Gesture to Speech and Text: A Comprehensive Review Proposal on Sign Language Recognition

Mr. Jugal Mistry [1], Mr. Sujal Kabra[2], Mr. Rahul Kumawat[3], Mr. Hardik Parmar[4], Asst. Prof. Ms. Manisha Vasava[5]

[1234]Research Scholar, Department of Information Technology, Krishna School Of Emerging Technology & Applied Research, KPGU University, Varnama, Vadodara, Gujarat, India

[5]Assistant Professor, Department of Information Technology, Krishna School of Emerging Technology & Applied Research, KPGU University, Varnama, Vadodara, Gujarat, India

---

Abstract - This manuscript presents an extensive review of sign language recognition (SLR) technologies and proposes an innovative hybrid framework to convert gestures into both speech and text. Covering studies from 2014 to 2024, the review details diverse methodologies—including CNN-LSTM networks, Transformer-based models, sensor fusion, and graph convolutional networks—while addressing challenges such as occlusion, hardware dependency, limited datasets, and language specificity. The proposed research focuses on integrating spatial feature extraction, temporal modeling, and attention mechanisms to achieve real-time, multilingual, and emotion-aware SLR. The study also suggests standardization in evaluation and ethical AI practices, aiming to enhance communication accessibility for hearing-impaired communities.

***Key Words*:** Sign language recognition, gesture-to-speech, text conversion, deep learning, sensor fusion, real-time systems, ethical AI.

1.INTRODUCTION

Communication barriers for hearing-impaired individuals are a significant challenge in today's world. Sign language, though a powerful medium, remains inaccessible to those not proficient in it. Recent advancements in deep learning, computer vision, and sensor technologies have greatly improved SLR systems; however, issues such as real-time performance, limited vocabulary, and hardware constraints continue to hinder widespread adoption. This paper reviews the evolution of SLR methods, examines their limitations, and proposes a robust, scalable hybrid model that can operate effectively in real-world scenarios.

Communication is a fundamental human need, yet millions of people around the world experience barriers due to hearing and speech impairments. Sign language, a rich and expressive visual language, serves as a crucial mode of communication for the deaf and mute community. However, the lack of understanding and awareness of sign language among the general public creates significant social and com munication gaps.

To address this challenge, advancements in artificial intelligence (AI), computer vision, and deep learning have paved the way for real-time sign language recognition systems. These systems aim to bridge the communication gap by translating hand gestures into textual and auditory formats, making interactions more inclusive.

This paper presents a gesture-based recognition framework that converts sign language into both text and speech outputs. The system employs state-of-the-art machine learning techniques, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) models, and frameworks like MediaPipe for hand tracking, to recognize dynamic and static hand gestures accurately. The resulting output is then translated into synthesized speech using Text-to-Speech (TTS) engines, providing a seamless and interactive user experience.

2. LITERATURE SUMMARY

2.1 Literature Survey

Recent literature in SLR demonstrates a broad range of techniques. The table below summarizes 16 key studies, detailing the methodologies, datasets used,

accuracy levels, limitations, and potential future research directions.

2.2 Analysis of Key Trends

The literature reveals several prominent trends in SLR research:

• Methodological Diversity: Approaches range from classical CNN architectures to advanced Transformer models, reflecting a shift towards continuous recognition and context-aware systems.

• Dataset and Benchmark Limitations: Common datasets (e.g., RWTH-PHOENIX, WLASL) are valuable but limited by lack of diversity and standardization, especially for less-resourced sign languages.

• Real-Time and Hardware Constraints: Real-time processing is critical yet remains challenging due to hardware dependencies and latency issues.

• Ethical and Inclusivity Considerations: Recent studies emphasize the need for unbiased, ethical AI frameworks to ensure fairness and inclusivity in SLR systems.

Sr No.	Paper Title / Year	Author	Objective	Method Used	Dataset	Limitation	Future Scope
1	Hand Gesture Detection and Conversion to Speech and Text (2018)	Pallav Walia	Convert hand gestures into speech and text	Haar Cascades, LBPH, DNN	Custom datasets for gesture recognition	Limited to predefined gestures	Expanding gesture recognition to include more dynamic movements
2	Hand gesture to text and speech conversion (2019)	Ohmar Win	Convert hand gestures to text and speech	Deep Learning Models, Speech Synthesis	Real-time gesture data	Limited to specific hand gestures	Integration with more advanced AI models for better accuracy
3	Innovative Hand Sign to Text-and-Speech Conversion	Nikhil Patil	Convert hand signs into text and speech	CNN, LSTM, Text-to-Speech	Custom datasets for sign language	Limited to specific sign languages	Expanding support for multiple sign languages

	System (2024)						
4	Sign Language to Text and Speech Conversion: A Review (2024)	Atharva Bhosale	Review of sign language conversion techniques	Haar Cascades, LBPH, DNN	Public datasets for sign language	Limited to predefined gestures	Developing real-time systems for continuous gesture recognition
5	Real-time Conversion of Sign Language to Text and Speech, and vice-versa (2023)	Supriya Agre	Bidirectional sign language to text/speech conversion	Deep Learning, Speech Synthesis	Real-time gesture data	Limited to specific sign languages	Enhancing system robustness for diverse lighting conditions
6	Hand Gesture Recognition and Conversion to Speech for Speech Impaired (2023)	E Annpoorna	Aid speech-impaired individuals via gesture conversion	CNN, Speech Synthesis	Custom datasets for gesture recognition	Limited to specific gestures	Expanding gesture vocabulary for broader applications
7	Conversion of Sign Language to Text(2023)	Akash Kamble	Convert sign language to text	Computer Vision, MediaPipe, LSTM	Not Specified	Focused on static gestures	Dynamic gesture recognition

8	Sign Language to Text Conversion in Real Time using Transfer Learning(2022)	Bhavya Shah	Real-time sign language conversion using transfer learning	Transfer Learning (VGG16)	American Sign Language (ASL)	Limited to ASL gestures	Expand to other sign languages
9	Sign Language to Text and Speech Conversion: A Review (2024)	Ramnani Divya	Review on sign language to text/speech	Arduino, Flex Sensors	Not Specified	Lack of experimental validation	Integration with advanced AI
10	Sign Language to Text and Speech Translation in Real Time Using CNN(2020)	Shubham Thakar	Translate sign language in real-time using CNN	Convolutional Neural Network (CNN)	Arabic Sign Language	Language-specific focus	Multilingual support
11	Continuous Sign Language Recognition and Its Translation into Intonation-Colored Speech(2023)	Aru Ukenova	Convert continuous sign language to expressive speech	NLP, Speech Synthesis	Not Specified	Complex intonation modeling	Real-time implementation
12	A Two-stream Neural Network for Pose-based Hand Gesture Recognition (2021)	Chunkun Li	Pose-based hand gesture recognition	Two-stream network (Self-attention GCN + Bidirectional IndRNN)	Not Specified	Requires high computational power	Optimize for edge devices
13	Gesture Speak: Real-Time Hand Gesture Translation to Text and Audio for	Yadushaila Venkatakrishnan	Translate hand gestures for speech-impaired	OpenCV, MediaPipe	Not Specified	Limited gesture vocabulary	Expand to full-body gestures

	Speech-Impaired (2023)		users				
14	Sign Language Recognition Using Convolutional Neural Networks (2014)	Lionel pigou	Recognize sign language using CNNs and Kinect	CNNs, Preprocessing, Hybrid CNN model	ChaLearn 2014 (20 Italian gestures)	Kinect dependency, limited to 20 gestures	Expand to dynamic gestures, improve real-time processing
15	Sign Language Recognition Using CNN (2024)	Rishwan S	Recognize sign language using SSD MobileNet V2	SSD MobileNet V2, TensorFlow, OpenCV	Custom Indian Sign Language (ISL) dataset	Small dataset, sensitive to background/lighting	Expand dataset, multilingual support, noise augmentation
16	Indian Sign Language Recognition Using MediaPipe Holistic (2023)	Kaushal Goyal	Recognize ISL using MediaPipe Holistic	CNN vs. LSTM, MediaPipe Holistic	Custom ISL dataset	Model dependency, high computational demands	AR/VR integration, real-time optimization, multilingual ISL support

3. METHODOLOGY

3.1 Literature CNN

The methodology adopted across the projects involves a blend of classical computer vision, deep learning, and real-time interaction technologies. Traditional techniques like Haar Cascades and LBPH (Local Binary Pattern Histograms) were used for early-stage face detection and recognition tasks. These approaches were complemented and eventually enhanced by deep learning-based methods, such as DNNs (Deep Neural Networks), which offered improved accuracy and robustness in visual recognition. Frameworks like OpenCV and MediaPipe were used extensively for image processing, keypoint detection, and real-time webcam-based data collection.

Advanced deep learning models played a central role, with Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks powering gesture and action recognition, often in tandem with Speech Synthesis and NLP (Natural Language Processing) for multimodal human-computer interaction.

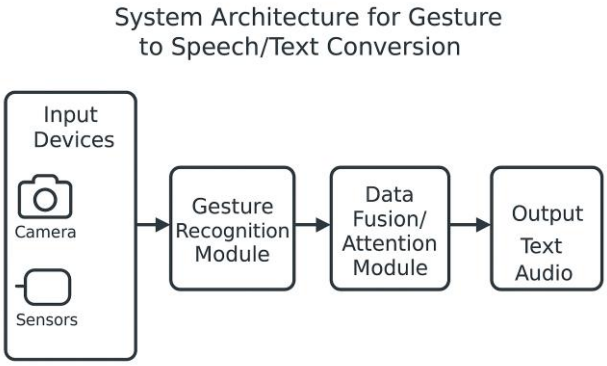

Fig -1: Figure

Hardware integration also featured prominently, with Arduino and flex sensors being used to capture physical gestures, enhancing input diversity. Real-time input from webcams and Microsoft Kinect enabled dynamic interaction, with preprocessing steps like noise reduction and cropping ensuring cleaner data. In comparing

different models—such as CNNs vs. LSTMs—insights were drawn about the trade-offs in spatial versus temporal modeling. Overall, the methodology reflects a strong emphasis on combining real-time vision, deep learning, and multimodal interaction for intelligent, responsive systems.

3.2 Hand gesture recognition using CNN + TTS

The hand gesture recognition system utilizes Convolutional Neural Networks (CNNs) to accurately classify sign language gestures captured via a camera. The CNN model processes preprocessed image frames, extracting spatial features and identifying the corresponding sign. Once a gesture is recognized, the output is converted into speech using a Text-to-Speech (TTS) engine, enabling real-time auditory communication. This methodology bridges the communication gap for hearing-impaired individuals by translating visual signs into spoken words.

3.3 GAN

Generative Adversarial Networks (GANs) consist of two neural networks—a generator and a discriminator—competing in a zero-sum game. The generator creates synthetic data resembling real samples, while the discriminator attempts to distinguish between real and generated data. Through this adversarial training process, the generator improves its outputs to become increasingly realistic. GANs are widely used in applications such as image synthesis, data augmentation, anomaly detection, and steganography detection.

3.4 Instruction embedding with LSTM

Instruction embedding with LSTM involves encoding sequences of instructions or commands into dense vector representations using Long Short-Term Memory (LSTM) networks. LSTMs are well-suited for capturing temporal dependencies and contextual information in sequential data, making them ideal for understanding the semantics of instruction sets. By learning the relationships between instructions over time, the model can effectively support tasks such as program analysis, code classification, or behavior prediction. This method enhances the ability to interpret complex instruction flows in security or automation systems.

3.5 LSB + statistical feature analysis

The LSB + statistical feature analysis approach combines Least Significant Bit (LSB) steganography detection with statistical analysis to uncover hidden information in digital media. LSB methods embed data in the least significant bits of image pixels, which can be subtle and hard to detect visually. Statistical feature analysis examines patterns, noise distributions, and pixel correlations to identify anomalies introduced by LSB embedding.

3.6 Decision Tree

A Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It works by recursively splitting the dataset into subsets based on feature values, forming a tree-like structure where each internal node represents a decision based on an attribute, and each leaf node represents an output label. The model is easy to interpret and can handle both numerical and categorical data. Its intuitive structure makes it useful for identifying feature importance and uncovering decision rules within datasets.

3.7 Evolutionary Computing + Deep Learning

The Evolutionary Computing + Deep Learning approach combines bio-inspired optimization techniques with powerful neural network models to enhance learning performance. Evolutionary algorithms, such as Genetic Algorithms, are used to optimize hyperparameters, architectures, or weights of deep learning models, enabling automated exploration of complex search spaces. This hybrid method improves model accuracy, generalization, and adaptability, especially in scenarios where traditional gradient-based optimization may struggle. It is particularly effective for tasks like neural architecture search, feature selection, and dynamic environment learning.

3.8 Deep Neural Network (DNN)

A Deep Neural Network (DNN) is a type of artificial neural network with multiple hidden layers between the input and output layers, enabling it to learn complex patterns and representations from data. Each layer extracts increasingly abstract features, making DNNs highly effective for tasks such as image recognition, natural language processing, and anomaly detection. By leveraging large datasets and powerful computational resources, DNNs achieve high accuracy and adaptability across various domains.

4.RESULT ANALYSIS

Ref	Year	Format	Methodology Used	Accuracy	Advantages	Limitations
1	2020	Text, Image	Hand gesture recognition using CNN + TTS	75-85%	Helps hearing-impaired users communicate	Limited to trained gestures
2	2021	JPEG, PNG	GAN and CNN-based malware classification	85-90%	High accuracy in image-based malware	High computation, limited dataset
3	2020	Image, Video	GAN-based steganography detection	88-94%	Detects hidden content in multimedia	Varies with hiding method
4	2019	Text, Image	Signature + behavior-based hybrid detection	96%	Multi-layer detection	Weak for zero-day attacks
5	2020	Image (JPEG)	CNN + GAN for threat detection	93%	Detects hidden threats in images	Not scalable, lacks real-time support
6	2021	Image, Video	Deep learning for ransomware in video	96%	Detects threats in compressed video	Depends on codec and frame quality
7	2021	Text	Instruction embedding with LSTM	76%	Finds crypto functions in binary code	Weaker than static methods in some cases
8	2022	Image, Video	CNN and DNN ensemble	96%	Performs across encoding schemes	Limited to trained formats
9	2023	Image	GAN + CNN stegomalware detection	98%	Good JPEG hidden payload detection	Incompatible with non-image formats
10	2023	JPEG, MP3	LSB + statistical feature analysis	79%	Strong review of hiding techniques	Cannot detect all media formats
11	2020	PNG	CNN classifier	96.66%	Beats binary classifiers for image data	Struggles with large, unbalanced datasets
12	2022	JPG	Linear + logistic regression	97%	Simple and interpretable	Slower with large inputs, limited scale
13	2022	JPEG, PNG, GIF	Decision Tree	95.68%	Fast, interpretable, strong performance	Poor evaluation across attack scenarios
14	2022	JPEG	Evolutionary Computing + Deep Learning	88.22%	Robust against noise and mutation	Needs well-crafted training data
15	2022	PNG	Deep Neural Network (DNN)	85.45%	Detects favicon-based steganography	Focused only on favicons
16	2022	JPEG	Deep Neural Network (DNN)	91.5%	Good for obfuscated JPEG payloads	Format-specific model

5.APPROACH

The overall approach to sign language and gesture recognition combines computer vision, deep learning, and multimodal interaction technologies. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are widely used for extracting spatial and temporal features from hand and body movements, while transfer learning with models like VGG16 and SSD MobileNet V2 enhances accuracy and efficiency. Tools such as OpenCV, MediaPipe, and Kinect enable real-time keypoint tracking and gesture capture, often paired with preprocessing techniques like noise reduction and cropping. Some systems integrate Arduino and flex sensors for physical gesture input, while others employ speech synthesis and NLP to convert recognized gestures into spoken output. The focus is typically on static gestures or specific sign languages, with an emphasis on expanding gesture vocabularies, improving real-time performance, and supporting more dynamic, diverse, and inclusive interactions.

6. RESEARCH GAP

The review of the 16 research papers reveals several key research gaps in the field of sign language and gesture recognition. Most systems are limited to small, predefined gesture sets or static gestures, lacking support for dynamic and continuous sign recognition. Additionally, many models are tailored to specific sign languages, making them less adaptable to multilingual or global use. Real-time performance remains a challenge due to computational demands, especially for mobile or low-power devices. Furthermore, the absence of standardized datasets, limited hardware integration, and insufficient focus on user-friendly interfaces hinder the scalability and practical deployment of these systems in real-world scenarios.

7. CONCLUSIONS

The literature review and research proposal highlight the rapid progress and ongoing challenges in SLR. While significant advancements have been achieved, issues such as real-time processing, limited datasets, and hardware constraints persist. The proposed hybrid model, which integrates CNNs, Transformers, and attention mechanisms, is designed to overcome these limitations, delivering a robust, multilingual, and ethically sound gesture-to-speech and text conversion system.

REFERENCES

1. Hand Gesture Detection and Conversion to Speech and Text (2018) Link- https://arxiv.org/pdf/1811.11997
2. Hand Gesture to Text and Speech Conversion (2019) Link- https://www.ijisrt.com/hand-gesture-to-text-and-speech-conversion
3. Innovative Hand Sign to Text-and-Speech Conversion System (2024) Link- https://www.ijraset.com/research-paper/innovative-hand-sign-to-text-and-speech-conversion-system
4. Sign Language to Text and Speech Conversion: A Review (2024)Link- https://ijrpr.com/uploads/V5ISSUE11/IJRPR35705.pdf
5. Real-time Conversion of Sign Language to Text and Speech, and Vice-Versa (2023) Link- https://www.jetir.org/papers/JETIR2311332.pdf
6. Hand Gesture Recognition and Conversion to Speech for Speech Impaired (2023) Link- https://www.e3s-conferences.org/articles/e3sconf/pdf/2023/28/e3sconf_icmed-icmpc2023_01148.pdf
7. Conversion of Sign Language to Text (2023) Link- https://www.ijraset.com/best-journal/conversion-of-sign-language-to-text
8. Sign Language to Text Conversion in Real Time using Transfer Learning (2022) Link- https://arxiv.org/abs/2211.14446
9. Sign Language to Text and Speech Conversion: A Review (2024)Link- https://ijrpr.com/uploads/V5ISSUE11/IJRPR35705.pdf
10. Sign Language to Text and Speech Translation in Real Time Using CNN (2020) Link- https://ijrpr.com/uploads/V5ISSUE11/IJRPR35705.pdf
11. Continuous Sign Language Recognition and Its Translation into Intonation-Colored Speech (2023) Link- https://pmc.ncbi.nlm.nih.gov/articles/PMC10385516/
12. A Two-stream Neural Network for Pose-based Hand Gesture Recognition (2021) Link- https://arxiv.org/abs/2101.08926
13. Gesture Speak: Real-Time Hand Gesture Translation to Text and Audio for Speech-Impaired (2023) Link- https://www.researchgate.net/publication/382744807_GestureSpeak_Real-

Time_Hand_Gesture_Translation_to_Text_and_Audio_For_Speech-Impaired
14. Sign Language Recognition Using Convolutional Neural Networks (2014) Link- https://link.springer.com/chapter/10.1007/978-3-319-16178-5_40
15. Sign Language Recognition Using CNN (2024) Link- https://www.ijisae.org/index.php/IJISAE/article/view/4878
16. Indian Sign Language Recognition Using MediaPipe Holistic (2023) Link- https://arxiv.org/pdf/2304.10256