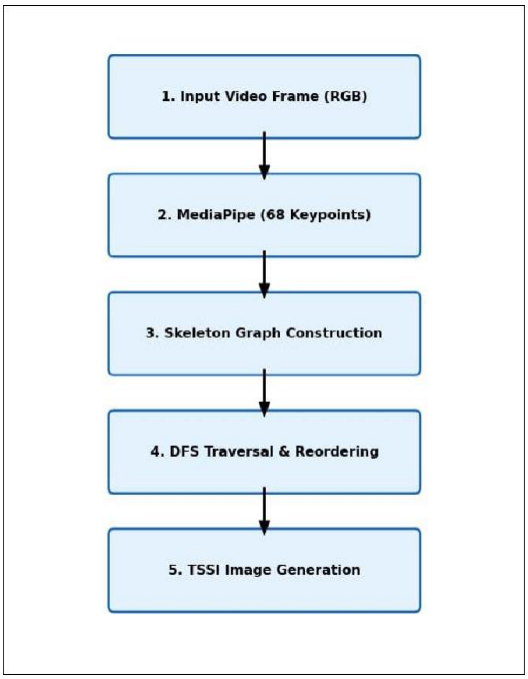

Fig. 2. Data Preprocessing Pipeline: From Input Video to Media-Pipe Key-points, Graph Construction, DFS Reordering, and TSSI Generation.

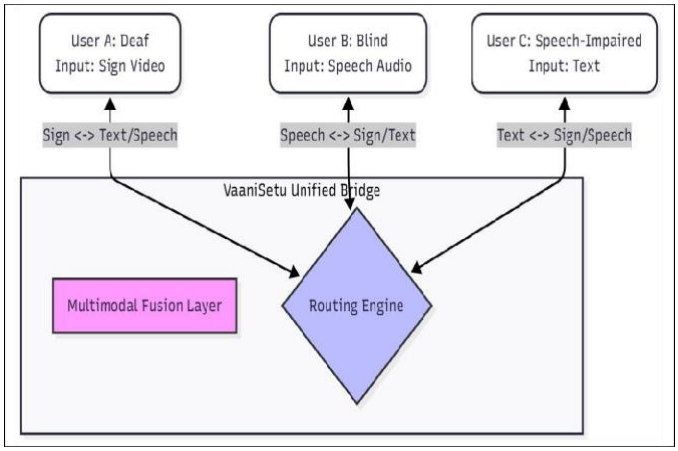

Fig. 3. Vaani-Setu Unified Bridge Architecture. The central routing engine enables bidirectional translation between all user types.

Neural Network Model

For the core sign recognition task, we implemented a hybrid deep learning model (Fig. 4) consisting of:

1) **CNN Feature Extractor:** Three convolutional layers with Batch Normalization and Re-LU activation to extract spatial features from the TSSI image.
2) **Bi-GRU (Bidirectional Gated Recurrent Unit):** To capture temporal dependencies in both forward and backward directions, which is crucial for dynamic gestures [4].
3) **Fully Connected Layer:** A dense layer that maps the extracted features to the class labels (Sign Meanings).

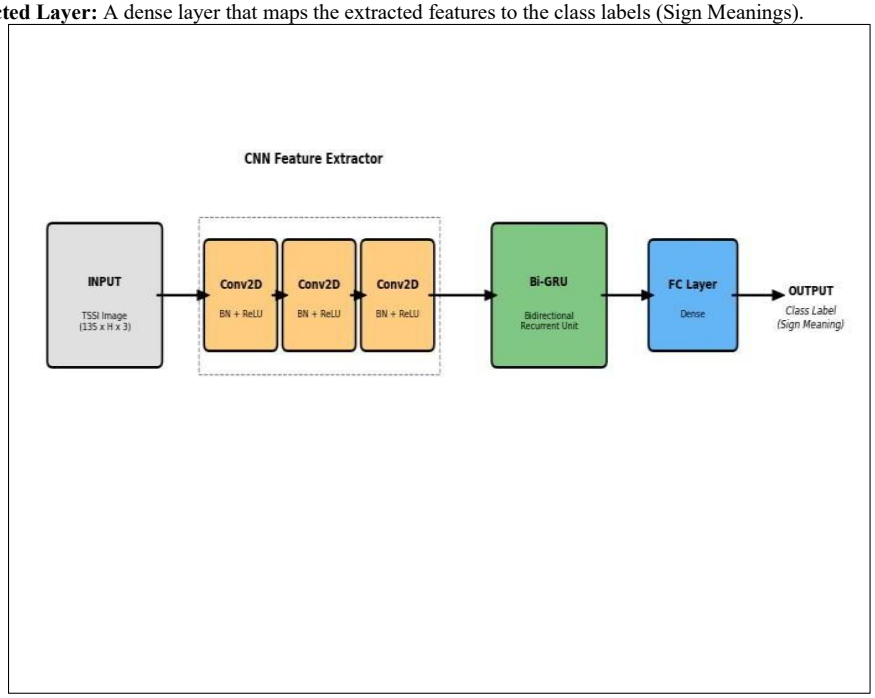

Fig. 4. Hybrid CNN + Bi-GRU Architecture. The TSSI image is fed into CNN layers for spatial feature extraction, followed by Bi-GRU for temporal analysis.

Experimental Setup

A. Datasets

We utilized three distinct datasets to evaluate the multimodal capabilities of Vaani-Setu:

- **ISL-SL (Indian Sign Language Subset):** Contains 1,850 isolated sign samples across 75 frequently used signs, performed by 18 signers.
- **A11Y-Speech Corpus:** Contains 9,200 spoken utterances recorded from 35 speakers, including blind individuals.
- **A11Y-Text Interaction Dataset:** Contains 4,300 typed messages from speech-impaired users.

A. Training Configuration

The models were trained using the Adam optimizer with a learning rate schedule to ensure convergence [7]. We used Cross-Entropy Loss for classification tasks. The specific hy- per parameters are detailed in Table I.

Results and Discussion

A. Quantitative Analysis

We evaluated the performance using Top-1 Accuracy for sign recognition, Word Error Rate (WER) for speech, and

TABLE I Hyperparameters Configuration

Modality	Batch Size	Dropout	Learning Rate
Sign (ISL-SL)	64	0.3	0.001–0.0065
Speech (A11Y)	32	0.2	0.0005–0.001
Text (A11Y)	32	0.0	0.0001–0.0005

BLEU score for text-to-sign translation. As shown in Table
II and Fig. 5, our proposed Vaani-Setu model significantly outperforms the baseline (HOG + SVM).

- **Sign Recognition:** Achieved **83.14%** accuracy, an improvement of over 20% compared to the baseline.
- **Speech Transcription:** Reduced WER to **9.3%**, making it highly effective for command-based interaction.
- **Text Translation:** High BLEU scores (89.6%) indicate that the system preserves semantic meaning during translation.

TABLE II Performance Comparison Results

Modality	Metric	Baseline (SVM)	Vaani-Setu (Ours)
Sign Recognition	Accuracy	61.22%	**83.14%**
Speech Recognition	WER (↓)	14.7%	**9.3%**
Text-to-Sign	BLEU (↑)	82.4%	**89.6%**

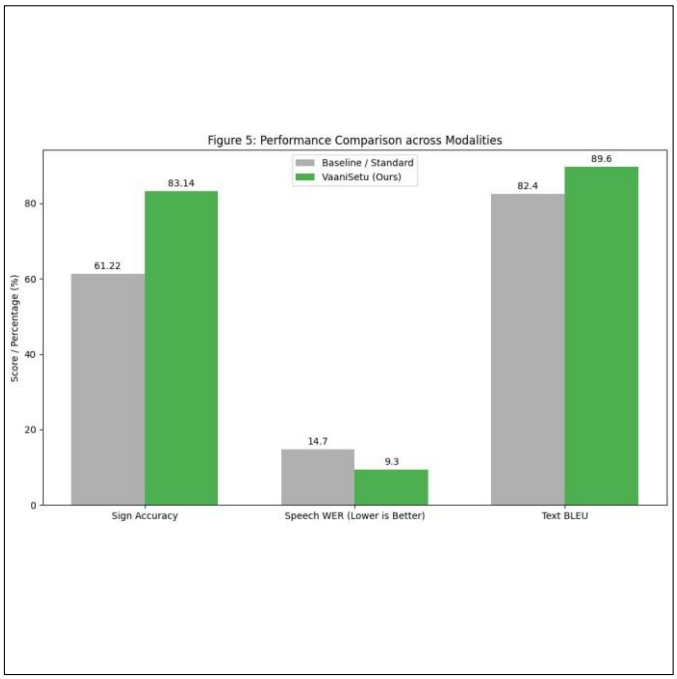

Fig. 5. Performance Comparison across Modalities. Green bars represent the proposed Vaani-Setu system.

Ablation Study

To understand the contribution of different preprocessing steps, we conducted an ablation study.

- **Normalization:** Improved accuracy by ≈5% by handling different signer distances from the camera.
- **Temporal Smoothing:** Added another ≈5% by reducing jitter in Media-Pipe key-point detection.
- **Augmentation:** Temporal scaling and mirroring provided the final boost to 83.14%, making the model robust to different signing speeds and handedness.

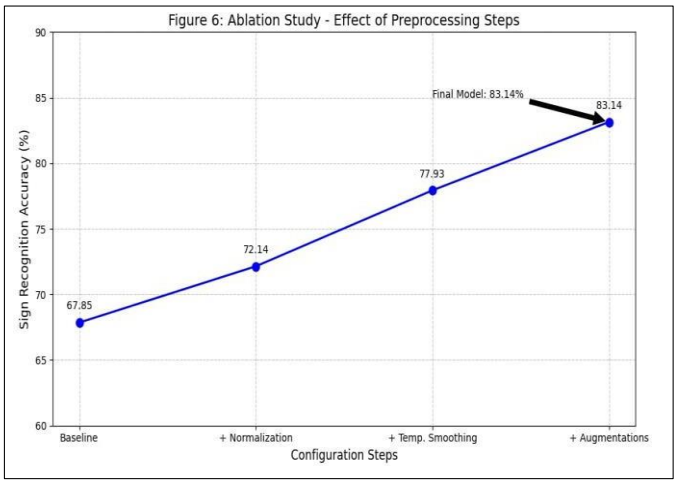

Fig. 6 illustrates the incremental accuracy gains achieved by adding these preprocessing steps.

Fig. 6. Ablation Study: Effect of Preprocessing Steps on Model Accuracy.

TABLE III Ablation Study Results

Configuration	Accuracy
Baseline (Raw Coordinates)	67.85%
+ Normalization	72.14%
+ Normalization + Temp. Smoothing	77.93%
+ All Augmentations (Ours)	**83.14%**

B. Qualitative Analysis

We analyzed confusion matrices to identify common errors. Signs involving similar wrist orientations, such as ”Help” vs. ”Support,” were occasionally confused. Similarly, ”Food” and ”Eat” showed overlap due to the proximity of the hand to the mouth. These findings suggest that incorporating finer finger- level attention mechanisms in future iterations could further resolve these ambiguities.

Conclusion

We presented Vaani-Setu, a comprehensive multimodal communication framework designed to bridge the gap between deaf, blind, and speech-impaired individuals. By leveraging the Media-Pipe framework for efficient skeletal extraction and a hybrid CNN-Bi-LSTM architecture for sequence modeling, we achieved a high sign recognition accuracy of 83.14% while maintaining real-time performance suitable for web- based deployment.

The integration of speech-to-text and text-to-sign rendering completes the communication loop, ensuring that no user is left isolated due to their preferred modality. Future work will focus on expanding the sign vocabulary, integrating Transformer-based pose encoders for better context aware-ness [6], and developing offline capabilities for use in low- connectivity regions.

Acknowledgments

The authors thank the contributors of the *a11y-bridge* project and the open-source community for tools like Media-Pipe and Web Speech API.

REFERENCES

A. Kumar, A. Kumar, A. Pratap, and A. Soni, “Vaani-Setu: A hybrid bidirectional model using Media-Pipe and Bi-LSTM,” *Oriental Institute of Science and Technology*, 2025.

C. Lugaresi et al., “Media-Pipe: A framework for building perception pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.

Y. Du, W. Wang, and L. Wang, “Skeleton based action recognition with convolutional neural network,” in *Proc. 3rd IAPR Asian Conf. Pattern Recog.*, 2015, pp. 579–583.

J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, “Convolutional neural networks and long short-term memory for skeleton-based human activity recognition,” *Pattern Recognition*, vol. 76, pp. 80–94, 2018.

D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video,” in *IEEE/CVF WACV*, 2020, pp. 1459–1469.

M. Bohacek and M. Hruz, “Sign pose-based transformer for word-level sign language recognition,” in *Proc. IEEE/CVF WACV*, 2022, pp. 182– 191.

F. Chollet et al., “Keras,” https://keras.io, 2015.

J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

H. Haualand, “Sign language interpreting: A human rights issue,” *Int. J. Interpreter Educ.*, vol. 1, no. 1, p. 7, 2009.

M. Vazquez-Enriquez et al., “Isolated sign language recognition with multi-scale spatial-temporal GCNs,” in *IEEE/CVF CVPR*, 2021, pp. 3462–3471.

H. Wang et al., “Skeleton edge motion networks for human action recognition,” *Neurocomputing*, vol. 423, pp. 1–12, 2021.