International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

AI-Powered Accessibility: Enabling Effective Communication for Hearing and Speech Impaired in Virtual Platforms

Dr.AB.Hajira Be[1], M. Thilothana [2]

[1] *Associate Professor*
Department of Computer Applications
Karpaga Vinayaga College of Engineering and Technology
Maduranthagam TK
[2]*PG Student*
Department of Computer Applications
Karpaga Vinayaga College of Engineering and Technology
**Corresponding Author: Thilothana M Email: thilothana0716@gmail.com*

Abstract - This project presents an AI- driven system to enhance inclusivity in virtual meetings for individuals with hearing and speech impairments. It features three core modules: a Sign Recognition Module (SRM) using Temporal Convolutional Networks to interpret Indian Sign Language, a Speech Recognition and Synthesis Module (SRSM) converting speech to text via Hidden Markov Models, and an Avatar Module (AM) that translates speech into sign language visually.

Key Words: AI for Accessibility, Indian Sign Language Recognition, Temporal Convolutional Networks, Speech-to-Text using HMM.

1. INTRODUCTION

Effective communication is a cornerstone of human interaction, especially in the context of online education, remote collaboration, and digital accessibility. While virtual platforms such as Zoom, Microsoft Teams, and Google Meet have revolutionized communication, they often overlook the needs of individuals with hearing and speech impairments. Traditional solutions like human sign language interpreters or closed captions fall short in terms of availability, real-time response, and comprehensive understanding of sign language expressions.

To address this gap, this paper proposes an artificial intelligence (AI)-powered system that enables inclusive virtual communication through the automated recognition and translation of sign language. The proposed solution integrates real-time sign language recognition using Temporal Convolutional Networks (TCNs), speech-to-text processing with Hidden Markov Models (HMMs), and avatar-based visual sign synthesis. The objective is to bridge the communication gap between sign language users and non-signers in a digital environment, empowering them to participate equally in virtual meetings, classrooms, and professional settings.

2. RELATED WORKS

Several research efforts have been made in the field of sign language recognition and accessibility enhancement. Traditional machine learning models, including Support Vector Machines (SVMs) and ensemble techniques, have shown moderate success in static gesture recognition. However, their limitations in recognizing dynamic, continuous gestures and adapting to real-time applications make them less effective in practical settings.

Deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have improved accuracy in gesture detection and classification. Nonetheless, their sequential dependency and computational complexity pose performance constraints. Temporal Convolutional Networks (TCNs), in contrast, offer a parallel, flexible architecture suitable for time-series tasks like sign recognition. Other works have explored animated avatars for sign language translation, but most fail to provide natural motion or contextual accuracy in real time.

The proposed system advances the state of the art by integrating TCN-based sign recognition, speech synthesis, and avatar-based visual feedback into one unified, real-time, web-based solution suitable for integration with mainstream communication tools.

3. SYSTEM ARCHITECTURE AND METHODOLOGY

The proposed framework, Sign Meet, is designed to enable inclusive virtual communication by recognizing sign language, interpreting speech, and translating between the two using an AI-based avatar. The system is implemented as a modular web application integrated with virtual conferencing platforms. The three key functional modules are:

1. Sign Recognition Module (SRM)

The Sign Recognition Module is responsible for converting live video-based sign gestures into textual or spoken output. This module leverages Temporal Convolutional Networks (TCNs) due to their ability to handle sequential data and extract temporal features from continuous sign gestures.

- Preprocessing: Video input is converted to grayscale, resized for uniformity, and filtered using Gabor filters. Region Proposal Networks (RPNs) isolate hand gestures from the background.
- Feature Extraction: Temporal patterns are identified using frame-wise analysis, enabling accurate classification of dynamic hand gestures.
- Classification: The processed gesture data is passed through a CNN followed by a TCN to recognize signs with high temporal context sensitivity.
- Output: Recognized gestures are translated into text or speech for non-signers.

2. Speech Recognition and Synthesis Module (SRM)

The Speech Recognition and Synthesis Module enables speech-to-text conversion and vice versa, using Hidden Markov Models (HMMs).

- Audio Preprocessing: Speech input is filtered using Wavelet Transform to reduce noise.
- Feature Extraction: Mel-Frequency Cepstral Coefficients (MFCCs) are used to extract voice features.
- HMM Processing: The extracted features are mapped to phonemes and converted into textual form.
- Synthesis: Text-based input can also be converted to natural-sounding speech for enhanced accessibility.

3. Avatar Module (AM)

The Avatar Module is a core component of visual communication for deaf users. It translates speech or text input into animated sign language gestures using a 3D avatar

- Sign Mapping: Recognized text or speech is mapped to corresponding sign language gestures.
- Animation Engine: A motion-synthesis system drives the avatar using predefined gesture datasets, ensuring natural and expressive sign rendering.
- Display: The avatar appears alongside the video interface, visually interpreting speech in real-time.

This module bridges the comprehension gap for users who rely primarily on visual sign language.

4. Platform Integration

The entire system is deployed via a Flask-based web application, connected to MySQL for user and meeting data management. It integrates with Jitsi Meet, enabling real-time virtual conferencing with embedded accessibility features. Technologies used include TensorFlow, OpenCV, Mediapipe, and Bootstrap.

4. IMPLEMENTATION AND RESULTS

1. Dataset and Model Training

To ensure robust gesture recognition, the SignNet Model was trained on a curated dataset of Indian Sign Language (ISL) gestures. Each video sample was preprocessed into frames, binarized, and segmented using Region Proposal Networks (RPNs). The dataset was expanded using image augmentation techniques to improve model generalization.

- Training Technique: The system utilized a CNN + TCN architecture. CNN layers extracted spatial features, while TCN layers modeled temporal sequences.
- Training Duration: Approximately 6 hours on a mid-range GPU (NVIDIA GTX 1660 Ti).
- Tools Used: TensorFlow, OpenCV, and Mediapipe for video capture and preprocessing.
- The model achieved an average classification accuracy of 92.4% on a test set of 500 samples, demonstrating strong recognition capabilities even with slight gesture variations

2. Real-Time Speech Recognition and Synthesis

- The Speech Recognition and Synthesis Module (SRSM) employed Hidden Markov Models (HMMs) for audio feature mapping and used MFCCs for extraction.
- Speech-to-Text Accuracy: 89.7% in normal conditions; 81.5% in moderately noisy environments.
- Text-to-Speech Quality: Generated via Google Text-to-Speech (gTTS) and pyttsx3, with low latency.
- Combined with noise filtering (Wavelet Transform), the module offered near-instantaneous responses during live sessions.

3. Avatar-Based Sign Rendering

The avatar module was implemented using 3D motion synthesis and predefined gesture models mapped to speech inputs.

- Rendering Speed: Average delay ~0.8 seconds per sentence.

- User Feedback: Participants described the avatar as "intuitive" and "easy to follow."
- Customization: Users could choose sign language dialect and speech language preferences (e.g., Tamil, Hindi, English).

4. End-to-End System Integration

The web platform allowed:

- Deaf users to host or join virtual meetings, with sign-to-text and avatar-based translation.
- Non-deaf users to speak or type, and receive responses in text or sign.
- Admins to manage training datasets, monitor usage, and approve new users.
- The entire system operated in real-time with minimal latency, making it highly effective for educational and professional settings.

5. DISCUSSION

The implementation of the Sign Meet system represents a significant advancement in the field of accessible virtual communication. By combining AI-powered sign language recognition with speech processing and avatar-based translation, the platform addresses several long-standing barriers faced by individuals with hearing and speech impairments.

1. Strengths and Contributions

- Real-Time Bidirectional Communication: Unlike traditional systems that focus solely on text captions or pre-recorded signs, Sign Meet allows live, two-way interaction between signers and non-signers.
- High Recognition Accuracy: The use of Temporal Convolutional Networks (TCNs) for gesture recognition delivers improved accuracy over conventional models like RNNs and SVMs.
- Modular Architecture: Each component—SRM, SRSM, and AM—can be individually updated or scaled without affecting the rest of the system.
- Platform Independence: Integration with Jitsi ensures that the system is lightweight, free to use, and easily adaptable to existing virtual platforms like Zoom or Microsoft Teams.
- Multilingual Support: The use of translation APIs allows the system to work across multiple Indian languages, making it contextually rich and regionally inclusive.

2. Challenges and Limitations

- Environmental Sensitivity: Recognition accuracy may degrade in low lighting or with background clutter. Gesture misclassification was more common in dim environments or with occluded hand movements.
- Avatar Naturalness: While effective, the avatar system currently lacks emotive expressions and complex grammar handling, which can limit nuanced communication.
- Resource Constraints: High-performance training of the TCN-based SignNet model requires dedicated GPU resources, which may not be available in all institutions.
- Limited Gesture Set: The current system is optimized for Indian Sign Language (ISL) and may not generalize perfectly to other national sign languages without retraining.

3. Ethical and Social Considerations

The system promotes digital equity by creating opportunities for the hearing and speech-impaired community to participate in mainstream virtual education and work environments. However, care must be taken to:

- Respect cultural variations in sign language.
- Avoid over-reliance on automated systems in contexts where human interpretation may still be necessary (e.g., legal settings).
- Ensure data privacy during video and audio processing.

6. CONCLUSION AND FUTURE WORK

This paper presented Sign Meet, an AI-powered virtual communication platform designed to empower individuals with hearing and speech impairments through accessible real-time interaction. By leveraging Temporal Convolutional Networks (TCNs) for sign recognition, Hidden Markov Models (HMMs) for speech processing, and a visual avatar-based translator, the system successfully bridges the gap between signers and non-signers in digital environments.

The integration of the proposed system with popular virtual platforms like Jitsi Meet demonstrates the feasibility of enhancing mainstream communication tools with AI-driven accessibility features. The system has shown promising results in terms of accuracy, usability, and social impact.

FUTURE WORK

- Expanding the gesture database to cover multiple sign languages and complex sentence structures.

- Improving the avatar's emotional expressiveness and grammatical accuracy.
- Enabling offline usage with lightweight models for low-bandwidth or remote areas.
- Incorporating AI-driven grammar correction for smoother translations between sign and spoken languages.

ACKNOWLEDGEMENT

The author would like to thank the faculty and project mentors at Karpaga Vinayaga College of Engineering and Technology for their guidance and support throughout the development of this research. Special thanks are extended to the Department of Computer Applications for providing the technical infrastructure and encouragement that made this work possible.

REFERENCES

[1] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2016). Temporal Convolutional Networks for Action Segmentation and Detection. IEEE CVPR.

[2] Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257–286.

[3] Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in Neural Information Processing Systems.

[4] OpenCV Library: https://opencv.org/

[5] TensorFlow Developers. (2023). TensorFlow: An end-to-end open source machine learning platform. https://www.tensorflow.org/

[6] Google Text-to-Speech API: https://cloud.google.com/text-to-speech

[7] Mediapipe by Google. https://mediapipe.dev/