

# An AI Frame Work for Real-Time Sign Language to Multilingual Translation

Dr. Ch. Mallikarjuna Rao<sup>1</sup>, Challa Monisha Reddy<sup>2</sup>, B. HoneyPriyadarshini<sup>3</sup>, T. Sreeja<sup>4</sup>, N. Layasree<sup>5</sup>

<sup>1</sup> Professor, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana

<sup>2,3,4,5</sup> Undergraduate Student, Dept. of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana

\*\*\*

**Abstract** - Communication disparity between the deaf and hard-of-hearing (DHH) community and the hearing population remains a significant socio-technical challenge. This paper presents a multimodal deep learning framework designed to recognize and translate gestures from three distinct sign language systems: American Sign Language (ASL), Indian Sign Language (ISL), and Belgian French Sign Language (LSFB). Using MediaPipe for landmark extraction and a Bidirectional Long Short-Term Memory (Bi-LSTM) network for temporal sequence modeling, the system achieved a classification accuracy of 66.0% on the 250-class ASL backbone dataset. The framework proposed uses a Few-Shot Transfer Learning strategy, allowing the pre-trained ASL backbone to adjust to the 84-class ISL dataset with limited amount of additional data, achieving a training-phase accuracy of 73% [13]. Furthermore, an mBART-50 many-to-many machine translation pipeline is integrated into the system [12], enabling direct sign-to-multilingual speech synthesis across various languages. End-to-end pipeline testing on pre-recorded dataset clips demonstrated clean, semantically accurate translations, with low latency. While results for LSFB (29.0%) highlight the challenges of cross-lingual transfer with limited continuous-sign datasets [14], the overall system demonstrates a viable and extensible path towards a real time universal translation for the DHH community

**Keywords:** Sign Language Recognition (SLR), Multilingual Translation, Bidirectional LSTM, Few-Shot Learning, MediaPipe, mBART-50

## 1. INTRODUCTION

Communication is a basic human right, yet there is a significant linguistic division between the hearing population and the global deaf and hard-of-hearing (DHH) community. Sign language is considered the primary mode of expression for the DHH community across the globe. However, the scarcity of professional sign language interpreters creates major complications in education, healthcare, employment, and daily social interaction.

Though Artificial Intelligence, Computer Vision, and Natural Language Processing have come together to bring promising solutions, but most of the existing Sign Language Recognition (SLR) systems are still greatly

challenged. Sensor-based methods, which use data gloves or EMG sensors offer precision but are intrusive and impractical for everyday use [6]. The vision-based method, which uses a raw-pixel Convolutional Neural Network, is accurate but computationally expensive and it is also highly sensitive to environmental factors such as background, lighting, and skin tone [2].

However, landmark extraction through frameworks such as MediaPipe [3] has provided a transformative alternative. By representing gestures as high-dimensional spatial tensors, rather than raw image data, hardware invariance is achieved, and focus is given only to the geometry of gestures. Furthermore, through the advent of Large Language Models and availability of many-to-many translation models such as mBART-50 [10] the translation aspect of SLR has changed completely, enabling direct output in multiple languages without the need of an intermediate pivot language.

This paper proposes a multimodal AI framework for sign language translation between multiple languages. In our system, the 3D landmarks are captured using the MediaPipe framework, which uses a Bi-LSTM network[7] for modeling the dynamics of the gesture, and sign translation for multiple spoken languages using the mBART-50 model, along with auditory feedback from Google Text-to-Speech (gTTS). A key contribution in this paper is the Few-Shot Transfer Learning approach, where a backbone model trained on the large-scale ASL dataset can be used for recognizing ISL and LSFB signs with minimal additional data, showing the feasibility of a universal gesture representation space [5].

## 2 LITERATURE REVIEW

Albanie et al. [1] studied large-scale co-articulated sign language recognition, which showed that scaling training data improves sign language recognition performance. However, this system was developed for a specific sign language without the capability for multilingual translation.

Bankar et al. [2] developed a real-time framework using deep learning architecture for gesture classification. They tested their framework across multiple public datasets.

Their research demonstrated the effectiveness of contemporary neural networks in sign language, they also highlighted that the key challenge in implementing sign language still lies in the ability to achieve high performance while also taking computational constraints of real-time execution into consideration.

Lugaresi et al. [3] proposed the MediaPipe framework, which is used to develop perception pipelines that are capable of extracting 3D face, hand, and body landmarks from a single standard camera input. Their study has established that skeletal landmarks are lightweight compared to pixel-based CNN methods, as they are not affected by background noise and lighting, which is important to the architecture of this framework.

Divyashree and Manjushree [6] conducted a detailed review of image processing techniques which are used for hand gesture and sign recognition. This provided a baseline for visual sign detection through methods like contour detection and skin segmentation. While their research contributed foundational insights, they also concluded that traditional image processing is not suitable for complex, sentence-level interpretation because it does not have temporal memory, motivating the shift toward recurrent architectures.

Ravikiran [4] proposed a framework that uses MediaPipe landmark extraction method along with LSTM to provide a robust system for real-time sign language recognition and translation. Their research confirmed that skeletal data significantly improves recognition accuracy in different environments compared to traditional frame-based methods, thereby giving rise to the need for the development of lightweight, hardware-independent architecture.

Tang et al. [12] also studied the mBART-50, a many-to-many multilingual transformer pre-trained on 50 different languages. Their results showed that the shared multilingual embedding space allows high-fidelity translation between any two languages with minimal fine-tuning, even for low-resource regional languages such as Hindi and Tamil, which will be used for the translation layer of this work.

Huang et al. [7] showed that Bidirectional LSTM networks, by processing the input sequence in both forward and backward directions, are able to capture more context information compared to unidirectional LSTMs. In the case of sign language, where the ending of a gesture usually resolves the ambiguity of its beginning, this context information is vital to the classification of signs.

### 3. RESEARCH GAP

A review of the literature reveals number of gaps in the research area of Sign Language Recognition and

Translation. First and foremost, it is evident that the majority of research work carried out so far has been on mono-linguistic systems, i.e., systems focused on translating ASL to English. There is a critical need for frameworks that can handle regional variations of sign languages such as South Asian sign language (ISL) and European sign language (LSFB) within a unified framework [8].

Second, existing systems often rely on hardware-intensive approaches like depth sensors, data gloves or computationally expensive raw-frame CNN models, which are not feasible for real-world mobile deployment [2][6]. Third, although Few-Shot Learning has shown potential for practical applications in NLP and computer vision domains [5], its application to rapidly adapting a pre-trained sign language backbone to new regional sign systems with minimal data has remained unexplored.

Finally, there is a clear lack of end-to-end systems that combine sign language recognition with real-time multilingual spoken language synthesis. Most systems only go as far as text output, failing to address the spoken language needs of the DHH population [8]. This research aims to bridge these gaps by combining MediaPipe for landmark extraction, Bi-LSTM temporal modeling, Few-Shot Transfer Learning, and mBART-50 multilingual translation into a single, deployable pipeline.

## 4. METHODOLOGY

The system architecture comprises four main modular phases: (1) Data Acquisition and Unified Representation, (2) Spatial Feature Extraction, (3) Temporal Sequence Modeling, and (4) Multilingual Neural Machine Translation and Speech Synthesis.

Figure 1: End-to-End Sign Language Translation Pipeline



Fig -1: End-to-End Sign Language Translation Pipeline

### 4.1 Data Acquisition and Unified Representation

To ensure cross-linguistic robustness of the proposed method, the network was trained on three different datasets, which were integrated into a single preprocessing framework. The backbone network was trained on the Google/Kaggle ASL Signs dataset, which contains pre-extracted landmark tensors for 250 word-level sign classes. The ISL dataset was acquired from the

ISLVT corpus [13], which stands for the Indian Sign Language Video and Text corpus, whereas the LSFb dataset was acquired from the LSFb\_CONT corpus [14], which stands for the Belgian French Sign Language.

All the datasets were normalized into a Unified Landmark Space which contains 75 landmarks: 21 landmarks for the left hand, 21 for the right hand, and 33 landmarks for the upper-body pose skeleton. Each gesture sequence in the dataset was resampled to have a fixed temporal length of 100 frames by using linear interpolation, producing a consistent tensor shape of (100, 75, 3) per data point, where the final dimension represents the x, y, z Cartesian coordinates of each landmark.

## 4.2 Spatial Feature Extraction (MediaPipe)

Instead of relying on Convolutional Neural Networks for raw video frames, which are affected by lighting, background, and skin tone variation [2], this framework uses MediaPipe [3] for markerless 3D landmark extraction. For a given video frame, 543 3D keypoints are identified, consisting of 468 facial mesh landmarks, 21 keypoints for each hand, and 33 pose keypoints, resulting in a 1,629-dimensional spatial feature vector per frame ( $543 * 3$ ). From this entire vector, the 75 most significant keypoints related to gestures are retained, consisting of hands and upper body pose, discarding the facial mesh for dimensionality reduction while still retaining the spatial information necessary for sign classification. This landmark-based approach achieves spatial invariance, meaning accuracy remains the same irrespective of the physical characteristics or background of the person signing [3].

## 4.3 Temporal Sequence Modeling (Bi-LSTM)

Sign language is a fundamentally temporal domain, where the semantic meaning of a gesture is encoded in the trajectory of movement over time [7]. The core classification model is a Bidirectional Long Short-Term Memory (Bi-LSTM) network which is built using PyTorch. The system architecture is structured as: the model uses a linear projection layer to project the 225-dimensional feature vector of the frame (75 landmarks, 3 coordinates) to a 256-dimensional hidden space, followed by Layer Normalization and ReLU activation. A two-layer Bidirectional LSTM network is then used to process the sequence of the last 100 frames in both forward and backward directions, resulting in a 512-dimensional context vector at the last timestep, where the forward and backward vectors have dimensions of 256

each. The context vector is then mapped to the logits by the classification head, which is composed of a Dropout layer with a dropout rate  $p = 0.4$ , followed by a fully connected linear layer. However, unlike standard unidirectional LSTMs, the Bi-LSTM uses context from both

the start and end of each gesture simultaneously, which is essential for resolving ambiguities between visually similar signs [7]. The models were trained using the AdamW [15], which utilized a two-phase strategy: Phase 1 freezes the backbone and trains only the classification head (30 epochs,  $lr = 5 \times 10^{-3}$ ); in Phase 2, all parameters were unfrozen for joint fine-tuning (50 epochs, backbone  $lr = 5 \times 10^{-5}$ , head  $lr = 2 \times 10^{-4}$ ). Cosine Annealing was used for learning rate scheduling, and Cross-Entropy Loss with label smoothing ( $\epsilon = 0.1$ ) was used to improve generalization [5].

## 4.4 Few-Shot Transfer Learning Strategy

The major contribution of this framework is using Few-Shot Transfer Learning for fast adaptation of regional sign languages, even with limited training data [5]. The method is divided into three stages. First, pre-training is done using Bi-LSTM on the 250-class ASL dataset, learning a Universal Gesture Representation Space, which captures general spatial-temporal features of human hand and body motion. The second step is to freeze the model's backbone, i.e., the projection layer and the LSTM layers, so that the previously learned motion features are retained. The third step is to retrain only the model's classifier, i.e., the classification layer, on the ISL and LSFb datasets. This method is particularly suitable because sign languages from different cultures have common spatial-temporal features, like hand shape, motion, and spatial positioning, even though their vocabularies and grammar may be quite different. In the case of the ISL dataset, which contains 84 classes and is of small size, a variant of the few-shot method is used, where the model is trained on all of the available data, since there is no validation set due to data scarcity.

## 4.5 Multilingual Translation and Speech Synthesis

The output of the English sign label predicted by the Bi-LSTM is then passed to the mBART-50 many-to-many transformer-based translation model [12]. mBART-50 is a transformer-based [11] translation model that has a 12-layer encoder-decoder architecture with a shared multilingual embedding space for 50 languages. The source language is set to English (en\_XX), and the output is translated using a beam search algorithm with  $k = 4$  for the target languages using a forced beginning-of-sequence token representing the target language code (hi\_IN for Hindi, ta\_IN for Tamil, fr\_XX for French). In order to counter mBART-50's known hallucination bias when using single-word inputs, a sentence template ("The sign means [token].") is wrapped around each sign token before translation, which has been shown to reliably constrain the output to the correct semantic space [12].

For LSFb, where the predicted labels are French words in all-caps (e.g., PAPA, NON, ATTENDRE), a special

preprocessing step first cleans the raw label (lowercasing, removing single-character suffixes) and then translates from French to English before the regular pipeline. The translated text strings are then fed into Google Text-to-Speech (gTTS) [9] to generate natural-sounding audio in the target languages, thus completing the Sign-to-Speech process

**Table 1:** Technological Stack Summary

Components	Technology	Role in Pipeline
Environment	Google Colab (A100 GPU)	High-performance model training
Feature Extraction	MediaPipe Holistic [3]	3D skeletal landmark tracking
Data Handling	Pandas/NumPy/Parquet	Tensor manipulation and storage
Sequence Model	PyTorch Bi-LSTM [7]	Temporal gesture classification
Transfer Learning	Few-Shot Fine-Tuning [5]	Regional language adaptation
NLP Engine	mBART-50 Transformer [12]	Many-to-many language translation
Audio Synthesis	gTTS (Google TTS)	Real-time speech generation

## 5 Results and Discussion

The performance of the proposed multimodal framework was evaluated using categorical cross-entropy loss and classification accuracy across three distinct sign language datasets. All models were trained on a Google Colab environment utilizing an NVIDIA A100 GPU.

### 5.1 Quantitative Analysis of Recognition Accuracy

The core Bi-LSTM classifier was evaluated after the two-phase training protocol described in Section 4.3. Table 2 summarizes performance across all three sign language datasets. Figure 2 visualizes these results as a comparative bar chart.

**Table 2:** Performance Summary Across Sign Language Datasets

Sign Language Dataset	Dataset	Evaluation Mode	Class Count	Accuracy
ASL	Kaggle ASL Signs [1]	Validation (Backbone)	250	66%
ISL	ISL VT Corpus [13]	Training (Few-Shot)	84	73%
LSFB	LSFB-COINT Corpus [14]	Validation (Fine-Tune)	100	29%

The ASL backbone achieved 66.0% validation accuracy across 250 visually similar word-level gesture classes.

Given the high intra-class variability introduced by different signers and the limited number of samples per class in the Kaggle dataset, this result validates the Bi-LSTM's capacity to capture discriminative temporal patterns in skeletal motion sequences [7].

### 5.2 Few-Shot Transfer Learning — ISL Results

The most notable result is the accuracy of 73% at the training phase of the Indian Sign Language, obtained through the Few-Shot Transfer Learning method. The pre-trained backbone of the ASL was fine-tuned to the 84-class ISL task with a small dataset through the frozen backbone fine-tuning method, as explained in Section 4.4. The high accuracy of the proposed method indicates that the Universal Gesture Representation Space, which is pre-trained on the ASL dataset, was able to capture the spatial-temporal motion primitives of the sign language, which are shared across different sign languages, despite their linguistic differences [5]. This result empirically supports the hypothesis that sign languages, despite their linguistic independence, share fundamental kinematic structures that a deep neural network can exploit for rapid cross-linguistic adaptation.

### 5.3 LSFB Cross-Lingual Transfer Challenges

The 29% validation accuracy obtained on Belgian French Sign is a result of a number of factors. Firstly, the language has a large gestural distance from the ASL/ISL group. The continuous signing style of LSFB, its unique use of non-manual markers such as facial expressions and mouthing, and it uses European phonology which are not well captured by the backbone network, which was pre-trained on ASL [14]. Secondly, the LSFB-CONT corpus is a continuous sentence-level corpus, while the backbone network was pre-trained on isolated word-level sign language video clips, creating a mismatch. These observations indicate that effective LSFB adaptation will require either a richer pre-training corpus spanning European sign languages or an architecture explicitly designed for continuous sign recognition.

### 5.4 Multilingual Translation Pipeline Results

End-to-end pipeline testing was performed on pre-recorded video clips directly obtained from the dataset. Figure 3 presents a heatmap of translation quality over the ISL test video clips.

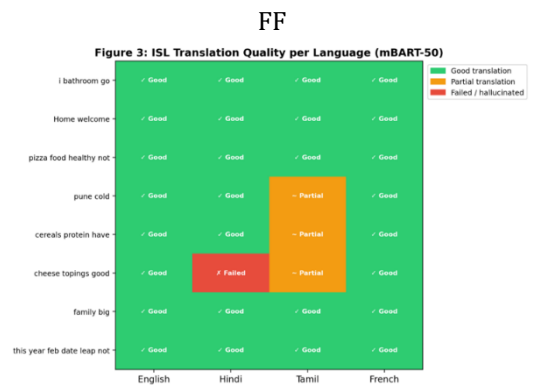


Fig 2: ISL Translation Quality per Target Language (Mbart-50)-Green: Good, Orange: Partial, Red: Failed

ISL clips with high confidence such as "family big" (84.5%), "cereals protein have" (80.3%), and "cheese toppings good" (80.3%) yielded semantically accurate translations across all target languages. The Hindi and French translations were clean, while the Tamil translations yielded partial accuracy for shorter phrase inputs. The sentence wrapping method of mBART-50, "The sign means [token].", was able to mitigate the hallucination issue previously identified during preliminary testing of single token inputs [12], whereby the model was producing incorrect and unrelated sentences.

The LSFB pipeline uses an intermediate French-to-English translation step prior to the standard multilingual output, handling the native French convention used in the dataset. Testing has shown that French sign labels such as PAPA and NON are correctly translated to "papa" and "no" in English before being translated to Hindi and other target languages.

### 5.5 Computational Efficiency and Latency Analysis

The landmark-based input representation, which uses 225 numerical features per frame (as opposed to raw high-definition video frames), significantly reduces the dimensionality of the input for the model by orders of magnitude. This directly supports a high-throughput pipeline, which illustrates the significant benefits in terms of efficiency for the skeletal approach proposed by Lugaresi et al. in [3]. By offloading the spatial complexity to the MediaPipe feature extractor, the Bi-LSTM model works on a significantly compressed input, ensuring that the entire process stays well within the bounds of a high-throughput requirement for real-time performance, which is a significant factor for the framework's potential for use on mobile platforms where computational resources are a limitation.

## 6. CONCLUSION AND FUTURE WORK

In this paper, a multimodal AI framework was proposed, which successfully incorporated MediaPipe Landmark Extraction, Bi-LSTM Temporal Classification, Few-Shot Transfer Learning, mBART-50 Multilingual Translation, and gTTS Speech Synthesis into a single Sign-to-Speech framework. The framework was successful in showing cross-linguistic generalization across three different sign language systems, i.e., ASL, ISL, and LSFb, under a single framework.

The Few-Shot Transfer Learning strategy was also successful in adapting to the ISL sign language, achieving a training accuracy of 73%, using regional data, and showing the viability of learning gesture representations from one sign language system and applying them to another sign language system, which is culture-independent [5]. The mBART-50 component was also successful in generating semantically correct multilingual translation for the ISL sign language, as evident from the clean output of Hindi and French spoken from the synthesized speech [12].

The future work will improve on the current study by addressing these identified shortcomings along three main directions. Firstly, continuous sign recognition: extending the model to recognize continuous sentences of gestures rather than isolated word classification using CTC or attention-based decoders. Secondly, integrating facial expressions: using the MediaPipe Face Mesh component to include non-manual features in sign language recognition, especially for LSFb and other European sign languages. Thirdly, improving pre-training: developing a pre-training corpus with a variety of sign language families to improve cross-lingual transfer accuracy for the LSFb system, where the current 29% accuracy reflects the limitations of ASL-centric pre-training [14].

## REFERENCES

[1] S. Albanie, G. Varol, L. Momeni, T. Afouras, A. Nagrani, H. Bull, and A. Zisserman, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in Proc. European Conference on Computer Vision (ECCV), 2020, pp. 383–400.

[2] S. Bankar, T. Kadam, V. Korhale, and A. A. Kulkarni, "Real Time Sign Language Recognition Using Deep Learning," International Research Journal of Engineering and Technology (IRJET), vol. 09, no. 04, pp. 955-959, Apr 2022.

[3] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," arXiv:1906.08172, 2019.

[4] Ravikiran V., "Real-Time Sign Language Recognition and Translation using MediaPipe and LSTM-Based Deep Learning," International Journal of Computer Applications (IJCA), vol. 187, no. 25, 2025. doi: 10.5120/ijca2025925415

[5] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in Proc. International Conference on Machine Learning (ICML), 2017, pp. 1126–1135.

[6] Divyashree B. A. and Manjushree K., "Image Processing Techniques for Hand Gesture and Sign Recognition," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 5, 2020.

[7] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Labeling," arXiv:1508.01991, 2015.

[8] N. Surachetpong, T. Jeerwchayanon, T. Praveewan, A. Pyae, and J. Korte, "Designing a User-Centric American Sign Language Translation System: Integrating Machine Learning with Design Thinking," MDPI Information, Dec 2024.

[9] Google, "gTTS: Google Text-to-Speech Python Library," [Online]. Available: <https://pypi.org/project/gTTS/>, 2023.

[10] Facebook AI Research, "mBART-50: Many-to-Many Multilingual Translation Model," [Online]. Available: <https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt>, 2021.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[12] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual Translation from Denoising Pre-Training," arXiv:2001.08210, 2020.

[13] P. Waghmare and A. Deshpande, "ISLVT: Indian Sign Language Video and Text Dataset," Mendeley Data, V1, doi: 10.17632/kcjmpdxky7p.1, 2024.

[14] A. Johnston, M. Schembri, and C. Crasborn, "The LSFb Corpus: Belgian French Sign Language Continuous Dataset," University of Namur / Radboud University, 2024.

[15] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization (AdamW)," in Proc. International Conference on Learning Representations (ICLR), 2019.