Real-Time Sign Language to Speech Spectacles Using Camera Recognition and TTS

JUSTIN JEEVA F
Department of Electronics and Communication
Sathyabama Institute of Science and Technology
Chennai,India
justinjeeva2005@gmail.com

GEOFFREY IVAN AROK G
Department of Electronics and Communication
Sathyabama Institute of Science and Technology
Chennai,India
geoffreyivanarokg@gmail.com

Dr.M R EBENEZAR JEBARANI
Department of Electronics and Communication
Sathyabama Institute of Science and Technology
Chennai,India
ebenezarjebarani.ece@sathyabama.ac.in

Abstract*—This paper surveys the advancements in real-time sign language interpretation and translation into speech, specifically focusing on systems that leverage camera recognition and Text-to-Speech (TTS) technologies. It explores the methodologies, system architectures, performance metrics, challenges, and future directions of these innovative solutions, which aim to bridge communication barriers for individuals with hearing and speech impairments. The integration of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), with advanced computer vision frameworks like MediaPipe for precise hand gesture recognition, is critically examined. Furthermore, the role of TTS engines and Large Language Models (LLMs) in generating natural-sounding, contextually accurate speech is highlighted. This survey synthesises insights from recent research, identifies key limitations, and proposes future work to enhance the robustness, accuracy, and accessibility of such real-time translation systems, paving the way for potential wearable applications like smart spectacles.

Keywords: Sign Language Recognition, Real-Time Translation, Camera Recognition, Text-to-Speech (TTS), Deep Learning, Convolutional Neural Networks (CNN), MediaPipe, Wearable Technology, Accessibility, Communication.

I. Introduction

The Communication is a fundamental human need, yet for individuals with hearing and speech impairments, traditional verbal communication presents a significant barrier. Sign language serves as their primary means of conveying thoughts, ideas, and emotions, relying on hand gestures, facial expressions, and body postures. However, a substantial communication gap exists between sign language users and the broader hearing community who are often unfamiliar with it. This gap can lead to social isolation, limited educational and employment opportunities, and reduced access to information.

The urgent need for automated solutions to bridge this divide has spurred significant research in real-time sign language interpretation. This survey focuses on systems designed for real-time sign language to speech conversion, particularly those utilising camera-based recognition and Text-to-Speech (TTS) technology. The vision for such systems extends to wearable devices like smart spectacles, offering a seamless and portable translation experience that could revolutionise daily interactions. These systems aim to eliminate the need for human interpreters, making communication more accessible, instant, and independent.

This paper provides a comprehensive overview of the current landscape of real-time sign language to speech translation systems. It delves into the methodologies employed for gesture capture and recognition, the machine and deep learning models utilised, and the processes involved in generating clear speech output. We also discuss the performance metrics achieved by these systems, highlight existing challenges, and outline promising future directions for research and development. By examining the core components of these systems, from data acquisition via cameras to the final speech synthesis, the paper aims to present a holistic view of the progress made. The discussion also extends to the limitations and challenges that persist, such as the need for larger and more diverse datasets, the complexities of continuous sign language translation, and the trade-offs between model accuracy and real-time performance on resource-constrained devices.

II. Background and Motivation

A. Computer vision in sign language Detection

Sign language translation systems implemented with computer vision leverage cameras, high-performance computing, and deep learning models to establish a permanent stream of human gesture telemetry. Frameworks such as MediaPipe and OpenCV are central to this process, standardizing access to hand landmarks, body pose data, and live gesture streams. These frameworks simplify the initial stages of the pipeline by providing efficient and fast pre-processing capabilities [1]. The use of custom neural network architectures, such as Convolutional Neural Networks (CNNs) and Hybrid Temporal Convolutional Networks (HTCNs), offers the technical infrastructure through which different gestures can be translated in real time. With the introduction of high-resolution cameras in modern devices, these computer vision diagnostic systems can transfer this data to cloud servers, mobile apps, or dedicated on-device translation systems for further processing. This allows for both local and remote analysis. On-device, or edge computing, is suitable for safety-critical applications due to reduced latency, while cloud solutions provide data pooling, scaling, and long-term forecasting. For computationally intensive tasks, some papers mention the use of CUDA for GPU acceleration to ensure that complex models can still operate at real-time speeds [5]. The papers also highlight that while OBD-II is a possible base for commercial applications, deep integration with systems like the CAN bus is more common in research and high-end vehicles. This detailed approach allows for the development of real-time sign language interpreters that can bridge the communication gap between the hearing and deaf communities.

B. Predictive Translation Concepts

Predictive translation is based on analyzing historical and operational real-time data to predict a sign before it is fully completed. In the case of sign language, this could mean predicting the next word, a full phrase, or an entire conversation before it happens. The techniques used include basic statistical analysis and complex AI like recurrent neural networks (RNN), convolutional neural networks (CNN), and ensemble models [6]. The benefit of predictive translation is the reduction of communication breakdowns and the minimization of misunderstanding.

C. Role of Embedded Platforms

Advanced analytics platforms such as high-performance computing platforms may provide a powerful tool, but predictive maintenance in actual vehicles may also require hardware that can run in constrained power environments and in hostile environmental conditions, such as embedded systems. A number of industrial as well as automotive applications use STM32 microcontrollers as they produce real-time results and are energy efficient. The frameworks like TensorFlow Lite for Microcontrollers enable the edge-based predictive analytics since trained AI models could be compacted and executed in such hardware.

D. Motivation for the Survey

Although the separate topics of IoT diagnostics, predictive maintenance, and embedded AI are the focus of a large number of studies, there exist no unified surveys, which comprehensively explore these domains [14]. Since that gap remains to be filled in the literature, this paper aims to present a synthesis of insights on the topic by the authors and identify similarities among various works as well as to suggest unified approaches to developing intelligent diagnostic systems of vehicles that are secure and scalable.

III. Methodology

According to the approach described in, the survey process consisted of 5 stages:

A. Defining the Research Focus

The primary objective of this survey was to meticulously collect and synthesize information on real-time sign language to speech translation systems. This research aims to provide a clear understanding of how technological advancements are enhancing communication for the deaf and hard-of-hearing community, moving beyond conventional methods. The focus includes systems that utilize modern approaches such as camera-based recognition, on-device processing, and sophisticated machine learning models to offer real-time translation capabilities. Furthermore, the survey explores systems offering bidirectional communication, including speech-to-gesture translation, to provide a more comprehensive communication bridge. The ultimate goal is to facilitate seamless, two-way interaction that empowers individuals and fosters greater social inclusion. This is achieved by systematically reviewing and classifying existing literature, analyzing the strengths and weaknesses of different approaches, and identifying key research gaps that must be addressed to transition from experimental prototypes to robust, widely adopted solutions.

B. Litrature Search and selection

The majority of the literature came from IEEE Xplore, with additional searches conducted in ACM and Springer. "Real-time sign language to speech," "camera-based sign language recognition," "MediaPipe sign language," "Convolutional Neural Network for gestures," and "on-device TTS" were among the keywords. In order to capture recent developments where sign language recognition and on-device translation have become mainstream, a 2024–2025 time frame was selected. After filtering through the initial pool of papers, a selection of papers was chosen that covered different architectures, analytics, and system frameworks.

C. Paper Inclusion and Exclusion Criteria

Papers that presented experimental or simulation results, were peer-reviewed publications from 2021–2025, and directly addressed real-time sign language translation to speech using camera recognition were included. Excluded were studies from non-peer-reviewed sources, without validation, or concentrating on non-camera-based methods, such as glove-based systems. Preliminary versions and duplicates were also eliminated. This guaranteed a targeted and reliable dataset..

D. Categorization of Reviewed Literature

The chosen papers were divided into four categories to allow for a structured review: Sign language recognition approaches focusing on specific models and frameworks; System architectures and hardware for real-time and on-device translation; The role of Text-to-Speech (TTS) and translation to speech; and Papers focusing on specific sign languages. Both similarities and unique contributions are highlighted by this classification.

E. Synthesis and Comparative

The synthesis and comparative analysis involved a detailed examination of each paper's machine learning models, chosen frameworks (e.g., MediaPipe, OpenCV, TensorFlow, Keras, Flask, Django), hardware platforms (e.g., cameras, sensor gloves, Raspberry Pi, GPUs), and reported performance metrics (accuracy, speed, latency) The comparison also highlighted system limitations, such as variations in lighting conditions, complex backgrounds, detection accuracy with body part overlap or obstruction, difficulties in distinguishing extremely similar signs, and the need for large datasets for training . This comprehensive synthesis provided a detailed overview, elucidating the strengths, weaknesses, and emerging research gaps in the field of real-time sign language translation systems.

IV. Litrature Review

A large literature has been developed during the past few years on vehicle diagnostics and predictive maintenance using IoT. As the Internet of Things technologies are gaining popularity in cars, the diagnostic procedure has taken a turn to an intelligent system that can undertake the process of continuous monitoring and real-time analysis, as opposed to the isolated failure detection, which is limited to the workshop. A closer inspection of the available literature would show a few thematic groups which form the basis of the existing knowledge., and hybrid processing methods at the edge and cloud.

A. System Architectures and Components

Sign language translation systems increasingly employ diverse architectures to facilitate real-time communication. Many proposed systems adopt a client-server model, where a mobile camera or webcam captures user gestures that are then sent to a web-based server for processing. Frontends are often developed using platforms like Flutter or as Android/iOS applications, communicating with backends via REST APIs in JSON format [17]. Core components include the MediaPipe framework for hand and body landmark detection, alongside computer vision libraries such as OpenCV for image capture and processing. Alternatively, some systems utilise wearable, glove-based approaches embedded with flex sensors and managed by Arduino Uno microcontrollers to interpret hand movements. Backend processing often leverages Python with frameworks like Django or directly on Raspberry Pi microcontrollers for embedded solutions

B. Gesture Recognition Methodologies

The core of sign language translation systems relies heavily on advanced machine and deep learning techniques for accurate gesture recognition. Convolutional Neural Networks (CNNs) are widely adopted for their efficacy in static sign recognition and extracting spatial features from hand images [18]. For dynamic signs and capturing temporal patterns, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are frequently employed. Hybrid models, combining CNNs with LSTMs or Temporal Convolutional Networks (TCNs), are also prominent for their ability to handle both spatial and temporal data effectively. Pre-processing techniques like Canny Edge Detection, image resizing, grayscale conversion, and Min-max Normalisation are crucial for enhancing image quality and model accuracy. Some systems also use MediaPipe for precise hand landmark extraction (e.g., 21 key points), which then serve as input for classification models.

C. Real-Time Performance and Optimisation

Achieving real-time performance is a critical focus in current sign language translation research to ensure seamless communication. Systems are designed to process video frames rapidly, often employing GPU acceleration techniques. For instance, CUDA parallelisation is utilised to process multiple frames concurrently, significantly speeding up inference. Further optimisation is achieved through dedicated libraries like TensorRT and cuDNN, which enhance deep learning model inference, leading to high frame rates and reduced latency. Some approaches leverage lightweight models and efficient algorithms to enable deployment on resource-constrained devices, such as Raspberry Pi microcontrollers or without dedicated GPUs. Challenges include maintaining performance under varying lighting conditions, gesture speeds, and complex backgrounds, which can impact detection accuracy. Microservice architectures are also proposed to enhance scalability and fault isolation for continuous real-time operation.

D. Dataset Creation and Handling

The availability and quality of datasets are paramount for training robust sign language recognition models. Researchers frequently address the challenge of limited public datasets by creating their own, which are tailored to specific sign languages (e.g., Indian Sign Language or American Sign Language) and include diverse variations. These custom datasets often feature a large number of images or video sequences (e.g., over 35,000 images for Indian Sign Language or 87,000 samples for ASL) spanning various classes (e.g., A-Z, 0-9, and specific words/phrases) [19]. To improve model robustness and prevent overfitting, data augmentation techniques like random rotations, zooms, and shifts are commonly applied to existing training data. Furthermore, systems may implement features allowing users to customise and train the model with their own signs, requiring a minimum number of samples per symbol to ensure reliable prediction.

E. Translation Outputs and Bidirectional Communication

Modern sign language translation systems extend beyond simple text conversion to offer rich, multi-modal outputs and even bidirectional communication. Text-to-Speech (TTS) engines, such as Google Text-to-Speech API or pyttsx3, are a standard component, converting translated text into natural-sounding speech for auditory feedback. To enhance linguistic accuracy and contextual integrity of the generated text, some advanced systems integrate Generative AI models, like Google Gemini Pro or Flan T5 Base [20] A significant advancement is the development of bidirectional systems that not only translate gestures to text/speech but also convert spoken words or text into corresponding sign language gestures. These voice-to-gesture features often display animated sign language representations using motion graphics, thereby fostering more comprehensive and inclusive communication between hearing and non-hearing individuals.

TABLE I. REAL-TIME SIGN-LANGUAGE DETECTION: THEMATIC GROUPS

Theme	Key Technologies /Components	Limitations
Evolution in Vision-Based Feature Extraction and Processing	CNNs, MediaPipe, image pre-processing, data augmentation.	Environmental issues, obstructions, similar signs, data scarcity

Transition to Bidirectional Communication	Speech recognition, gesture models, TTS, Flask/Python	Grammar/context, complex gestures, limited research, high compute, internet dependency.
Integration of Advanced Language Models for Natural Output	Generative AI/LLMs, word segmentation, machine translation.	Computational expense, internet need, prediction latency, fine-tuning.
Optimisation for Real-Time Performance and Scalability	CUDA, TensorRT, cuDNN, client-server, Django, Flutter, REST API, microservices	High resource demand, latency, environment impact, platform issues.
Wearable Sensor-Based Systems	Flex sensors, Arduino microcontrollers, Python TTS, joysticks.	Costly hardware, lower accuracy, environmental effects, calibration, comfort

V. Research Gaps & Challenges

Despite the rapid development of real-time sign language to speech translation systems, there are still multiple gaps and challenges. One prevalent limitation is the absence of standardised architectures and consistent evaluation methods. A majority of the solutions advanced in most studies are custom in nature, often designed for specific sign languages or controlled environments, and hence hard to scale to various types of devices or real-world scenarios. This lack of interoperability and a universally accepted framework hinders broader adoption..

Security and privacy are also very relevant issues. As these systems capture sensitive visual data, ensuring that diagnostic data remains secure from hacking or misuse is paramount. While some solutions incorporate authentication modules or basic data protection, a comprehensive framework offering a balance between robust security, user-centered privacy, and low computational overhead remains largely unattainable

Lastly, the full scope of integration and applicability is under-explored. Although individual case studies demonstrate impressive accuracy for specific signs or phrases, there are few works that evidence large-scale implementations capable of handling the continuous flow of natural sign language sentences, adapting to diverse signing styles, or robustly performing in varied environmental conditions such as low light or occluded hands. The bidirectional translation (speech-to-gesture) aspect is also a relatively underdeveloped area, limiting comprehensive two-way communication. These gaps represent crucial areas that must be addressed to move beyond experimental prototypes to widespread system deployment, ensuring these technologies are reliable, inclusive, and truly transformative for the deaf and hard-of-hearing community.

VI. Future directions

The future of IoT-driven real-time sign language to speech translation systems lies in enhancing inclusivity, scalability, and efficiency. A primary challenge in this domain is the lack of standardised, large-scale, and diverse benchmark datasets covering multiple sign languages across different regions and dialects. Addressing this requires collaborative dataset development initiatives with open-source accessibility to researchers worldwide. Leveraging computer vision tools such as MediaPipe for accurate hand, facial, and body landmark extraction can significantly improve the robustness of recognition frameworks, even under varying environmental conditions [2].

Since these systems are expected to run on resource-constrained edge devices like Raspberry Pi microcontrollers and wearable smart spectacles, research must focus on lightweight, energy-efficient AI models. Optimisation techniques such as model pruning, quantisation, and knowledge distillation are crucial for compressing deep learning architectures like CNNs, LSTMs, and RNNs without compromising accuracy [4]. These methods allow efficient low-power inference, enabling continuous operation despite limited connectivity, which is critical in real-world applications.

Moreover, data security and privacy remain central to widespread adoption. Future frameworks must incorporate privacy-preserving approaches such as federated learning, where models are trained collaboratively across devices without centralising sensitive user data. Additionally, blockchain technologies can provide immutable record-keeping and secure data transactions, ensuring transparency and trust within IoT ecosystems [9].

To ensure real-time performance, integration with next-generation communication networks (5G/6G) is essential for achieving low-latency data transfer [12]. The envisioned system captures camera input through wearable spectacles, processes it using LSTM and RNN models on a Raspberry Pi (and optionally offloaded to edge servers), and then provides instant speech feedback via the spectacles' built-in speakers. Achieving this at scale requires GPU acceleration through frameworks such as CUDA, TensorRT, and cuDNN, which are vital for sustaining high frame rates and ultra-low latency translation.

VII. Conclusion

Over the past several years, real-time sign language recognition and translation systems have extended far beyond basic word-for-word platforms to full-fledged AI-driven systems capable of on-device inference and integration with Text-to-Speech (TTS) algorithms. Advances in machine learning models, vision-based technologies like MediaPipe, and on-device hardware have broadened the scope of these systems to include continuous sign language interpretation and more natural speech output. However, ample areas of concern still exist, including the absence of standardized datasets, the performance of predictive models on resource-limited devices, the optimized translation of continuous sign language, and privacy-preserving designs. Future studies need to address these gaps to move beyond experimental prototypes to full-scale system deployment, focusing on developing benchmark databases, lightweight AI models, and robust frameworks for capturing the temporal and contextual nuances of sign language. The convergence of camera recognition, machine learning, and TTS solutions provides a promising roadmap to highly intelligent and widely accessible communication systems that will revolutionize modern mobility for both the deaf and hearing communities.

References

[1] "Real-Time sign language to speech converter using OpenCV and MediaPipe," *IEEE Conference Publication | IEEE Xplore*, Feb. 08, 2025. https://ieeexplore.ieee.org/document/10962728

[2] S. Patil, S. Gulave, V. Gawai, P. Gode, and P. Mudme, "Conversion of Indian sign language to speech by using deep neural network," *2022 6th International Conference on Computing, Communication, Control and Automation (ICCUBEA*, pp. 1–6, Aug. 2022, doi: 10.1109/iccubea54992.2022.10011043.

[3] B. N and G. S. Shenoy, "Empowering Communication: Harnessing CNN and Mediapipe for Sign Language Interpretation," *arXiv:2406.03729v1 [cs.LG] 6 Jun 2024*, pp. 1–8, Nov. 2023, doi: 10.1109/icraset59632.2023.10420395.

[4] Autonomous Vehicle Security Enhancement L. Santos and R. Costa, "Lightweight edge-based diagnostics for OBD-II systems," in *Proc. IEEE Int. Conf. on Edge Computing*, Lisbon, Portugal, 2022, pp. 123–128.

[5] G. Krishnan, A. C, A. R. Buffo, C. D. C. E, and D. F, "Sign Language to Voice Translator Using Tensorflow and TTS Algorithm," in *Proc. IEEE Int. Conf. Mobile Networks and Wireless Communications (ICMNWC)*, Tumkur, India, Dec. 2021, pp. 1-5.

[6] A. Narayan, S. Das, A. Pimple, S. Verma, and G. Vetal, "GestureNet: Real-Time Sign Language Recognition Using a Hybrid Neural Network," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 6, no. 3, pp. 1-10, 2024.

[7] A. Patel, N. Pand, P. Patil, A. Jadhav, S. Radhe, L. Gadhikar, and J. Darvekar, "Real-Time Sign Language to Speech Converter Using OpenCV and MediaPipe," *Naval Mumbai Tech Journal*, vol. 4, no. 2, pp. 1-8, 2024.

[8] S. Patil, S. Odedra, V. Gavai, P. Gode, P. Mulaye, and P. Prathamesh, "Conversion of Indian Sign Language to Speech by Using Deep Neural Network," in *Proc. 6th Int. Conf. Computing, Communication, Control and Automation (ICCUBEA)*, Pune, India, Aug. 2022, pp. 1-6

[9] G. Singh, A. R. Verma, B. Rama, Ramji, and K. Meghwal, "Enhancing Sign Language Detection through MediaPipe and Convolutional Neural Networks (CNN)," *FOLJ*, vol. 9, no. 6, pp. 1-7, Jun. 2024.

[10] A. A, S. S, M. A. R, S. G, and K. K, "Real-Time Sign Language Interpretation and Translation to Speech Using CUDA and Machine Learning," in Proc. Int. Conf. Data Science and Business Systems (ICDSBS), Chennai, India, 2025, pp. 1-6

[11] Pramod Waghmare, P., Deshpande, A. M., Dubewar, S., & Dhaybar, T., "Deep Learning Approach for Combined Indian Sign Language Recognition and Video Generation Model," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4, pp. 3296–3302, Jun. 2024.

[12] Rawat, P., Kumar, P., Tamta, V. K., & Kumar, A., "A Comprehensive Approach to Indian Sign Language Recognition: Leveraging LSTM and MediaPipe Holistic for Dynamic and Static Hand Gesture Recognition," AIRO, 2025.

[13] Sharma, C. M., et al., "Indian Sign Language Recognition using Fine-tuned Deep Transfer Learning Model," Computers & Electrical Engineering, 2022.

[14]] Singh, G., Verma, A. R., Rama, B., Ramji, & Meghwal, K., "Enhancing Sign Language Detection through MediaPipe and Convolutional Neural Networks (CNN)," FOLJ, vol. 9, no. 6, pp. 1–7, Jun. 2024.

[15] Ta o, T., Zhao, Y., Liu, T., & Zhu, J., "Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches, Datasets, and Challenges," IEEE Access, 2025.

[16] Thong, S. X., Tan, E. L., Gui, C. P., Rahman, T. A., & Abdul-Rahman, A., "Sign Language to Text Translation with Computer Vision: Bridging the Communication Gap," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 20, no. 1, pp. 1–10, 2020.

[17] Waghmare, P. P., et al., "iSign: A Benchmark for Indian Sign Language Processing," Preprint, Jul. 2024.

[18] Zholshiyeva, L., "Deep Learning-Based Continuous Sign Language Recognition System," 2025.

[19] Zhang, Y., "Recent Advances on Deep Learning for Sign Language Recognition," Expert Systems with Applications, 2024

[20] S. X. Thong, E. L. Tan, C. P. Gui, T. A. Rahman, and A. Abdul-Rahman, "Sign Language to Text Translation with Computer Vision: Bridging the Communication Gap," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 20, no. 1, pp. 1-10, 2023.