Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH & REVIEWS

journal homepage: www.ijmrr.online/index.php/home

# SIGNORA: AN AI BASED MULTIMODAL SPEECH TRANSLATION

**Dr. Rumana Anjum[1], Aravind S Gudi[2]*, Bindushree D S[3], Chethan K[4] & Harsha M R[5]**

[1] Assistant Professor, Department of Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology (VVIET), Mysuru, Karnataka, India.

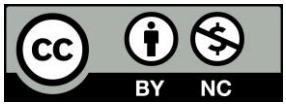
[2,3,4,5] Department of Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology (VVIET), Mysuru, Karnataka, India.

*Corresponding E-mail: aravindsgudi28@gmail.com

**How to Cite the Article:** Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

https://doi.org/10.56815/ijmrr.v5i10.2026.75-90

| *Keywords* | *Abstract* |
|---|---|
| SIGNORA, Artificial Intelligence (AI), Indian Sign Language (ISL), Bidirectional Communication, Sign Language Recognition, Computer Vision, Deep Learning, Natural Language Processing (NLP), Real-Time Translation, Assistive Technology. | There are various communication barriers in communicating between people who use Indian Sign Language (ISL) and people who do not use this language. The paper introduces an intelligent solution called SIGNORA that will help in bidirectional ISL translation using multimodal methods. SIGNORA will extract spatiotemporal landmarks and use deep learning, natural language processing, and text-to-speech systems to improve the accessibility of people within healthcare, education, and public service facilities.<br>SIGNORA employs a deep learner for 3D sequence classification called ONNX Runtime Web to extract hand, body pose and facial landmarks. These recognized gestures will be translated into meaningful sentences in English using grammatical rules and language models. The speech recognition technology will play an important role in translating to ISL using GIFs for phrases and still images for fingerspelling. The adaptation of the rendering algorithm will be done in such a way that priority will be given to phrases, then words, and finally characters. SIGNORA has been designed in such a way that it will be available online. It is made up of JavaScript frontend and Flask-SocketIO backend. |



The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

## 1. INTRODUCTION

Communication is required to engage oneself socially. The capability to communicate is necessary in order to engage in society, education, employment, and healthcare. People who rely on ISL in order to communicate may face difficulties in communicating with people unfamiliar with sign language. This may result in the transfer of important information being delayed in hospitals, educational institutions, and government bodies because of the lack of an interpreter.

Thanks to recent breakthroughs in computer vision, deep learning, NLP, and speech technology, camera-based technologies for sign language translation can now be developed. But sign language recognition is quite a complex process. Not only does it require the detection of signs by the shape of hands, but also by movement of hands, position of both of the hands, body language, and facial expressions. Moreover, grammar rules should be taken into account while translating sign language into text.

### Concept of Multimodal Sign Language Translation

"Multimodal" means a system that employs more than one mode as input or output. These include video, landmarks, audio, text, and ISL signs. For instance, when it comes to the SIGNORA system, the video recognition technique is used for recognition of signing whereas the audio recognition technique is utilized for understanding spoken languages. The data can be represented in text form or speech synthesis."

### Problem Statement and Objectives

The communication between hearing impaired and hearing people remains a significant issue owing to the lack of efficient means of translation of Indian Sign Language (ISL). Existing means of solutions focus on translating sign language to speech or vice versa but do not provide any assistance in terms of interactive translation or pre-recorded sign language translation. Moreover, the existing solutions face difficulties in recognizing gestures and creating coherent sentences using sign language. Hence, there arises a need for an AI-based communication system.

### Objectives and Scope

The primary focus of SIGNORA is on creating an AI-based web platform for two-way communication between Deaf and Hard of Hearing, speech impaired and hearing individuals using Indian Sign Language. This system involves four basic functions including real-time ISL to speech translation, real-time speech to ISL translation, single clip video translation and multi-clip video translation to grammatically correct English sentences. This system employs MediaPipe for landmark detection, Bi-LSTM with Attention model for gesture recognition and NLP for language processing and sentence reconstruction. This system involves ISL GIFs, fingerspelling sequence, speech output and Two-way Communication dashboard involving QR code pairing. Gestures processing via browser helps in minimizing the number of raw webcam frames sent to the server.

## 2. LITERATURE REVIEW

Both the recognition process and the translation process have transitioned from conventional image processing methods to advanced AI approaches that involve computer vision, feature extraction, deep learning, and NLP. Recent research work has mostly focused on dynamic gesture recognition, real-

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

time translation, speech-to-sign language translation, and bidirectional communication systems. This forms the basis for SIGNORA.

### 2.1 Vision-Based Indian Sign Language Recognition

Several research studies have been conducted during recent years to identify ISL through the use of image processing, feature extraction, and machine learning methods to enable automatic sign recognition utilizing cameras that are readily available.

In 2022, Katoch, Singh and Tiwary [1] proposed an Indian Sign Language Recognition System Using SURF with SVM and CNN. This solution used the Bag of Visual Words (BoVW) approach along with Speeded-Up Robust Features (SURF) for extracting visual features. Skin colour segmentation and background subtraction were employed during pre-processing. Support Vector Machine and Convolutional Neural Network are two models used for classification. The system recognized ISL alphabets and digits in real-time and produced text and audio output. However, this primarily focused on recognizing individual signs.

### 2.2 Landmark-Based Recognition and Temporal Sequence Modelling

With recent innovations in computer vision, researchers have utilised landmark based methods to extract movements from hands, body, and facial features. Landmarks provide a way to represent sign language in spatial format, enabling sequence analysis.

An ISLR system based on OpenPose was developed by Malhotra, Sankpal, Maniar, and Thakkar [2] in 2019. Their system used pose detection from OpenPose to extract movements of the body and hands. The Dynamic Time Warping (DTW) algorithm was implemented to compare gesture sequences. As this method does not require any gloves or extra wearable technology, it provides a simple form of vision based ISLR. This method highlighted that it is important to account for body and hand movements when recognising gestures. However, it still require improvements to be made to increase the systems vocabulary.

An Integrated MediaPipe-Optimised GRU Model was developed by Subramanian et al [3] in 2022. This model uses MediaPipe for landmark extraction and an optimised GRU model to recognise ISL gestures. The model was designed to learn and understand temporal patterns. This shows that landmark based features can be used alongside a recurrent neural network to recognise dynamic gestures. However, the landmark quality, variety in the data set, and range of gestures will impact the recognition performance.

A Comprehensive Approach to ISLR: Using LSTM and MediaPipe Holistic for Dynamic and Static Hand Gesture Recognition was presented by Rawat et al [4] in 2025. This system uses MediaPipe Holistic and a Sequential LSTM network. The models extracts hand, body and facial landmarks. Two LSTM layers are used to classify gestures. The model was tested on a customised data set containing 11 gestures and achieved an accuracy of 96.97%. This shows that combining holistic landmark extraction with sequence learning can be effective. However, the amount of gestures available is limited and could be tested on more people who have not signed before.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

### 2.3 Deep Learning and Transformer-Based Recognition

Recent works have shown that deep learning models like CNN, LSTM, GRU and Transformers can be used to improve upon gesture recognition by analysing complex spatial and temporal information. Hybrid models have been investigated.

SIGNFORMER: DeepVision Transformer for Sign Language Recognition [5] was proposed by Kothadiya et al., in 2023. This study focused on using a Transformer model to extract spatial and temporal data from sign language gestures. Thus attention based models can work well for sign recognition. However, only using gesture recognition will not provide a full communication solution without speech processing and sign generation.

Indian Sign Language Detection for Real-Time Translation Using Machine Learning [6] was presented by Singhal et al., in 2025. This study used a Convolutional Neural Network for gesture recognition and integrated the use of MediaPipe to provide hand tracking. The study achieved a classification accuracy of 99.95%. This shows how well a CNN can be used for sign recognition but the accuracy depends on factors like the dataset structure, testing process, variety of signers and quality of translation at sentence level.

Hybrid LSTM-Transformer Architecture with Generative AI for Enhanced Indian Sign Language Recognition [7] was introduced by Goel, Rani and Sarangi in 2026. The experiment conducted involved the use of a transformer model that had LSTM along with gesture recognition through MediaPipe. Text-to-speech services were given through Generative AI. There was a high accuracy of over 97% in the case of 7 gestures that were selected for recognizing ISL. Speech is generated in order to help people communicate with the ones who can listen. This is how effective the combination of time learning, attention, and speech can be when using a certain model. However, only a limited number of gestures have been considered in this study.

### 2.4 Bidirectional Translation and Linguistic Processing

Beyond gesture recognition, effective communication requires translation from verbal language to appropriate sign language and gesture to meaningful sentences. Various studies have been conducted on Natural Language Processing, Grammar Translation, Speech Recognition and Visual Sign Recognition.

In 2021, a Speech to Indian Sign Language Translator was proposed by Monga, Bhutani, Ahuja, Maida and Pande [8]. This used Speech Recognition technology to convert spoken English into text. The Natural Language Processing technology was used to translate sentences. Rules of ISL were also considered while translating sentences and Stop Words were removed. Stemming and Lemmatisation were done. Corresponding sign videos were then extracted from the dictionary. This highlighted the need to use Grammar Translation in Natural Language Processing instead of Word-for-Word translation. However, this was limited to translation of speech into ISL only and did not consider a complete gesture recognition system.

In 2025, Adithyaraaj et al. proposed an Indian Sign Language (ISL) Translator: AI-Powered Bidirectional Translation System [9]. Using Computer Vision, Convolutional Neural Networks and

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

Natural Language Processing based on Transformers, this system could be used to communicate between ISL users. The aim was to translate verbal and textual language into ISL, and translate recognised gestures into text/speech. This work is very similar to the aim of SIGNORA, however there is more experimentation required to ensure the model can be generalised, and that the translations are linguistically correct.

Real-Time Indian Sign Language Recognition and Multilingual Sign Generation was presented in 2025 by Ahmed Aslam, Mohamed Aabidh and N. Sabiyath Fatima [10]. Gestures were identified through the use of Deep Learning and Computer Vision in addition to the LSTM model. The generation of multilingual signs was done by the use of animated 3D models. This highlights the need for gesture identification and visualization. Nevertheless, the availability of signs and their linguistic correctness should be considered.

### 2.4 Datasets, Continuous Translation, and Research Surveys

The availability of different data sets and ways of evaluating them plays an important role in the creation of efficient ISL translation systems. Recent approaches have mainly concentrated on continuous sign translation, sentence level translation, and problems in the current recognition techniques.

Joshi, Agrawal and Modi [11] created the ISLTranslate: Dataset for Translating Indian Sign Language in 2023. It provided a dataset of 31,222 sentence and phrase pairs translating from ISL into English. This helps support research into continuous translation from ISL to English. However there are still issues with having a dataset for translation due to the variation of signers, errors with recognition and the real time application.

Liang, Li and Chai [12] produced Sign Language Translation: A Survey of Approaches and Techniques in 2023. The work covered many different tasks within sign language translation including sign to gloss to text, sign to text and gloss to text. This covers transformation based methods and highlights the complexities of sign language due to the independent grammar structure and nonmanual features such as facial expressions, body postures and movement. This work is an important base for understanding the difference between sign language recognition and full translation.

Rashmi J, Sahani and Yadav [13] provided a review of recent advancements in Indian Sign Language Recognition. They discussed many different approaches including using CNNs, Hidden Markov Models, transfer learning and HOG feature extraction. They also discussed using advanced models and NLP to improve sign recognition and translation. The work identified many issues with sign language recognition including variation between signers, computational complexity, environment and linguistic differences. This highlights the need for efficient and practical recognition systems.

Damdoo and Kumar [14] gave an integrative survey of Indian Sign Language Recognition and Translation in 2025. The work discusses methods using machine learning, deep learning and gesture based methods. This includes looking at building datasets and dictionaries for ISL. The work highlighted the issues with current research and the need for large datasets, increased variation in signers and better support for continuous sign language translation. This work gives an overall understanding of current research and shows need for integrated ISL translation tools.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

Negi [15] provided a review of Deep Learning Models for Indian Sign Language Recognition in 2025. The work discusses using CNNs, RNN, LSTM networks, hybrid methods and transformer based methods. They also looked at advances in continuous sign recognition. The work identified issues with limited training data, variation between signers and real time issues. This highlights the need for multimodal systems and end to end translation models to increase accessibility of ISL technology.

## 3. SUMMARY OF ANALYSIS

The literature review indicates the extent to which we have reached in the field of Intelligent Sign Language Recognition and Translation by applying methods like AI, Computer Vision, Deep Learning, and Natural Language Processing. Recent literature reviews have applied various methods such as Convolution Neural Networks, Long Short Term Memory, Gated Recurrent Units, Transformers, Mediapipe, and Openpose to recognize both static and dynamic gestures.

Research has been conducted on converting recognized gestures into meaningful text and speech. Research has been done on intelligent translation of speech into ISL. The research has been done on the importance of considering the structure of ISL for translating speech into ISL. Other researches have applied ISLTranslate datasets for translating continuous signs.

There are also some limitations within the current research. The majority of research focuses on single gestures and translates in just one direction; therefore, we lack two-way communication. Also, the number of researches which focus on sentence formation, error correction, and translation in real-time and video is also rather limited. There are also some limitations regarding the lack of gesture dataset, different signers, and computational needs. A platform for complete integration for such communication is required.

The presented SIGNORA platform will try to tackle the aforementioned limitations with the use of the AI, Computer Vision, Deep Learning, and NLP. Within this platform, the fast translation of ISL gestures into text and speech is possible. The platform will allow for translating speech into visual representation of ISL. Translating single pre-recorded videos is also provided. Several short videos may be assembled in order to form a grammatically correct sentence in English along with the speech. The Two-way Communication Dashboard will provide for the synchronization of the communication with the help of QR codes. Gestures are processed with the help of the browser-based approach.

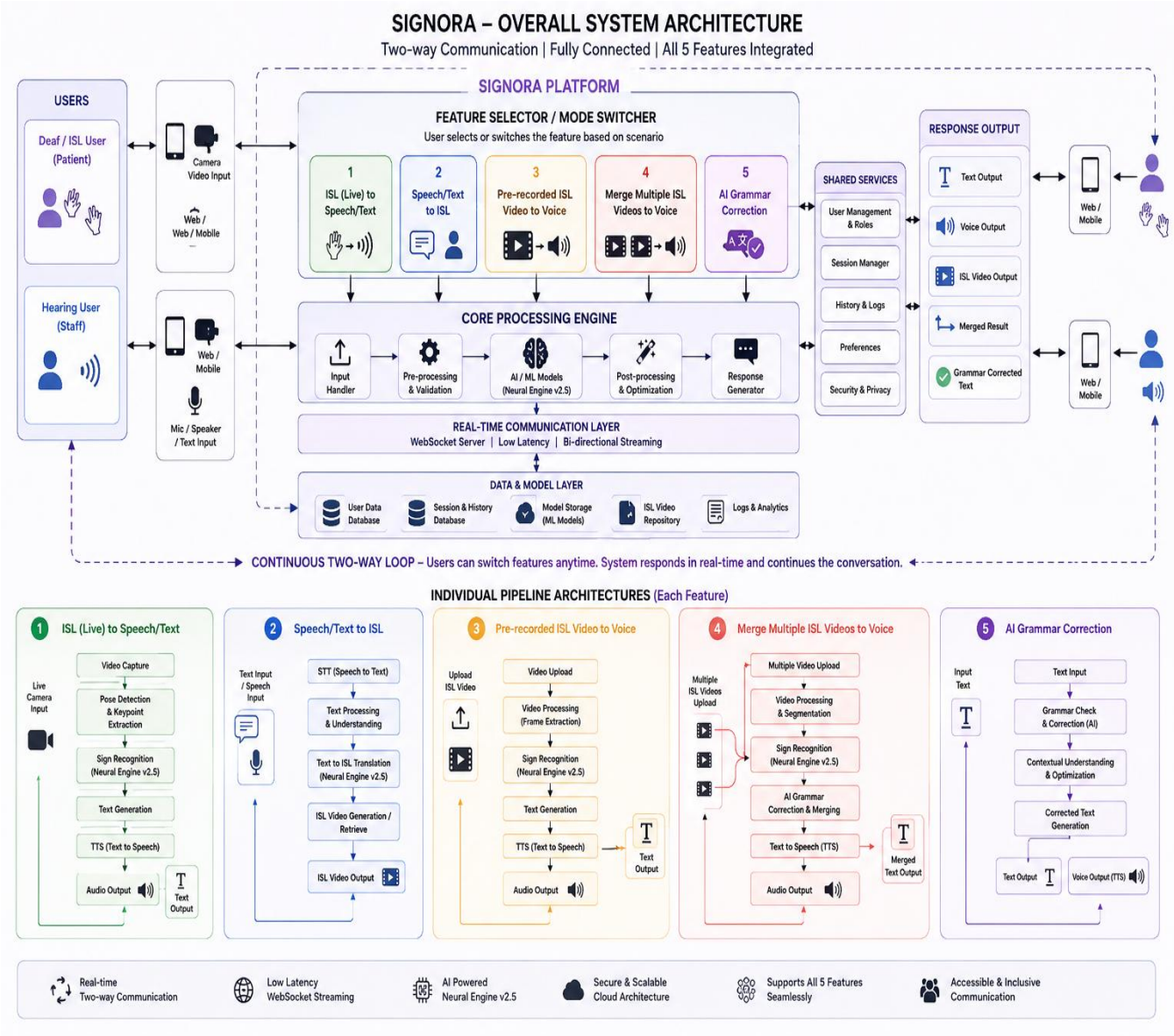
## 4. PROPOSED METHODOLOGY

The proposed system SIGNORA attempts to provide bidirectional communication between the Deaf and Hard of Hearing(DHH) community, speech-impaired and the hearing community using Indian Sign Language(ISL). The proposed system uses Artificial Intelligence(AI), Computer Vision, Landmark detection, Deep Learning, Natural Language Processing(NLP) and Agentic AI techniques for real-time and pre-recorded translation of the Sign Language. The system has four main communication modules that are supported by a Two-way Communication Dashboard. The system uses browser based AI to detect gestures and a multi-level Natural Language Processing engine to generate meaningful and grammatically correct outputs.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

**Figure 1: System Architecture of Signora**




### 4.1 System Architecture and Input Processing

The SIGNORA system begins with an interactive web-based interface allowing the user to choose which form they would like to use. It supports webcam feeds for instant ISL recognition, microphone feeds for converting speech to ISL and uploading video files. The input section will take the chosen input and pass it onto relevant process.

With webcam input, frames are taken from the camera and processed in real time. With microphone input, the Web Speech API will convert the audio into text. With uploaded ISL videos, they will be decoded into frames. Multiple video clips will be ordered according to the user's choice. Once data has been received, it will be validated and prepared for the correct recognition/translation. The input section allows users to use various communication methods.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

### 4.2 Computer Vision, Landmark Detection and Gesture Recognition

When using sign language input, Google MediaPipe Tasks Vision will extract the hand, body pose and facial landmarks from either the live webcam feed or pre-recorded video feed. It will extract 21 3D points from each hand, 6 upper body pose points and 10 specific facial points.

These points will then be passed onto the SIGNORA Landmark Engine. Missing data will be interpolated. Points will be scaled according to the correct body position and hand size. Kinematic features will be calculated to capture the change in movement between frames. These will be combined to create a feature vector of 198 dimensions to be used for each frame. Up to 60 frames of data will be kept and passed onto the gesture recognition model.

A Bidirectional Long Short-Term Memory (Bi-LSTM) with Multi-Head Self-Attention is used in the SIGNORA Custom Deep Learning Model to identify patterns within the data. The trained model will be exported as a ONNX file and run using the ONNX Runtime Web. A confidence prediction will help filter out unreliable predictions. Prediction will then be passed onto the next section. This will allow for instant gesture recognition with no webcam data being sent.

### 4.3 Speech Recognition and ISL Translation

The speech to ISL process can be used by hearing individuals for communication with ISL users. Audio will be taken from the microphone using the Web Speech API and converted into text. This will then be processed using the spaCy NLP to identify important parts of speech such as Subjects, Objects, Verbs, Time, Negation, Question and Named Entities.

This will then be reordered to fit the format used by the system. Unnecessary words will be removed, verbs will be converted into their base form and proper nouns will be identified. This will be matched against the ISL visual dictionary. If there is a match, the relevant animated GIF will show. If there is no match, alphabet fingerspelling will be used. A visual output will be created using the dashboard.

### 4.4 Grammar Processing and Agentic Language Engine

Once gesture recognition has been completed, ISL predictions will be stored as gloss tokens. These will then be passed onto the SIGNORA Grammar Module. Multiple layers of processing will be used to convert the recognised signs into meaningful English sentences.

Firstly, a deterministic mapping will be used to identify any set phrases and generate sentences. Secondly, spaCy will be used to identify grammatical aspects. Subjects and verbs will be extracted along with tense. This will then be used to build meaningful English sentences. If there is a complex, ambiguous or colloquial phrase that cannot be recognised using the previous methods, the Agentic AI language engine will be used.

The large language model (LLM) from the Agentic AI will be used via the OpenRouter gateway. This will identify the gloss sequence and generate a fluent English sentence. A main language model will be used with a failover to a secondary model if there is a timeout or network issue. The response will then be cleaned up to remove any extra formatting and conversational text. This multi-layered approach will help with generating sentences, especially when translating multiple clips.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

### 4.5 Video Processing and Multiple Clip Translation

SIGNORA can translate single video files or multiple short clips. Single video files can be uploaded using the video processing section. This will decode the video into frames and MediaPipe will detect the necessary hand and body landmarks. The points will be scaled and processed using the same ONNX gesture recognition. Predictions will be passed onto the grammar processing engine to generate meaningful English text. This will then be converted into audio using the Web Speech API.

Multiple video clips can be uploaded. This can be a series of short videos showing each word/concept. Each video will be processed and the predictions will be stored. These will be combined to create a gloss sequence and will be ordered according to the user's choice. This will be passed onto the multi layered language engine. Deterministic mapping, spaCy and Agentic AI language generation will be used as needed. The content will then be put together to create a full grammatical sentence. The video section will allow the user to view the clips and reorder them.

### 4.6 Output and Two Way Communication Dashboard

Once all processing has been done, outputs will be created based on what communication the user wants to use. If ISL to speech, gestures will be recognised and converted into meaningful English text and audio using the Web Speech API. If speech to ISL, the text will be shown using animated ISL GIFs and fingerspelling. If a pre-recorded video is used, outputs will be created.

A two way communication dashboard will be included. This will consist of a Staff section and Patient section. They will be linked using a QR code with the same session ID. Flask SocketIO will allow for real time communication between the sections. The system will have alternative methods of real time communication using HTTP polling and Firebase Realtime Database. If the patient uses ISL gestures, the text will be sent to the staff section. If the staff member uses speech, the ISL visuals will be sent to the patient section. This will allow users to communicate together.

### 4.7 Main Processing

All functions of SIGNORA will be integrated into one system. The input will be processed by the input component. Computer vision will be used through MediaPipe Tasks Vision. Landmarks will be extracted by MediaPipe Tasks Vision and gestures will be recognized using ONNX Runtime Web. The frontend will be created using HTML5, CSS3, JavaScript and Bootstrap. The backend will be created using Python Flask. Language processing will be performed using spaCy and complex sentences will be generated using LLM from the OpenRouter gateway. The language processing will be performed using Flask SocketIO. The system will contain several layers of processing. These layers will be input, computer vision, landmarks extraction, gestures recognition, language processing, Agentic AI and output. All four communication approaches will contain common recognition and translation components. The adaptive communication approach will enable real-time communication using WebSockets, HTTP and Firebase. It will allow edge gesture recognition, multilayered NLP and Agentic AI sentence generation.

## 5. IMPLEMENTATION

### 5.1 Dataset Preparation & Sign Vocabulary

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

SIGNORA uses a bespoke dataset of over 5,400 ISL samples. This dataset contains 4320 samples to train a model and 1080 samples to validate/test the model. Each gesture sample can have up to 60 frames and extracts 198 features including 174 spatial and 24 kinematic features. Augmentation techniques have been implemented including spatial rotation, scaling, horizontal flipping, temporal scaling and adding noise.

The system includes a visual sign vocabulary which contains around 90 ISL word and phrase level GIFs and 26 alphabetical fingerspelling images. These are used to show the user the recognised signs and create an ISL representation.

### 5.2 Real-Time ISL to Speech

The real-time ISL to speech functionality takes live ISL gestures and converts them into meaningful text. The video stream is captured from the user's webcam and the MediaPipe Tasks Vision is used to detect hand, upper body pose and specific landmarks. The landmarks are sent to the Landmark Detection and Feature Normalisation Pipeline. This normalises the data and handles any missing landmarks. It also extracts the spatial and kinematic features.

The landmark sequence is then sent to a bespoke Bidirectional Long Short-Term Memory (Bi-LSTM) with Multi-Head Self Attention model. The model will detect what gesture the user is making and output a prediction and confidence value. The model has trained and export as an ONNX file. This runs in the browser using ONNX Runtime Web.

The gesture labels are stored in a gloss buffer. This helps to filter out any false predictions by validating the confidence values and applying temporal filtering. When the sentence is complete the linguistic processing engine will take the gloss sequence and produce meaningful English text. Finally the sentence is converted into speech.

### 5.3 Real-Time Speech to ISL

The real-time speech to ISL converts speech into an ISL representation. The system records the speech from the microphone and uses the Web Speech API. The text is then sent to the Natural Language Processing (NLP) module where spaCy is used to extract key linguistic features including subjects, objects, verbs, tense, negation, questions and proper nouns.

The sentence is organised in the order defined by the system. This follows the Time-Subject-Object-Verb-Negation-Question (TSOVNQ) order. The words and phrases are then matched to the visual dictionary of ISL. If there is a full phrase that can be shown then this is prioritised over individual words. If there isn't then individual words are shown. Finally if there is no matching word animation then alphabetical fingerspelling is implemented.

### 5.4 Single Video to Speech

The single video to speech translates an ISL video into meaningful text. The user uploads a video or supported GIF. The media is checked for validation and the video frames are extracted. Using browser based techniques the video frames are extracted using an offscreen canvas.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

MediaPipe Tasks Vision is used for extraction of the necessary landmarks. The landmarks go through the same feature normalisation and sequence creation process as the real-time gestures. This sequence then goes to the trained Bi-LSTM model with Multi-Head Self Attention using ONNX Runtime Web.

The recognised gloss sequence is sent to the linguistic processing engine. This will produce meaningful English text. The text is then displayed to the user and converted into speech. This uses the same recognition and linguistic processing as the real-time functionality.

**5.5 Multiple Clip to Grammar Correct Sentence**

The multiple clip functionality allows the user to select multiple short videos and combine them into one full English sentence. The user can upload multiple clips and organise them in the correct order.

Each clip goes through the video frame extraction, landmark detection, feature normalisation and recognition process. The recognised gloss sequences are then combined in the correct order. This is then sent through the linguistic processing engine. This will analyse and recreate the words into a meaningful English sentence.

If the sentence is complex or unclear then an optional Large Language Model can be utilized to help recreate the sentence. The sentence will then be displayed. This functionality allows SIGNORA to process multiple short sign language inputs.

**5.6 Agentic AI & Linguistic Processing**

SIGNORA has an Agentic AI assisted linguistic processing engine. This will help create meaningful English sentences from the recognised signs. The system uses a well orchestrated process to process the input. It uses a combination of deterministic processing, NLP and optional processing from Large Language Models. The correct stage will be selected depending on what the input is.

The following steps are performed:

1. Input & Gloss Extraction: Recognised gesture labels are extracted and are sent to linguistic processing as gloss tokens.

2. Linguistic Orchestration: The controller will determine the sequence of the processing. It will send the recognised glosses to the proper step depending on whether mapping of the input is possible or additional processing is required.

3. Deterministic Mapping: At first, the system will try to perform the mapping of the gloss sequence to the predefined phrase/sentence mappings.

4. NLP Processing: In case additional processing is required then the SpaCy library will be used for performing analysis of the sentence. The relevant parts of speech will be extracted and used in the construction of the sentence. This step helps in converting the sign-based glosses into readable English sentences.

5. LLM Processing: In case the gloss sequence is complex and/or incomplete, then an optional Large Language Model will be used. This model will be using the OpenRouter API. This step will help to

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

interpret the gloss sequence and create the sentence out of it. It provides additional linguistic processing.

6. Output: Lastly, the output sentence is checked and converted into speech.

Overall the process can be summarised as:

Recognised Glosses → Linguistic Orchestration → Deterministic Processing → NLP with spaCy → Optional LLM Processing → Output → Meaningful English Sentence → Speech Output

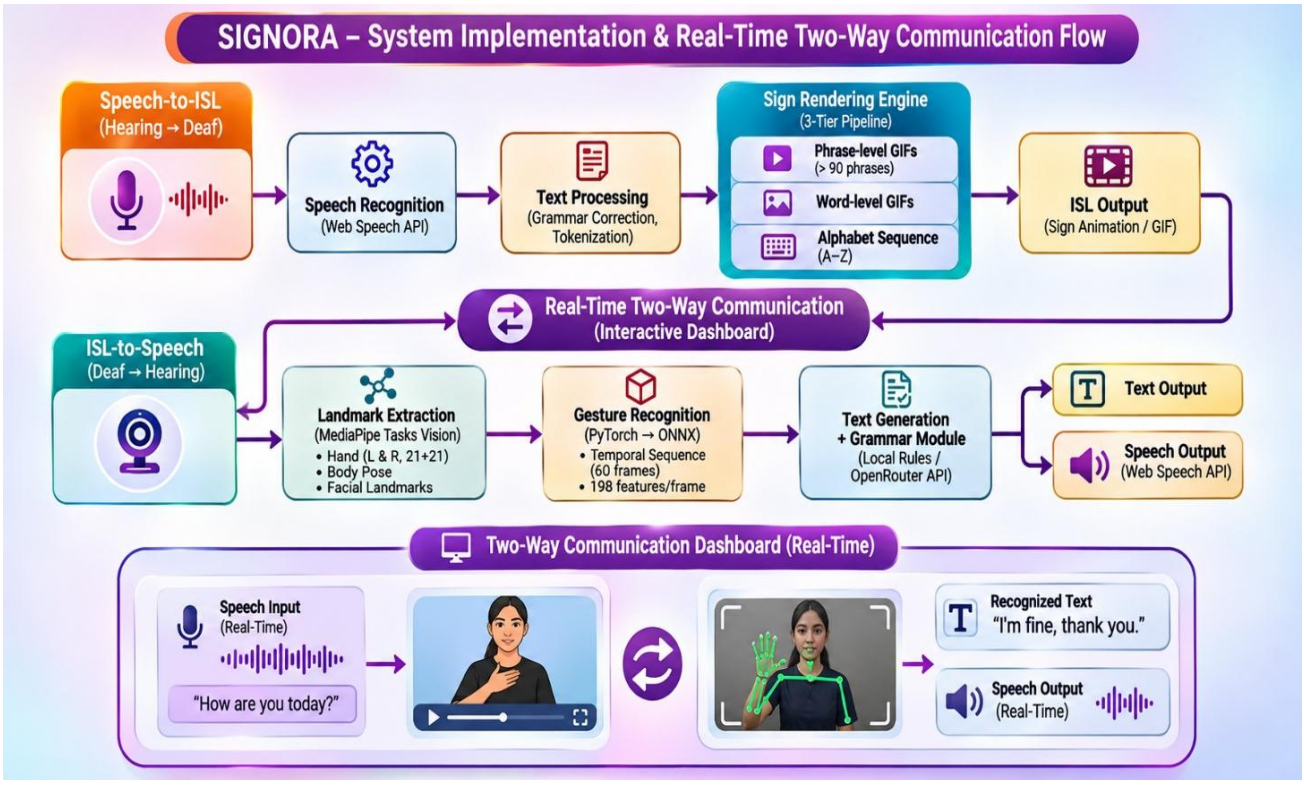
This has been implemented throughout the system. The real-time ISL to speech, single video translation and multiple clip sentence recreation all use this architecture. This creates a well structured system to process the linguistic features. The architecture is an orchestrated process and should not be viewed as independent agents performing tasks.

### 5.7 Two-Way Communication Dashboard

SIGNORA has a Two-Way Communication Dashboard. This allows two users to communicate with each other. The system pairs the communication using QR codes. One user will generate a QR code and the other will scan this. The systems are then paired using a unique session ID.

Once pairing is completed the users can start to communicate with each other using the two communication methods. If speech is used then speech recognition is performed. This will convert the text into an ISL representation. This is synchronised with the other system. If the user is using ISL gestures then the gesture recognition will identify the gestures. These will then be converted into meaningful text and speech. This will be shared with the other system.

**Figure 2: Real Time Communication flow in Signora**




The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

The system uses Flask and Flask-SocketIO to enable the backend communication. Events such as speech_to_isl and sign_to_speech will be used to identify what type of communication is being used. There is also option to use HTTP Polling instead. Firebase Realtime Database can be used to sync the sessions.

This allows continuous information to be shared between the users. The output will be synchronised and the four translation functions are brought together.

### 5.8 Output

Finally, the last stage of SIGNORA generates outputs from the various modules of the system. Based on the type of input used, meaningful English text, speech, ISL GIFs or alphabetical fingerspelling will be generated. The outputs will be presented via the web interface.

The frontend of the system is developed using HTML5, CSS3, JavaScript and Bootstrap 5. The backend is implemented using Python 3.11, Flask and Flask-SocketIO. MediaPipe Tasks Vision is used to detect the landmarks. The Bi-LSTM with Multi-Head Self Attention model is executed using ONNX Runtime Web. The linguistic processing module is developed using spaCy. The speech recognition and synthesis is performed using the Web Speech API. The optional processing module uses the Large Language Models with the help of OpenRouter. Firebase Realtime Database is used to give an alternative method of session synchronization.

It implies that SIGNORA is capable of performing real-time bidirectional translation, single and multiple clip video translation, AI assisted sentence recreation and two-way communication.

## 6. RESULTS AND DISCUSSION

### 6.1 Results

The four features that have been integrated in our proposed SIGNORA system are:

- ISL to Speech Real Time Conversion: The ISL to Speech module uses the MediaPipe and Bi-LSTM along with Multi-Head Self-Attention model to generate meaningful text and speech from webcam gestures.
- Speech to ISL Real Time Conversion: The Speech to ISL module uses speech recognition and NLP to generate animated ISL from the spoken word.
- Single Video Pre-recorded Conversion: Single video conversion into sentences in English.
- Multiple Clip Videos to Grammar Correct Sentence Conversion: Multiple video conversion into a grammar correct sentence.

Agentic AI-assisted Linguistic Engine relies upon deterministic mappings, spaCy and optional LLM processing for the purpose of generating sentences. The Two-way Communication Dashboard synchronizes the process via QR code pairing. The system works with up to 60 frames with 198

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

features each. The number of parameters in the model is about 245,000, and the size of the ONNX model is 1.92 MB.

### 6.2 Discussion

These findings show that the proposed model is able to enable two-way communication between users through the use of real-time and pre-recorded ISL. Various videos can be analyzed via the multiple clips function. Linguistic engine enables improved sentence formation with the help of deterministic mappings, NLP, and LLM.

The two-way communication dashboard provides a complete system solution. Landmark extraction via the browser makes sure that less video data gets transmitted in its raw format on the back end and keeps the information private. The model is trained at very high levels of accuracy. SIGNORA shows that computer vision, deep learning, speech processing, and NLP all can work together for enabling communication through ISL.

## 7. CONCLUSION

SIGNORA includes four properties such as Real-time ISL to speech conversion, Real-time Speech to ISL conversion, Single video to speech conversion, and Multiple clip video to grammatically correct sentence conversion. MediaPipe is used as a solution for landmark detection. Sign Language Neural Orchestration & Real-Time Adaptation uses the Bi-LSTM model with Multi-Head Self-Attention along with ONNX Runtime Web and NLP to convert sign language gestures and speech into information.

Sentence Reconstruction is done with the help of Agentic AI assisted language engine, and Two-Way Communication Dashboard is used for real-time interaction using QR code pairing. Landmark detection and model inference are performed using the browser. Therefore, SIGNORA can be seen as an assistive communication solution.

The system recognizes nine gesture classes with controlled signing and network connectivity. With further improvements, the system could recognize more gestures, could become more robust through gesture recognition in various signing conditions, could incorporate real-time ISL translation and could enhance multilingual and offline capabilities. This would improve SIGNORA and show its potential usage in real-world scenarios.

## 8. AUTHOR(S) CONTRIBUTION

The writers affirm that they have no connections to, or engagement with, any group or body that provides financial or non-financial assistance for the topics or resources covered in this manuscript.

## 9. CONFLICTS OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

## 10. PLAGIARISM POLICY

All authors declare that any kind of violation of plagiarism, copyright and ethical matters will take care by all authors. Journal and editors are not liable for aforesaid matters.

## 11. SOURCES OF FUNDING

The authors received no financial aid to support for the research.

## 12. REFERENCES

S. Katoch, V. Singh, and U. S. Tiwary, "Indian Sign Language Recognition System Using SURF with SVM and CNN," *Array*, vol. 14, Art. no. 100141, 2022. DOI: 10.1016/j.array.2022.100141.

P. Malhotra, C. K. Maniar, N. V. Sankpal, and H. R. Thakkar, "Indian Sign Language Recognition System Using OpenPose," *i-manager's Journal on Computer Science*, vol. 7, no. 2, pp. 43–50, 2019. DOI: 10.26634/jcom.7.2.15993.

B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K.-H. Park, and J. Kim, "An Integrated MediaPipe-Optimized GRU Model for Indian Sign Language Recognition," *Scientific Reports*, vol. 12, Art. no. 11964, 2022. DOI: 10.1038/s41598-022-15998-7.

P. Rawat, P. Kumar, V. K. Tamta, and A. Kumar, "A Comprehensive Approach to Indian Sign Language Recognition: Leveraging LSTM and MediaPipe Holistic for Dynamic and Static Hand Gesture Recognition," *EAI Endorsed Transactions on AI and Robotics*, vol. 4, 2025. DOI: 10.4108/airo.8693.

D. R. Kothadiya, C. M. Bhatt, T. Saba, A. Rehman, and S. A. Bahaj, "SIGNFORMER: DeepVision Transformer for Sign Language Recognition," *IEEE Access*, vol. 11, pp. 4730–4739, 2023. DOI: 10.1109/ACCESS.2022.3231130.

R. Singhal, J. Gupta, A. Sharma, A. Gupta, and N. Sharma, "Indian Sign Language Detection for Real-Time Translation Using Machine Learning," *arXiv* preprint arXiv:2507.20414, 2025.

N. Goel, L. Rani, and P. K. Sarangi, "Hybrid LSTM-Transformer Architecture with Generative AI for Enhanced Indian Sign Language Recognition," Research Square preprint, 2026. DOI: 10.21203/rs.3.rs-8827413/v1.

H. Monga et al., "Speech to Indian Sign Language Translator," in *Recent Trends in Intensive Computing*, IOS Press, 2021. DOI: 10.3233/APC210172.

A. R. P. Adithyaraaj et al., "Indian Sign Language (ISL) Translator: AI-Powered Bidirectional Translation System," *International Journal for Research in Applied Science & Engineering Technology*, 2025. Complete publication details to be verified.

M. Ahmed Aslam, A. Mohamed Aabidh, and N. Sabiyath Fatima, "Real-Time Indian Sign Language Recognition & Multilingual Sign Generation," in *Proceedings of the Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2025. DOI: 10.1109/ICAISS61471.2025.11041851.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License

Rumana Anjum, Aravind S Gudi, Bindushree D S, Chethan K & Harsha M R (2026). Signora: An Ai Based Multimodal Speech Translation. International Journal of Multidisciplinary Research & Reviews, 5(10), 75-90.

A. Joshi, S. Agrawal, and A. Modi, "ISLTranslate: Dataset for Translating Indian Sign Language," in *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10466–10475, 2023. DOI: 10.18653/v1/2023.findings-acl.665.

Z. Liang, H. Li, and J. Chai, "Sign Language Translation: A Survey of Approaches and Techniques," *Electronics*, vol. 12, no. 12, Art. no. 2678, 2023. DOI: 10.3390/electronics12122678.

R. J., S. Sahani, and P. K. Yadav, "A Review on Advances in Indian Sign Language Recognition: Techniques, Models, and Applications," *International Journal for Research in Applied Science & Engineering Technology*, vol. 12, no. 11, pp. 2332–2338, 2024. DOI: 10.22214/ijraset.2024.65539.

R. Damdoo and P. Kumar, "An Integrative Survey on Indian Sign Language Recognition and Translation," *IET Image Processing*, vol. 19, no. 1, 2025. DOI: 10.1049/ipr2.70000.

L. Negi, "Indian Sign Language Recognition Systems Using Deep Learning Models," *International Journal of Progressive Research in Engineering Management and Science*, vol. 5, no. 11, pp. 1290–1294, 2025.

The work is licensed under a Creative Commons Attribution Non Commercia 4.0 International License