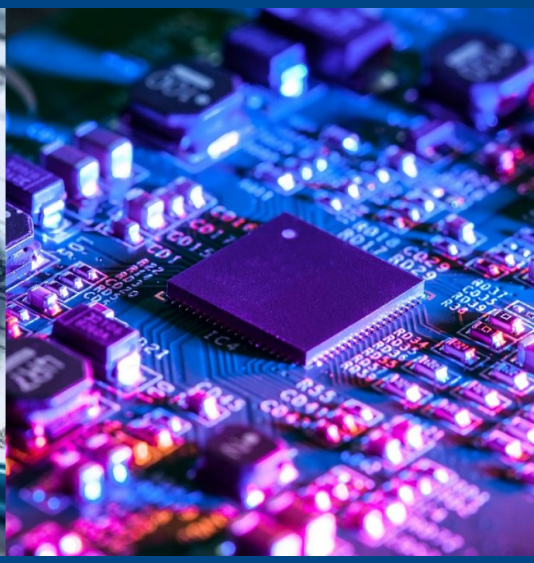
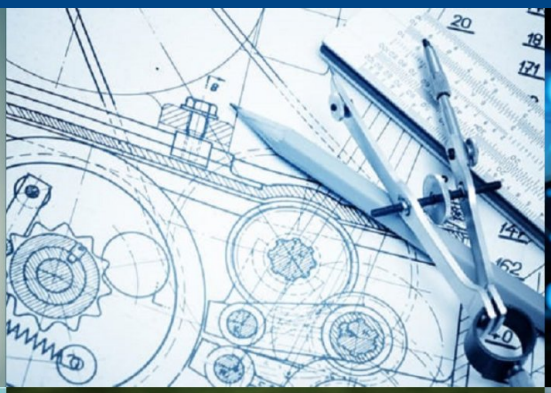
ISSN: 2582-7219

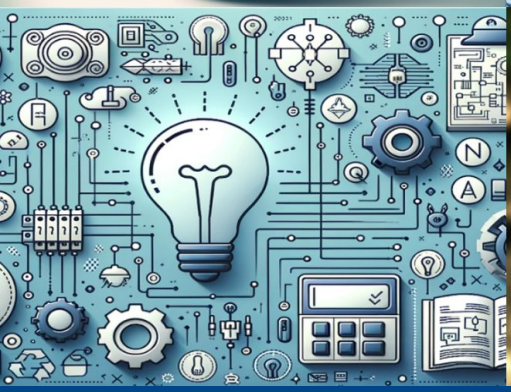




# International Journal of Multidisciplinary Research in Science, Engineering and Technology

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*



**Impact Factor: 8.206** **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904392

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# Real Time Multilingual Video Call Translation System

**Pradeep Kumar P, Dr. Jose Reena K**

MCA 2nd Year, Department of Computer Applications, B.S Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India

Assistant Professor, Department of Computer Applications, B.S Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India

**ABSTRACT:** This project presents the design and implementation of a Real-Time Multilingual Video Call Translation System that enables seamless communication between users speaking different languages during live video conferencing. Traditional video call systems lack integrated real-time translation, creating communication barriers and limiting effective interaction in global environments such as online meetings, remote education, telemedicine, and international collaboration. The proposed system integrates WebRTC technology for peer- to-peer audio and video communication with AI-powered speech processing and translation services to provide instant multilingual interaction. The system captures live audio from users and utilizes Speech-to-Text (STT) technology to convert spoken language into text. The recognized text is then processed using a Machine Translation module to convert it into the target language. The translated text is further converted into natural speech using a Text-to-Speech (TTS) module, enabling users to hear the translated output in real time. A Node.js and Socket.io-based signaling server manages session establishment, user connections, and real-time communication coordination. The modular architecture ensures low latency, high scalability, and efficient performance. The system improves communication accessibility, eliminates language barriers, enhances collaboration, and provides an intelligent solution for real- time multilingual communication in modern video conferencing applications.

## I. INTRODUCTION

The rapid expansion of global communication through online meetings, remote work, virtual education, and telemedicine has increased the need for effective interaction between people speaking different languages. However, traditional video conferencing systems do not provide built- in real-time translation, creating communication barriers that limit collaboration, understanding, and accessibility. Users often rely on external translators or subtitles, which interrupt the natural flow of conversation and introduce delays. With the advancement of artificial intelligence, speech processing, and real-time communication technologies, it is now possible to develop intelligent systems that enable seamless multilingual communication. This project introduces a Real- Time Multilingual Video Call Translation System that integrates WebRTC for peer-to-peer audio and video communication with AI-based Speech-to-Text, Machine Translation, and Text-to-Speech technologies to provide instant translation during live video calls. The system captures spoken audio, converts it into text, translates it into the target language, and delivers translated speech to the receiving user in real time. By combining real-time media streaming, intelligent translation, and efficient signaling using Node.js and Socket.io, the proposed system enhances communication accessibility, reduces language barriers, and improves collaboration in modern digital communication environments.

## II. LITERATURE REVIEW

Bali et al. (2023) presented a comprehensive review of real- time spoken language translation systems integrating Automatic Speech Recognition, Neural Machine Translation, and Text-to-Speech technologies using deep learning architectures. The system focuses on low-latency streaming translation using transformer-based neural networks to improve translation accuracy and speech naturalness. The proposed framework demonstrates the importance of combining ASR, NMT, and TTS modules for effective multilingual communication. However, the study primarily focuses on speech translation pipelines and does not address integration with real-time video conferencing systems or synchronization challenges. This work highlights the need for unified system architectures that combine translation with

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

live communication platforms.Ahmed et al. (2025) proposed a streaming Automatic Speech Recognition and alignment-based Neural Machine Translation system designed to reduce latency in real-time speech translation applications. The system uses incremental speech processing and alignment techniques to generate translated output before the speaker finishes the sentence, improving response time. The architecture improves translation efficiency and supports continuous real-time communication. However, the study reports reduced accuracy when deployed on low-power edge devices due to limited computational resources and memory constraints. This research emphasizes the importance of optimizing models for efficient real-time deployment.

Jia et al. (2022) introduced a direct speech-to-speech translation system using an end-to-end deep learning framework that eliminates intermediate text generation. Thesystem uses encoder-decoder neural networks with attention mechanisms to directly convert source language speech into target language speech. This approach reduces processing latency and improves natural communication flow in multilingual environments. However, the model requires large training datasets and high computational power for effective performance. This study highlights the potential of end-to-end architectures for real-time multilingual communication systems..

Prabhavalkar et al. (2017) developed a sequence-to-sequence speech recognition system using deep neural networks and attention-based encoder-decoder models for accurate speech transcription. The system improves speech recognition accuracy by learning contextual relationships in speech signals. The architecture enhances the performance of speech recognition systems used in translation pipelines. However, the model requires extensive training data and high processing power, which may limit deployment on low- resource devices. This work emphasizes the importance of efficient model optimization for real-time applications.

Google Research (2020) proposed a multilingual neural machine translation system using transformer-based architectures to support high-quality translation across multiple languages. The system uses attention mechanisms to improve contextual understanding and translation accuracy. The architecture enables scalable multilingual communication and supports real-time translation services. However, the system relies heavily on cloud infrastructure, which introduces latency and dependency on network connectivity. This study highlights the importance of optimizing translation models for real-time communication environments.

Zhang et al. (2021) introduced a simultaneous speech translation system using streaming neural networks and incremental decoding techniques. The system processes speech input in small segments and generates translated output in real time without waiting for complete sentences. The architecture improves translation speed and enhances user interaction in live communication systems. However, maintaining translation accuracy while reducing latency remains a major challenge. This research emphasizes the importance of balancing speed and accuracy in real-time multilingual communication systems..

**PROPOSED SYSTEM**

The proposed system is an AI-based Intelligent Surveillance System designed to detect criminal and abnormal activities in real time. It uses deep learning-based object detection and automated alerts to improve public safety and reduce manual monitoring.

The system consists of the following major modules:

- **User Interface Module**
- **Authentication and Room Management Module**
- **Signaling Module**
- **WebRTC Streaming Module**
- **Speech Recognition Module**
- **Machine Translation Module**
- **Text-To-Speech Module**
- **Chat and Messaging Module**
- **Language Selection and Synchronization Module**
- **Screen Sharing Module**

The system accepts real-time video and audio input from webcams and microphones. The audio is converted into text using Speech-to-Text and translated into the selected language. The translated text is delivered as subtitles and speech to other participants in real time, ensuring seamless multilingual communication.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904392

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

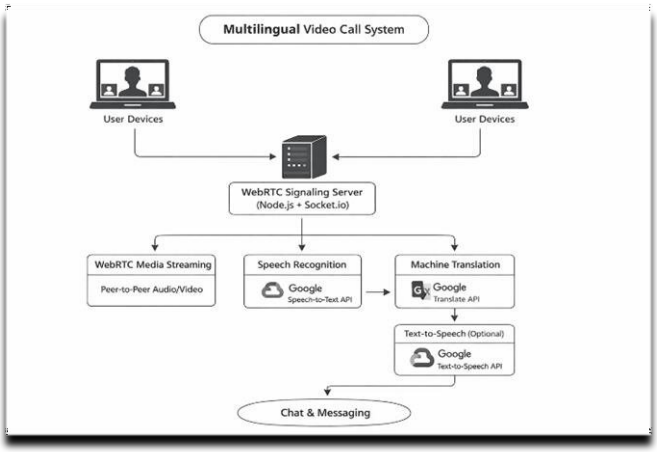
## III. SYSTEM ARCHITECTURE

The architecture diagram illustrates the complete workflow of the proposed Real-Time Multilingual Video Call Translation System. The system begins with the input layer, where real-time video and audio are captured from user devices such as webcams, microphones, laptops, or mobile devices. These inputs allow multiple users from different locations to join a common meeting room and communicate seamlessly. The captured media streams are transmitted through the WebRTC communication layer, which establishes peer-to-peer connections and ensures low- latency, secure audio and video transmission between participants..

Once the audio stream is received, it is forwarded to the Speech-to-Text processing module, where spoken language is converted into text using cloud-based speech recognition APIs. This extracted text is then sent to the Translation Module, which translates the recognized text into the target language selected by the receiving user. The translated text is further processed by the Text-to-Speech module to generate synthesized speech and is also displayed as real-time subtitles on the user interface. This ensures both visual and audio understanding of translated communication.

Finally, the translated audio and subtitles are delivered to other participants through the real-time delivery module. The system also includes session management and signaling components using Node.js, Express, and Socket.IO to manage room creation, user connections, and synchronization. This complete architecture enables real- time, accurate, and seamless multilingual communication during video conferencing.

Fig. 4.1




## IV. METHODOLOGY

Real-Time Audio And Video Capture

The system begins by capturing live audio and video streams from user devices such as webcams and microphones during a video call. WebRTC technology is used to ensure secure, low-latency, peer-to-peer media transmission. This module enables real-time communication between participants in different locations..

Speech-To-Text Conversion

The captured audio stream is sent to the Google Speech-to- Text API, where the spoken language is automatically recognized and converted into text. The API analyzes the audio in real time using advanced speech recognition models to generate accurate and reliable textual output. This extracted text serves as the input for the translation module, enabling seamless multilingual communication.

Machine Translation Processing

The extracted text is sent to the translation module, where it is translated into the target language selected by the user using the Google Cloud Translation API. This API provides fast and accurate Neural Machine Translation while

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904392

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

preserving context and meaning. This module enables seamless real-time multilingual communication between participants.

**A. Text-To-Speech Synthesis And Subtitle generation**

The translated text is sent to the Text-to-Speech API, where it is converted into natural-sounding audio in the selected language. The generated speech is played back to the user in real time, and the translated text is simultaneously displayed as subtitles on the user interface. This API-based approach ensures synchronized audio and visual output, enabling users to both hear and read the translated content for clear and effective communication.

**B. Real-Time Delivery And Synchronization**

The translated audio and subtitles are delivered to other participants through the WebRTC communication channel. Node.js and Socket.io are used for signaling, session management, and synchronization. This ensures seamless real-time multilingual communication during video conferencing. The system maintains low latency and synchronized delivery so that all participants receive the translated speech and subtitles simultaneously for smooth interaction.

**ALGORITHM**

**A. Speech-To-Text Conversion Algorithm**

The Speech-to-Text algorithm converts spoken language into text using the Google Speech-to-Text API. The captured audio stream from the user's microphone is transmitted to the API, where advanced deep learning models analyze speech patterns, phonemes, and acoustic features. The API processes the audio in real time and generates accurate textual output. This algorithm enables the system to extract meaningful text from live conversations, which serves as the input for the translation module.

**B. Neural Machine Translation Algorithm**

A Convolutional Neural Network (CNN) is a deep learning model designed to analyze visual data and extract important features from images. It uses convolution and pooling operations to identify patterns such as shapes, edges, and textures. In the proposed system, CNN processes frames extracted from surveillance videos and captures visual characteristics related to human activities. These features help the system understand interactions occurring in the scene. CNN is selected because it automatically learns relevant image features and improves activity recognition performance.

**C. Text-To-Speech Synthesis Algorithm**

The Text-to-Speech algorithm converts the translated text into natural-sounding speech using the Google Text-to-Speech API. The API processes the translated text and generates human-like audio output using advanced speech synthesis models. The generated audio is played to the user, and the translated text is also displayed as subtitles on the interface.

**D. Real-Time Communication And Streaming Algorithm**

The WebRTC algorithm enables real-time audio and video communication between participants. It establishes a peer-to-peer connection using signaling through Node.js and Socket.IO. WebRTC captures media streams from the camera and microphone, transmits them securely, and ensures low- latency communication. This algorithm enables smooth and uninterrupted real-time video call functionality.

**E. Signaling And Date Exchange Algorithm**

The Socket.IO algorithm manages signaling and real-time data exchange between connected users. It handles room creation, joining, session negotiation, and transmission of translation data. This ensures synchronization between participants and supports seamless communication. The algorithm enables reliable connection establishment and real-time translation delivery.

## V. RESULTS

The proposed system implements a real-time multilingual video communication platform that enables users to connect through a web interface, create or join rooms, and communicate seamlessly using live audio and video powered by WebRTC, while connection setup and room management are handled efficiently using Socket.IO. Once connected, users can interact with low latency and also access additional features such as chat messaging and screen sharing for enhanced collaboration. A major contribution of the system is its integrated speech processing and translation

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904392

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

capability, where spoken input is captured and converted into text using Speech-to-Text (STT), translated into a target language based on user preference, and then converted back into audio using Text-to-Speech (TTS). This allows users to receive both translated subtitles and synthesized speech in real time, enabling effective communication between participants who speak different languages. The system also ensures reliability through features like automatic reconnection of media devices and stable peer-to-peer streaming, while its hybrid architecture—combining client-server signaling with direct media transmission—reduces server load and improves performance. Overall, the project successfully demonstrates a scalable and efficient solution for breaking language barriers in video conferencing, making it highly suitable for applications in education, business communication, and global collaboration. Furthermore, the system is designed to be easily extendable with advanced AI models and additional language support in the future. It also provides a user-friendly interface, ensuring accessibility even for non-technical users.

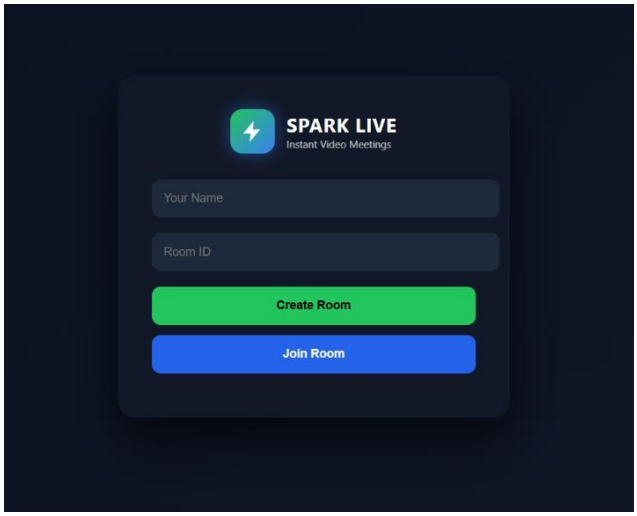



Fig 7.1

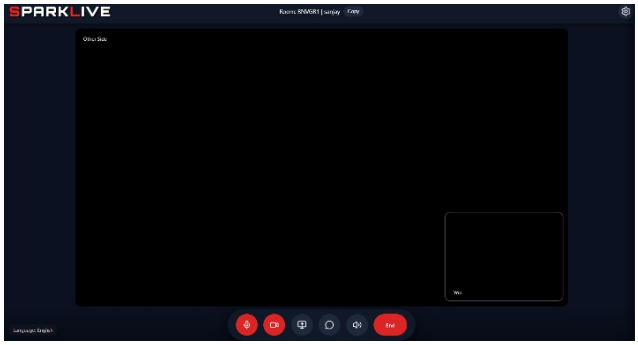



Fig 7.2

## VI. FUTURE ENHANCEMENT

Several future enhancements can be incorporated to further improve the performance, scalability, accuracy, and overall usability of the proposed Real-Time Multilingual Video Call Translation System. These improvements aim to enhance translation quality, reduce processing latency, support more users and languages, and ensure reliable real-time multilingual communication across diverse platforms and network conditions.

- Integration of additional language support to enable communication across a wider range of global languages
- Development of a mobile application to provide real-time translation support on smartphones and tablets
- Optimization of latency to achieve faster speech recognition, translation, and audio delivery
- Enhancement of speech recognition accuracy in noisy and multi-speaker environments
- Integration of AI-based noise reduction and voice enhancement techniques

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

• Deployment on scalable cloud infrastructure to support a large number of concurrent users and ensure reliable real-time performance

These enhancements will improve translation accuracy, reduce latency, and enhance system scalability and efficiency, making it more reliable for real-time multilingual video communication.

## VII. CONCLUSION

The proposed Real-Time Multilingual Video Call Translation System was developed to enable seamless communication between users speaking different languages during video conferencing. The system integrates Speech-to-Text, Google Translation API, and Text-to-Speech technologies with WebRTC-based video communication to provide real-time translated audio and subtitles. By combining speech processing, translation, and real-time communication modules, the system ensures efficient and synchronized multilingual interaction between participants.

The implementation demonstrates that automated speech translation can significantly reduce language barriers and improve accessibility in global communication. The system provides accurate translation, real-time performance, and an intuitive interface for users. Although the system depends on internet connectivity and cloud-based APIs, the results show its effectiveness in practical video conferencing scenarios. With future improvements in translation accuracy, latency optimization, and multi-platform deployment, the system can become a reliable and scalable solution for real-time multilingual communication.

## REFERENCES

[1] J. F. Kurose and K. W. Ross, Computer Networking: A Top-Down Approach, 8th ed., Pearson Education, 2022.
[2] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson Education (Draft), 2023.
[3] A. Graves et al., “Speech Recognition with Deep Recurrent Neural Networks,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013 (widely cited foundational work).
[4] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems (NeurIPS), 2020 (core LLM reference still valid).
[5] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems (NeurIPS), 2017 (transformer foundation).
[6] Microsoft Azure, “Cognitive Services: Speech and Translation APIs Documentation,” 2024.
[7] Google Cloud, “WebRTC and Real-Time Communication Architecture Documentation,” 2024.
[8] M. B. Hoy, “The State of Speech Recognition Technology,” Medical Reference Services Quarterly, vol. 37, no. 3, pp. 321–329, 2022.
[9] Y. Zhang et al., “Recent Advances in End-to-End Speech Recognition,” IEEE Access, vol. 11, pp. 1–20, 2023.
[10] S. Latif et al., “A Survey on Deep Learning Techniques for Speech Recognition,” IEEE Access, vol. 11, pp. 1–25, 2023.
[11] A. Baevski et al., “wav2vec 2.0: A Framework for Self- Supervised Learning of Speech Representations,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1–12, 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

## IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com