# Signspeak Speech To Sign Language Convertor

Joshua Sheron D
Artificial Intelligence & Data Science
Sri Venkateswara College of Engineering
Chennai, India

Rishab S
Artificial Intelligence & Data Science
Sri Venkateswara College of Engineering
Chennai, India

Siddhartha P
Artificial Intelligence & Data Science
Sri Venkateswara College of Engineering
Chennai, India

Uma P
Computer Science
Sri Venkateswara College of Engineering
Chennai, India

## *ABSTRACT*

This project aims to develop an advanced communication system called SignSpeak, designed to bridge the gap between spoken language and sign language, thereby promoting greater inclusivity for individuals with hearing and speech impairments in public interactions and official functions. In many real-world scenarios, the absence of sign language interpreters poses a significant communication barrier, often excluding hearingimpaired individuals from accessing essential information. The existing solutions involve stitching together video clips, based on the input text. These solutions are not scalable and moreover the actions rigid and lacks real time feedback loop.

SignSpeak addresses this challenge by leveraging Automatic Speech Recognition and Natural Language Processing to capture and semantically interpret real-time speech input. The processed text is then translated into sign language using a Sign Language Generation model, which is rendered through an expressive virtual avatar capable of conveying sign language gestures accurately and naturally. By automating the entire speech-to-sign language conversion pipeline, the system eliminates the dependency on human interpreters, ensuring scalability, cost-effectiveness, and consistent accuracy across diverse environments such as government offices, hospitals, educational institutions, and public events. Ultimately, SignSpeak offers a robust, real-time, and scalable solution that enhances communication accessibility, promotes digital inclusivity, and empowers the hearing-impaired community by enabling equal participation in society through improved access to spoken information. ***A live demonstration of the system is available at: https://speech-to-sign-language.vercel.app***

***Keywords***— *Indian Sign Language, Speech to Sign Language,*

*ISL, Avatar-based Sign Language, Text to Sign Language, Speech Recognition, Digital Accessibility, Hearing Impairment Support, Human Computer Interaction*

## I. INTRODUCTION

SignSpeak addresses the communication barriers faced by the deaf community by creating a real-time speech-to-sign language converter. While spoken communication is dominant in public domains, the lack of interpreters often excludes hearing-impaired individuals. SignSpeak integrates speech recognition, natural language processing, and avatar-based rendering to automatically convert spoken input into animated Indian Sign Language. This provides accessibility, independence, and inclusion for the deaf in real-world interactions.

## II. RELATED WORKS

Speech-to-sign language translation has gained significant momentum due to advancements in computer vision, deep learning, and natural language processing.

Miah et al. proposed GmTC, a model that integrates Graph Convolutional Networks (GCNs) and Multi-Head Self-Attention (MHSA) to handle multicultural sign language recognition effectively. It achieved impressive results across datasets like

KSL, ASL, and BSL, though it struggled with low-resource sign languages [1]. Paneru et al. developed an AI-driven ASL-to-Nepali translator that leverages pre-trained CNN models such as ResNet50 and VGG16 for gesture recognition and integrates gTTS for realtime speech synthesis. Their system achieved over 99% accuracy in controlled test scenarios [2]. Natarajan et al. introduced an end-toend deep learning framework combining MediaPipe for pose estimation, CNN-BiLSTM for gesture recognition, Neural Machine Translation (NMT), and GANs for video generation. Though powerful, the approach demands considerable computational power [3]. Talaat et al. presented a YOLOv8-based Arabic Sign Language (ArSL) avatar system, which achieved 99.4% accuracy. Despite its high performance, the system was limited in recognizing nuanced facial expressions and dynamic hand gestures [4]. Luqman proposed a Dynamic and Accumulative Motion Network (DMN + AMN) architecture for isolated sign recognition. By encoding gestures into a single image using the AVM method, the system improved signerindependent accuracy, although continuous sign translation remained a limitation [5]. Maruyama et al. proposed a Multi-Stream Neural Network (MSNN) for word-level sign recognition that fuses global body motion, detailed hand/face inputs, and skeleton data. It improved recognition accuracy across benchmarks but faced difficulties with fast or overlapping gestures [6]. Faisal et al. designed the Saudi Deaf Companion System (SDCS) that integrates real-time sign recognition, avatar rendering, and speech synthesis for two-way communication. While effective, it is limited to Saudi Sign Language [7]. Abbas et al. developed a multimodal dataset for ArSL, combining synchronized video, audio, and text of religious content. The dataset ensures accurate translation but is limited in scope to structured domains like sermons [8]. Unlike these systems, SignSpeak integrates self-supervised speech recognition (Wav2Vec 2.0), syntax-adaptive NLP, and expressive 3D avatar animation, focusing on Indian Sign Language (ISL) for real-time translation in public settings.

## III. DATASETS

The project employs the Kaggle Indian Sign Language dataset, which comprises annotated video recordings of various ISL gestures, including alphabets, numerals, and commonly used terms. These video samples are processed using MediaPipe Holistic to extract three-dimensional skeletal key points corresponding to joints such as wrists, elbows, and fingertips. The resulting landmark data provides the foundation for training the system's gesture classification model and driving the avatar's animation engine with accurate motion patterns.

For the speech recognition component, the Wav2Vec 2.0 model is fine-tuned using the LibriSpeech corpus, a widely used ASR benchmark dataset composed of high-quality English speech derived from public domain audiobooks. The diversity in speaker accents, speech rates, and acoustic conditions within LibriSpeech enables the ASR model to achieve robust and accurate transcription performance in varied real-world scenarios. This fine-tuning process significantly enhances the system's ability to convert spoken input into text, which is subsequently translated into sign language.

## IV. PROPOSED METHODOLOGY

The system consists of three core modules:

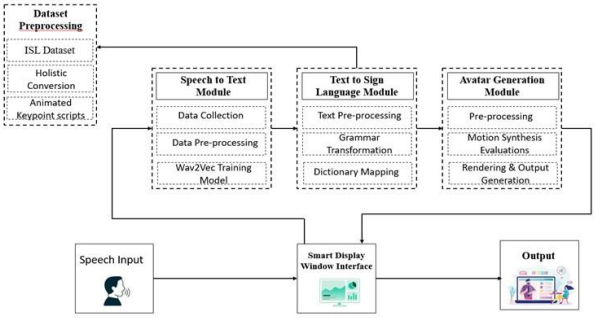


Figure 1 – Architecture Diagram

### *A. Automatic Speech Recognition (ASR):*

The process begins with real-time Speech Input, typically captured using a high-quality microphone. The choice of microphone and the acoustic environment play a significant role in the quality of the initial audio signal. To ensure optimal performance of the ASR model, the raw audio stream undergoes a series of essential preprocessing techniques. These techniques are designed to enhance the signal-to-noise ratio and standardize the audio input. Background noise reduction algorithms are employed to filter out unwanted ambient sounds, such as hums, static, or distant conversations, which could interfere with the speech recognition process. Volume level normalization ensures that the amplitude of the audio signal is consistent over time and across different speakers, preventing variations in loudness from negatively affecting the model's performance.

Furthermore, techniques for silence removal and elimination of other irrelevant acoustic elements (like coughs or lip smacks) are applied to focus the model's attention solely on the meaningful speech segments. These preprocessing steps are critical in providing a clean and consistent acoustic signal, which is essential for achieving high accuracy in the subsequent speech recognition task. For the core task of converting this pre-processed audio into text, the SIGNSPEAK system leverages the power of Wav2Vec 2.0, a stateof-the-art, self-supervised Automatic Speech Recognition (ASR) model pioneered by Facebook AI. What makes Wav2Vec 2.0 particularly powerful is its ability to learn rich and contextualized representations of speech sounds and patterns from vast amounts of unlabeled audio data. This self-supervised pre-training allows the model to capture the intricate acoustic characteristics of human speech without requiring extensive manual transcriptions. The architecture of Wav2Vec 2.0 incorporates convolutional neural networks (CNNs) to effectively extract latent acoustic features directly from the raw audio waveforms.

The convolutional layers learn hierarchical representations of sound, capturing both low-level acoustic details and higher-level phonetic information. Following the feature extraction, the model is further refined through fine-tuning with labeled transcriptions. This supervised training on specific language data (in this case, likely English) allows the model to map the learned acoustic features to their corresponding textual units (phonemes, words). This two-stage training process – self-supervised pre-training followed by supervised fine-tuning – enables Wav2Vec 2.0 to achieve remarkable accuracy, even in challenging real-world scenarios characterized by noisy environments or diverse accents and speech styles. Its robustness to variations in pronunciation and acoustic conditions makes it highly suitable for deployment in practical applications where consistent and reliable speech-to-text conversion is paramount. The immediate output from this ASR module is a verbatim transcription of the spoken English sentence. This textual representation captures the spoken words in their original sequence. However, it's important to recognize that spoken language often contains elements that are not directly or efficiently represented in sign language. These can include fillers (e.g., "um," "uh," "like"), auxiliary verbs (e.g., "is," "are," "have"), and casual speech patterns (e.g., informal phrasing, redundant words). Directly translating such verbatim transcriptions into sign language would likely result in cumbersome, unnatural, and potentially confusing signed output for sign language users.

Therefore, to ensure a meaningful and linguistically appropriate translation into the visual domain of sign language, the raw text necessitates further processing. This crucial step involves aligning the textual representation with the grammatical structures and pragmatic conventions of the target visual language, which is the primary responsibility of the subsequent module in the pipeline: linguistic simplification and text processing. This next stage refines the textual output of the ASR module, preparing it for the mapping to sign language glosses and ultimately, the generation of natural and understandable sign gestures by the avatar.

### *B. Text Processing and Simplification:*

The module serves as the essential bridge between the grammatical structures of spoken English and the distinct linguistic characteristics of sign languages, specifically Indian Sign Language (ISL). The fundamental rationale for this module stems from the inherent syntactic and semantic the divergences between these two forms of communication. Spoken English typically adheres to a subject-verb-object (SVO) structure, relies heavily on function words (e.g., articles, prepositions, auxiliary verbs), and incorporates complex grammatical inflections. In contrast, ISL often employs a more flexible word order, frequently utilizes topic-comment structures, and tends to omit grammatical elements considered redundant in the visual-gestural modality. Consequently, a direct, unmediated translation from English to ISL would likely yield outputs that are grammatically inaccurate, semantically ambiguous, and ultimately challenging for ISL users to comprehend (Klima & Bellugi, 1979). The primary objective is to mitigate this structural incongruity by performing a series of linguistic simplification and restructuring operations on the transcribed English sentence, thereby preparing it for accurate and meaningful translation into sign language.

The initial phase of this module involves a series of text normalization procedures aimed at creating a consistent and errorfree input for subsequent processing. This includes cleaning the raw text by removing punctuation marks, which hold less significance in sign language representation. Furthermore, all characters are converted to a lowercase format to ensure uniformity and simplify lexical lookup. Critically, this stage also addresses potential inaccuracies arising from the ASR module by correcting common transcription errors, such as the misidentification of homophones (words with identical pronunciation but different meanings, e.g., "see" and "sea") or mispronounced words.

Following these normalization steps, the system employs advanced NLP tools to perform a comprehensive analysis of the sentence's linguistic

structure. Libraries such as spaCy (Honnibal et al., 2020) and Stanza (Qi et al., 2020) are instrumental in this process, enabling the system to perform Part-of-Speech (POS) tagging, which annotates each word with its grammatical category (e.g., noun, verb, adjective). Moreover, these tools facilitate dependency parsing, which identifies the syntactic relationships between words in the sentence, revealing the underlying grammatical structure and dependencies. By identifying key linguistic elements such as the subject, verb, and object, the system gains a deeper understanding of the sentence's semantic core.

Leveraging this structural analysis, the system then proceeds with linguistic simplification and restructuring. This involves strategically reordering or eliminating words that are deemed redundant or grammatically superfluous in the context of ISL. For instance, a declarative English sentence like "She is going to the market" undergoes a transformation to a more concise form, such as "She go market." This simplification involves the removal of the auxiliary verb "is" and the article "the", as these grammatical elements are typically not explicitly represented in ISL. This process results in a more telegraphic style of language that aligns better with the visual-spatial nature of sign communication. The outcome of this stage is a concise and semantically accurate version of the original sentence, specifically optimized for efficient and accurate translation into sign language glosses.

After the sentence has been syntactically and semantically simplified, the next step is to convert the processed text into a structured representation that corresponds to specific signs. This is achieved through the generation of sign glosses. A gloss is a written representation of a sign and serves as an intermediary between the spoken/written word and its signed equivalent. Glosses are typically written in uppercase (e.g., BOY, GO, MARKET) and act as symbolic cues that trigger specific gestures in the sign language domain.

The simplified sentence from the previous module is tokenized into individual words, each of which is looked up in a predefined gloss dictionary. This dictionary maps common English words to their corresponding signs in Indian Sign Language. If a word is found, its gloss equivalent is added to the output sequence. In cases where a word is not found due to it being a proper noun, a rarely used term, or a slang word the system either applies fingerspelling or chooses the closest semantically similar word that does exist in the dictionary.

### *C. Avatar Rendering:*

The third module involves the Avatar-Based Gesture Rendering module, where the processed linguistic information is transformed into a visual representation of sign language through a 3D animated avatar. The avatar functions as the system's visual output, conveying sign language through the dynamic articulation of hand configurations, facial expressions, and body postures.

The initial step in this process is the creation of the avatar itself, typically achieved through the utilization of 3D modeling software "Blender". This software facilitates the design of a virtual robot with a more detailed and anatomically plausible form. A critical component of this avatar is the implementation of a skeletal rig. This underlying digital framework establishes a hierarchical structure of interconnected joints, enabling the avatar to execute a wide range of fluid and nuanced movements that are essential for the accurate depiction of sign language. The animation of the avatar is directly driven by a custom-built dataset that encodes the skeletal movements required for accurate Indian Sign Language (ISL). For each gesture in the dataset covering individual alphabets and a predefined vocabulary set skeletal movement sequences are generated using MediaPipe Holistic, which performs full-body pose estimation. MediaPipe tracks and records the three-dimensional coordinates of critical anatomical landmarks, such as hand joints, elbows, shoulders, and head positions, while a signer performs each gesture. This process results in a rich temporal dataset that captures the spatial structure and dynamic flow of signs.

The collected motion data forms a comprehensive key point dataset representing each sign's motion trajectory. These skeletal sequences are then mapped to a virtual 3D avatar's rig, ensuring the avatar can mimic human-like ISL gestures. Using inverse kinematics, the motion synthesis engine creates smooth transitions between signs while preserving the semantic and grammatical integrity of the gesture sequence. This ensures that the final animated output is both intelligible to ISL users and visually coherent. The avatar's movement is dynamically generated based on real human motion data, enhancing realism and authenticity in sign representation.

The motion data derived from these recordings is subsequently mapped onto the corresponding joints of the 3D avatar. This mapping procedure ensured a precise correspondence between the movements of the human signer and the virtual signer. The resulting animation sequences for each sign were stored as discrete animation clips within the avatar's motion library. Upon receiving a sequence of linguistic input, the rendering engine retrieves and orchestrates the playback of these animation clips. The precise timing and synchronization of these clips are crucial for replicating the natural rhythm and flow of sign

language. Consequently, the system is designed to generate smooth and continuous transitions between signs, avoiding jerky or unnatural movements. In addition to hand movements, the avatar's animation encompasses facial expressions and head movements, which constitute integral components of sign language communication.

## V. RESULTS AND DISCUSSION

The application is designed to generate corresponding sign language output from user-provided linguistic input. This input can be delivered to the system in one of two ways: either as live spoken input, captured as a real-time audio stream, or as direct textual input, entered via a keyboard or other text-based interface.

When the input is provided as live spoken audio, the system initially processes it using Wav2Vec for feature extraction. This crucial step transforms the raw audio signal into a detailed representation of its acoustic characteristics. Wav2Vec, a selfsupervised learning model, excels at capturing the nuances of speech, including phonemes, intonation, and rhythm. This transformation allows the system to effectively analyze and interpret the spoken words, converting them into a format suitable for further linguistic processing. The extracted acoustic features represent the spoken content in a way that preserves its linguistic information.

Alternatively, if the user provides direct textual input, this initial audio processing stage is bypassed. The system directly receives the text, which is assumed to already be in a format suitable for the subsequent linguistic analysis. This text input option allows for scenarios where the user might have pre-written sentences they want to translate into sign language, or if the system is integrated with other text-generating applications.

Once the spoken input is converted into a textual representation, it is processed by subsequent modules that perform linguistic simplification and mapping to a movement dataset. This dataset contains pre-recorded motion capture data of human signers performing various signs corresponding to words and phrases. The simplified text is used to index and retrieve the appropriate sequences of movements from this dataset.

Finally, these movement sequences are processed by the avatar generation module, which animates a virtual avatar to perform the corresponding signs. The avatar, created with a detailed skeletal structure, precisely replicates the movements stored in the dataset. The avatar's very articulated hands and natural body language are all carefully controlled based on the retrieved motion data to accurately and naturally represent the sign language translation of the original spoken input. The Output of the SIGNSPEAK system is the visual display of this avatar seamlessly performing sign language, effectively converting the live speech into a visually accessible form for sign language users.

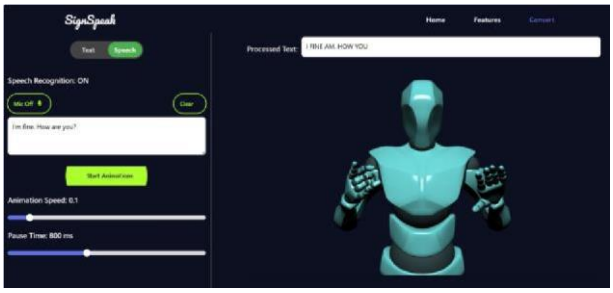


Figure 2 – Project live result from deployed website (https://speech-to-sign-language.vercel.app)

## VI. CONCLUSION AND FUTURE WORKS

In the evolving digital landscape, the need for effective speechto-sign language translation systems is becoming increasingly apparent, as it can significantly enhance communication accessibility for individuals who are deaf or hard of hearing. This project addressed this need by utilizing Wav2Vec for detailed acoustic feature extraction from spoken input, coupled with a series of modules to process and render the extracted features into a visual sign language representation. Through a carefully curated dataset of sign language movements, our system has demonstrated significant potential in translating spoken language into meaningful sign gestures, supporting the goal of enhanced communication access.

Looking ahead, there are several key areas for advancing this speech-to-sign language translation technology. First, to strengthen the system's ability to operate in real-world conditions, we can introduce audio augmentation and variation during training. This approach would simulate various acoustic environments, making the model more robust to background noise, varying microphone quality, and other real-world audio imperfections. Additionally, fine-tuning Wav2Vec configurations to capture even more granular acoustic features could further improve the system's sensitivity to the nuances of spoken language, leading to more accurate sign language translation.

Another avenue for improvement lies in refining the sign language generation process. This could involve incorporating a more sophisticated avatar with a wider range of facial

expressions and body language, or exploring different rendering techniques to enhance the fluency and naturalness of the avatar's signing. Moreover, adding contextual analysis to the translation process, enabling the system to better handle ambiguities and idiomatic expressions, could further enhance the system's ability to generate accurate and comprehensible sign language output.

Finally, to facilitate the scalability of this technology, future work can explore optimizing the model's processing speed without compromising accuracy, making it viable for real-time applications. This would enable practical implementation in areas requiring immediate communication, such as classrooms, public events, and customer service interactions. By continually refining these aspects, this project aims to provide a comprehensive and adaptable solution to the challenges posed by communication barriers between spoken and signed languages, with implications for more inclusive communication across a range of domains.

## REFERENCES

[1] Abu Saleh Musa Miah et al., "GmTC: A Graph Convolutional and Multi-Head Self-Attention Model for Multi-Cultural Sign Language Recognition," 2024.

[2] Biplov Paneru et al., "AI-driven ASL to Nepali Text and Speech Translator," 2024.

[3] B. Natarajan et al., "End-to-End Deep Learning for Sign Language Translation," 2022.

[4] Fatma M. Talaat et al., "YOLOv8-based Arabic Sign Language Avatar," 2024.

[5] Hamzah Luqman, "Dynamic and Accumulative Motion Networks for Isolated Sign Recognition," 2022. [6] Mizuki Maruyama et al., "Multi-Stream Neural Network for Word-Level Sign Language Recognition," 2024. [7] Mohammed Faisal et al., "Saudi Deaf Companion System," 2023.

[8] Samah Abbas et al., "Multimodal Dataset for Arabic Sign Language," 2024.