Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Next-Gen Communication: AI-driven Speech-to-Sign Language Translation

Vinaya Kulkarni1, Pranoti Kale2, Sanika Chaudhari3, Shruti Bhumkar4, Manasi Deshmukh5, Samruddhi Deshmukh6

[1]*Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra.*
[2]*Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra.*
[3]*Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra.*
[4]*Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra.*
[5]*Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra.*
[6]*Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra.*
Emails: sanikac450@gmail.com, bhumkarshrut@gmail.com, manasitd2004@gmail.com, deshmukhsamu17@gmail.com

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Accessibility to digital content is challenging for individuals with hearing impairments as it lacks sign language translation. This limitation impacts their ability to consume digital resources available in audio or video format, which creates an accessibility gap. To bridge this gap, our project introduces an AI-powered system that converts Video to sign language animation using an avatar. The system integrates NLP, CNN, Speech recognition API and various libraries in its internal conversion steps from processing inputs till generating gestures by the avatar to ensure accurate and efficient translation. This paper discusses the system's design, development and evaluation for its effectiveness in making digital content more accessible.

Keywords: Sign Language Translation, Video Processing, Audio Extraction, Audio-to-Text Conversion, Text Preprocessing, Contextual Analysis with BERT, Gesture Synchronisation, LSTM Networks, Speech Recognition, Natural Language Processing (NLP).

1. INTRODUCTION

With rapid digitization, multimedia content in audio and video formats has become the primary mode of information. Multiple platforms like YouTube utilize audio-video formats that have transformed education, entertainment, and social interaction. This increasing dependency on video-based digital content presents a major challenge for the Deaf and hard-of-hearing community, which restricts their access to information. According to WHO estimates, over 430 million people worldwide face hearing impairments, which represent 5% global population, and this figure is expected to rise in future. Due to a lack of inclusive formats, many hearing-impaired individuals are unable to engage with essential digital resources. The challenge is even greater in educational and professional domains where audio and video-based communication is prevalent in lectures, meetings, and e-learning.

The communication gap exists between hearing and non-hearing individuals due to the absence of an automated sign language translation system. Captions do provide some level of accessibility, but do not replace sign language. Existing sign language translation solutions depend on human interpreters, which are less scalable and impractical. Direct word-to-word translation is an ineffective solution as sign language has its grammatical structure and syntax different from spoken language. The traditional captioning system faces challenges in conveying meaning accurately, leading to potential misunderstandings for DHH viewers.

This research proposes an AI-powered system that translates the spoken content of video into sign language animations. Initially, the video input is processed to extract the audio using the FFmpeg library. This audio undergoes pre-processing steps such as noise reduction and normalisation to improve clarity. The cleaned audio is then transcribed into text using the IBM Watson Speech-to-Text API. Next, natural language processing techniques are applied—using libraries like spaCy and NLTK—for tokenization, lemmatization, and removal of stop-words. The

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

system further uses a BERT-based model to restructure the text grammatically to align with Indian Sign Language (ISL) syntax. This processed text is then mapped to corresponding ISL gestures using a Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN), ensuring smooth temporal sequencing of gestures. Finally, a Three.js-powered 3D avatar renders these gestures visually, with dynamic time warping (DTW) and inverse kinematics (IK) enhancing gesture fluidity and realism.

Unlike traditional text-based conversion methods, this project emphasises video-to-sign translation that ensures spoken language is effectively represented through sign language animations. This innovation enhances digital inclusivity by providing sign language translation for a broader range of digital resources. This project sets the foundation for future advancements in automated sign language interpretation, promoting greater inclusivity in digital communication.

2. LITERATURE REVIEW

Debashis Das Chakladar [1] proposed a 3D avatar-based system that translates English speech/text into Indian Sign Language (ISL) using NLP techniques and Blender animation. The system employs IBM Watson for speech-to-text conversion, context-free grammar parsing for translation. Finally, the motion list was used to synthesize avatar gestures.

Modi and Jain [2] developed a system that translates YouTube video transcripts into American Sign Language (ASL) using deep learning and video processing. Their approach trains a Convolutional Neural Network (CNN) which recognises ASL alphabets. Further, the system automatically extracts video transcripts and places corresponding ASL sign images in the video frames to generate subtitles.

Another paper introduces a speech-to-ISL system that utilises Google Speech API for transcription, NLTK for preprocessing, and Blender for 3D avatar-based sign animation. The system incorporates a Django-based web interface and handles isolated words and complete sentences as well, by using stored animation clips.

Ritika Bharti et al. [4] have proposed an automated speech-to-sign language translation system using the Google Speech-to-Text API and Natural Language Processing. Their pipeline consists of speech acquisition, text tokenization, and matching the text with a visual sign language video library based on ASL.

Furthermore, a paper presents a rule-based English-to-Indian Sign Language translation system using grammar restructuring, detection of multi-word expressions, and synonym substitution. The system preprocesses text into ISL glosses using Stanza and WordNet, which generates videos with over 95% accuracy.

A method that creates 3D sign language subtitles for videos is shown in another work [6] with the goal of improving accessibility for the deaf and hard-of-hearing community. This method uses a 3D avatar to translate spoken text into animated motions. They have used Speech Recognition tools for audio transcription and animation software to map them with a 3D avatar. The authors address issues with avatar realism and gesture synchronization, offering insights on how to combine motion mapping and speech recognition for efficient sign language representation.

A prototype pipeline for creating a 3D sign language avatar from 2D video inputs is shown in this [7] work. It is planned to be displayed on augmented reality (AR) glasses. The system produces the 3D avatar in an augmented reality environment, records human posture motions while signing, and uses joint coordinates to control the skeleton of the avatar. The work addresses issues with avatar realism and hand gesture detection, providing a basis for next improvements that include machine translation and speech recognition.

A neural machine translation model improved with Generative Adversarial Networks (GANs) is presented in this[8] work to generate realistic sign language videos from the text inputs. This method improves gesture fluidity and avatar realism by learning from parallel text-sign datasets to construct sequences of sign gesture movements. The authors assess the quality of generated outputs by using metrics like SSIM, PSNR, and MSE, to demonstrate the model's capability to synthesise sign language videos that closely resemble ground truth data.

This paper presents a system[9] that translates spoken English into Indian Sign Language (ISL) using speech recognition and animation tools. This approach is of converting audio input into text, then applying ISL grammar rules and generating corresponding gestures displayed via an animated avatar as output. The system integrates speech recognition APIs for transcribing audio, employs Natural Language Processing (NLP) techniques for text processing, and utilises animation software to render ISL gestures through a 3D avatar.

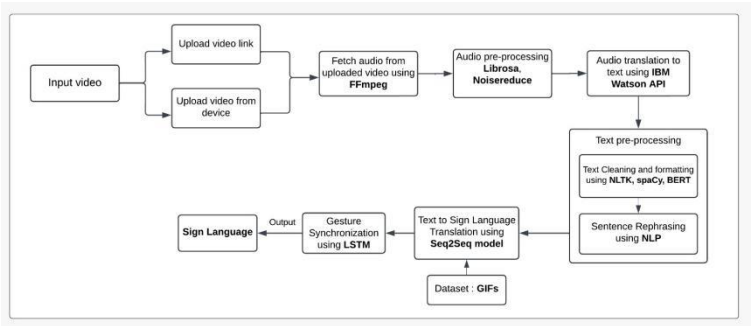
Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

U. S. Prasad presents a method that introduces a system designed to translate both audio and text inputs into Indian Sign Language (ISL) gestures[10]. This system uses speech recognition APIs to convert spoken language into text, then processes it using Natural Language Processing (NLP) techniques to align with ISL grammar. Subsequently, the processed text is mapped to the corresponding sign language gestures using a predefined ISL dataset. This integrated approach provides real-time translation that aims to enhance communication for individuals with hearing impairments.

3. IMPLEMENTATION AND RESULTS:

The developed system is an Avatar-driven Indian Sign Language system intended to bridge the communication gap between hearing-impaired users. The system utilises new web technologies, machine learning and processing to offer a complete set of ISL translation and learning tools. The following are the main features and functionalities of the system:



3.1. USER INPUT AND PROCESSING

The system begins with the user providing a YouTube video link or uploading a video directly from their device as the input. The system even accepts Speech or text as input. This input serves as the main source of speech data. Once the user provides the video, audio or text. This library makes sure that the video is downloaded, maintaining the original audio quality. On the other hand, if the user uploads the local video file, it is saved in an assigned directory for further processing.

Furthermore, the next step involves extracting the audio from the video. This is employed using the FFmpeg library, which separates the audio stream from the video file without affecting its quality. The extracted audio is temporarily stored in MP3 format using the pydub library to maintain compatibility for the next processing steps. To enhance the audio quality, the system performs various audio pre-processing steps such as Noise Reduction, Normalisation and Segmentation. These steps are essential to eliminate the background noise and adjust the volume levels for clearer audio.

After pre-processing the audio, it is now ready for the next phase, which is Speech-to-Text Conversion.

3.2. SPEECH-TO-TEXT CONVERSION:

Before sending the audio to Google's Speech Recognition API, further audio pre-processing is performed to ensure optimal transcription quality. The extracted audio is checked for sampling rate consistency, and non-speech segments (such as background noise or silent pauses) are filtered out. This step enhances recognition performance by removing unnecessary audio artefacts that might interfere with the transcription process.

The processed audio file is then passed as input to the IBM Watson Speech-to-Text API through an HTTP POST request. The API processes the audio and returns the recognised text in JSON format. The system extracts the relevant transcribed text and stores it for further pre-processing. In cases where the speech recognition model detects multiple possible transcriptions, confidence scores are used to select the most probable text output.

One of the key advantages of using Speech Recognition API is its ability to recognise multiple languages and dialects, allowing flexibility in extending the system's capabilities. Additionally, the API supports speaker diarization, which can be useful in scenarios where the input video contains multiple speakers. This ensures that speech from different speakers can be accurately transcribed and mapped to the appropriate gestures.

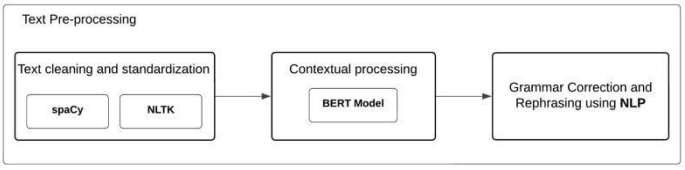
Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Once the text transcription is obtained, it is stored in a structured format and passed to the text pre-processing module, where it undergoes further refinement to enhance translation accuracy.

3.3. TEXT PREPROCESSING

Once the speech-to-text conversion is completed, the system processes the transcribed text to ensure that it is clean, structured, and optimised for accurate mapping to Indian Sign Language (ISL) gestures. This step is crucial because raw speech-to-text output often contains grammatical inconsistencies, filler words, punctuation errors, and redundant phrases, which can affect the accuracy of gesture translation.



The first stage of text pre-processing involves tokenization, where the transcribed text is broken down into individual words or phrases. This is achieved using spaCy, a widely used Natural Language Processing (NLP) library. Tokenization helps in identifying key linguistic components that need to be mapped to sign language gestures. Following tokenization, stop-word removal is performed using NLTK (Natural Language Toolkit). Stop-words, such as "is," "the," "a," and "of"—are removed because they do not have direct equivalents in ISL and do not contribute to the meaning of gestures.

Next, lemmatization is applied to convert words into their base forms. For example, words like "running" or "ran" are transformed into "run." This normalization step ensures that variations of the same word do not lead to incorrect gesture mapping. Additionally, named entity recognition (NER) is implemented using spaCy's pre-trained models to identify proper nouns, dates, or special terms that may require different gesture mappings or need to be preserved in their original form.

After these fundamental steps, the system applies contextual grammar correction using the BERT (Bidirectional Encoder Representations from Transformers) model. This deep learning-based approach helps refine the structure of the transcribed text, ensuring that fragmented or misinterpreted phrases are corrected before sign language translation. BERT analyzes the context of each word in a sentence, making intelligent corrections while maintaining the intended meaning.

Once the text is cleaned and structured, the final step is mapping the processed text to ISL-compatible phrases. Since Indian Sign Language does not follow the same grammatical structure as spoken languages, the system rearranges words to match ISL syntax. For instance, the English sentence "What is your name?" would be rearranged to "Your name what?" to align with ISL's grammatical structure. This transformation ensures that the generated sign language gestures are both accurate and natural.

At the end of this stage, the pre-processed text is ready for gesture mapping, where each word or phrase will be linked to corresponding sign language gestures stored in the ISL dataset. The refined text improves translation accuracy, ensuring that the final sign language output is both meaningful and contextually correct.

3.4. GESTURE SYNCHRONIZATION & DEEP LEARNING MODEL

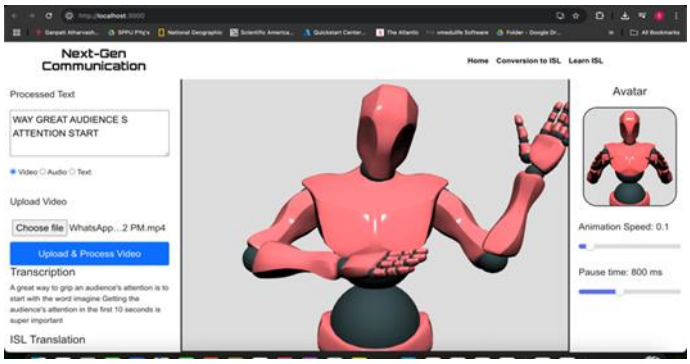
After the text has been pre-processed, the system moves to the crucial phase of gesture synchronisation and deep learning-based sign generation. This step ensures that the final output—an animated avatar performing sign language—is fluid, natural, and synchronised with the original speech. Unlike simple word-to-gesture mapping, where signs are displayed in isolation, this stage integrates deep learning models to generate smooth and coherent gesture transitions that reflect real-world sign language communication.

The system utilises a Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN) to maintain temporal consistency in gesture sequences. Since sign language is inherently sequential, the LSTM model is trained on a dataset containing continuous sign sequences, ensuring that gestures are aligned properly with the spoken

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

sentence's rhythm and flow. Each sign gesture is stored as a sequence of frames in the dataset, and the LSTM model learns how to transition between them naturally, preventing abrupt or robotic movements.



At the end of this stage, the system successfully generates a synchronised, AI-driven sign language avatar that translates spoken content into fluid and grammatically accurate ISL gestures. This marks a significant step toward bridging the communication gap for the Deaf and Hard-of-Hearing (DHH) community, making spoken content more accessible and inclusive.

3.5. AVATAR-BASED SIGN LANGUAGE OUTPUT

Gesture synchronisation ensures that sign language gestures are generated in a smooth, natural, and contextually accurate manner. The system employs a Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN) to maintain temporal consistency, ensuring that gestures transition seamlessly without abrupt movements. To align signing speed with speech, dynamic time warping (DTW) is applied, adjusting the duration of each gesture based on the pace of spoken content. The system utilises Three.js for realistic avatar rendering, with Bezier curve-based interpolation to smooth transitions between gestures. Additionally, inverse kinematics (IK) refines joint movements, preventing unnatural distortions in hand and arm positions. The model is fine-tuned on the ISL-CLTR dataset, enhancing accuracy by learning context-specific variations of Indian Sign Language (ISL). By integrating these techniques, the system ensures a fluid, expressive, and grammatically accurate sign language output, bridging the communication gap for Deaf and Hard-of-Hearing (DHH) individuals. Epoch graphs represent the training process of our LSTM-based gesture generation model. Each epoch indicates one full pass through the training dataset, where the model continuously learns to map text to corresponding ISL gesture sequences. As the epochs progress, the model's loss decreases and accuracy improves, showing that it is effectively learning gesture transitions and refining its performance with each cycle.

These visual results confirm that the model converges over time, ensuring smoother and more accurate avatar-based sign language output during actual translation.

3.6. RESULTS AND DISCUSSION

The evaluation focuses on two key metrics:

translation accuracy and processing efficiency, measured across videos of varying durations. This section shows quantitative analyses to assess the robustness and scalability of the proposed system.

3.6.1. QUANTITATIVE EVALUATION

The below table demonstrates a quantitative For every video, the word count of transcriptions, word-level accuracy, and overall processing time were measured.

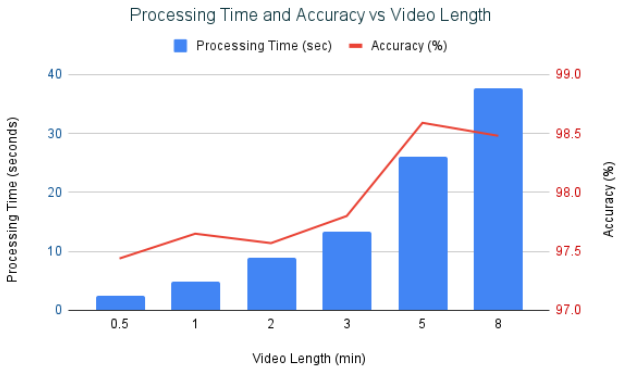
Video Length (min)	Transcribed Words	Accuracy (%)	Processing Time (sec)
0.5	55	97.44	2.56
1.0	158	97.65	4.96

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.0	331	97.57	8.84
3.0	390	97.80	13.35
5.0	435	98.59	26.09
8.0	988	98.48	37.61

With a peak at 98.59% at the 5-minute point, the data show that accuracy regularly stays above 97%, even for longer videos. Processing time scales almost linearly with video length; the longest (8-minute) video processed in just 37.61 seconds, proving the system's capacity to run far faster than real time.



A. Processing Time vs Video Length:

The bar plot indicates a quasi-linear growth of processing time with video duration. This is to be expected because the pipeline process is frame by frame and token by token. Each module—speech recognition, text preprocessing, gloss prediction, and gesture rendering—is input sequentially processed, hence scaling linearly in time with video length.

Efficiency: The longest video (8 minutes) was processed only in 37.61 seconds with an RTF of 0.0785, which confirms that the pipeline can process input at over 12× real-time speed on a hardware-based CPU.

Implication: This responsiveness makes the system suitable for use in applications such as live interpretation in a classroom, a courtroom, or in business meetings.

B. Accuracy vs Video Length:

Precision is always high, between 97.44% and 98.59%, with marginal differences across video lengths. The small variations can be traced to the following reasons:

Improved Context: As the video becomes longer, the system gains more sentence-level context, making the decoder better at choosing suitable glosses. This is why precision is at its highest at 98.59% at the 5-minute mark.

Diminishing Returns: Once some length (say, 8 minutes) has passed, context saturation can occur and precision slightly decreases, possibly

due to model fatigue or ASR drift.

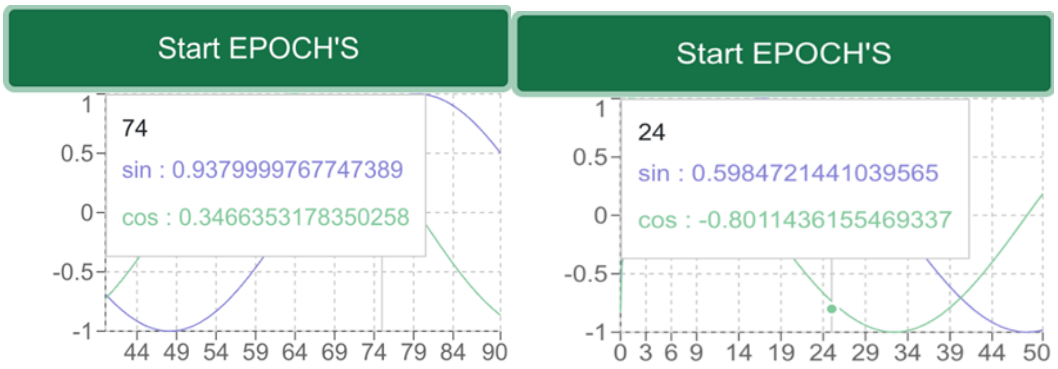
Robustness: The consistency over different lengths

shows that the attention-based Seq2Seq model performs well under long dependencies and avoids error propagation.

3.6.2. EPOCH VISUALIZATION DURING GESTURE GENERATION

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



After the user provides speech or text and triggers the sign language translation model, the base model starts producing ISL gestures in a sequence of inference epochs. In order to give visual feedback for this process, an epoch-based visualization module is activated. As shown in the above figures, sine and cosine curves are being plotted dynamically, where every point on the curve corresponds to the current number of epochs. These visualizations provide a sense-giving indication that the model is in the process of computing the input and adjusting internal weights for gesture validity. The curves themselves are not substantively significant, but they verify the commencement, continuity, and responsiveness of the model during real-time animation. This functionality adds to user experience by making it certain that gesture rendering is not a static but an adaptive, learning-based backend process.

4. CONCLUSION

This project presents an AI-powered system designed to bridge the communication gap between the hearing and Deaf or Hard-of-Hearing (DHH) communities by converting spoken video content into Indian Sign Language (ISL) using animated avatars. By combining technologies such as audio processing, speech recognition, natural language processing, and deep learning through LSTM networks, the system ensures accurate, context-aware, and grammatically structured sign translations. In contrast to conventional approaches relying on word-level or static gestures, this model generates smooth, expressive animations for sign languages that simulate real signing. The

usage of Three.js in rendering avatars, combined

with dynamic time warping and inverse kinematics, it contributes to the realism of movement. This work opens doors to inclusive, scalable digital communication with greater access for DHH individuals in education, media, and public spaces.

5. FUTURE SCOPE

The proposed system opens several avenues for future enhancement and real-world deployment. One potential direction is the integration of real-time live translation for streaming platforms, online classrooms, and video conferencing tools, which would allow Deaf and Hard-of-Hearing users to access content instantly. Additionally, expanding the system's capabilities to support multiple regional sign languages and multi-language speech input can significantly increase its inclusivity across diverse linguistic communities. Enhancing the avatar's expressiveness with facial expressions and body language cues could bring it closer to real-life human interpreters. Furthermore, incorporating feedback-based learning models where users can correct or improve translations will help the system adapt over time. With the advancement of wearable AR/VR technologies, the solution can also be extended to immersive sign language learning environments for both DHH individuals and interpreters in training. Overall, this system can evolve into a complete accessibility platform that transforms the way the DHH community interacts with digital content.

REFERENCES

[1] D. Chakladar, P. Kumar, S. Mandal, P. P. Roy, M. Iwamura, and B.-G. Kim, "3D Avatar Approach for Continuous Sign Movement Using Speech/Text," Applied Sciences, vol. 11, no. 8, pp. 1–13, Apr. 2021

[2] R. Jain, "Converting YouTube Video to American Sign Language Translation Using Convolution Neural Network and Video Processing," iJournals: International Journal of Software & Hardware Research in Engineering, vol. 10, no. 5, 2022.
[3] H. Kotha, S. D. Ponugoti, and V. Krishnan, "Audio to Sign Language Using NLTK," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 10, no. 6, pp. e404–e408, Jun. 2023.
[4] R. Bharti, S. Yadav, S. Gupta, and B. Rajitha, "Automated Speech to Sign Language Conversion using Google API and NLP," in Proc. Int. Conf. on Advances in Electronics, Electrical & Computational Intelligence (ICAEEC), Jun. 2019.
[5] Ghosh and R. Mamidi, "English To Indian Sign Language: Rule-Based Translation System Along With Multi-Word Expressions and Synonym Substitution," in Proc. 19th Int. Conf. on Natural Language Processing (ICON), New Delhi, India, Dec. 2022, pp. 123–127
[6] N. Mehta, S. Pai, and S. Singh, "Automated 3D Sign Language Caption Generation for Video," Universal Access in the Information Society, vol. 19, no. 3, pp. 725–738, 2020, doi: 10.1007/s10209-019-00668-9
[7] L. T. Nguyen, F. Schicktanz, A. Stankowski, and E. Avramidis, "Automatic generation of a 3D sign language avatar on AR glasses given 2D videos of human signers," in Proc. 1st Int. Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Virtual, Aug. 2021, pp. 71–81.
[8] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks," International Journal of Computer Vision, vol. 128, no. 4, pp. 891–908, Jan. 2020
[9] H. Monga, J. Bhutani, M. Ahuja, N. Maida, and H. Pande, "Speech to Indian Sign Language Translator," in Recent Trends in Intensive Computing, M. Rajesh et al., Eds. IOS Press, 2021, pp. 55–60. doi: 10.3233/APC210172.
[10] U. S. Prasad, K. A. Anuradha, E. Siddartha, S. S. Reddy, and N. P. Reddy, "Audio/Text to Sign Language," International Journal of Creative Research Thoughts, vol. 11, no. 5, pp. 1450–1456, May 2023.
[11] Khare, V. Kulkarni, and A. Upadhayay, "A Collaborative Augmented Reality System Based on Real Time Hand Gesture Recognition," Global Journal of Computer Science and Technology, vol. 11, no. 23, pp. 1–5, Dec. 2011.
[12] Pathak, A. Kumar, Priyam, P. Gupta, and G. Chugh, "Real-Time Sign Language Detection," Intl. J. for Modern Trends in Science and Technology, vol. 8, no. 1, pp. 32– 37, 2022.
[13] O. M. Foong, T. J. Low, and W. W. La, "V2S: Voice to Sign Language Translation System for Malaysian Deaf People," presented at the Intl. Conf. on Modern Trends in Intensive Computing, Nov. 2009.
[14] L. Goyal and V. Goyal, "Text to Sign Language Translation System: A Review of Literature," Intl. J. Synthetic Emotions, vol. 7, no. 2, pp. 20–28, Jul.–Dec. 2016.
[15] Z. Yu, S. Huang, Y. Cheng, and T. Birdal, "Sign Avatars: A Large-Scale 3D Sign Language Holistic Motion Dataset and Benchmark," in Proc. English to Indian Sign Language Rule-Based Translation System, May 2023.
[16] S. Kumar, K. Saraswat, and S. Prasad, "Audio Extraction from Video," International Journal of Research in Engineering and Science (IJRES), vol. 11, no. 3, pp. 45–49, May 2023.
[17] Pandipati and R. Praveen Sam, "Speech to Text Conversion Using Deep Learning Neural Net Methods,"Turkish J. Computer and Mathematics Education, vol. 12, no. 5, pp. 2037–2042, 2021.
[18] S. Sasavade, T. Sutar, K. Baral, and D. Kambale, "Extract the Audio from Video by Using Python," Intl. Research J. Engineering and Technology, vol. 10, no. 6, pp. 120–123, Jun. 2023.
[19] R. Zuo, F. Wei, Z. Chen, B. Mak, J. Yang, and X. Tong, "A Simple Baseline for Spoken Language to Sign Language Translation with 3D Avatars," in Proceedings of the European Conference on Computer Vision (ECCV), 2024
[20] R. M., S. A. R., S. Shristi, S. P. S., and M. Kumar B. H., "Video-Based Sign Translation Model,"