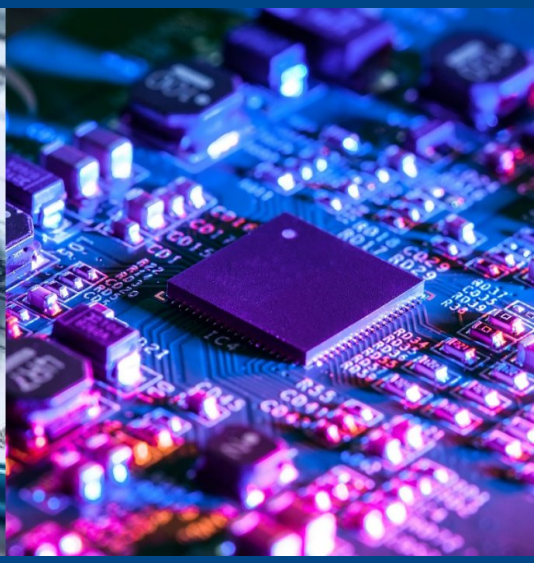
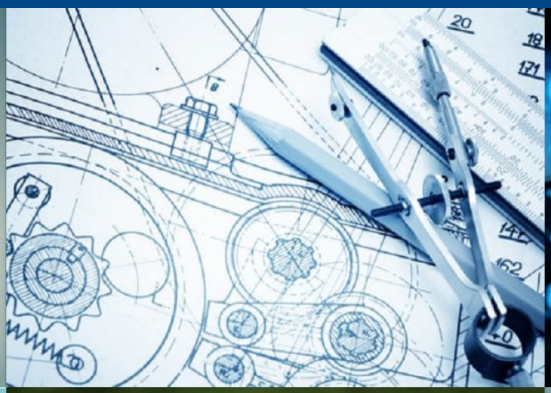
ISSN: 2582-7219

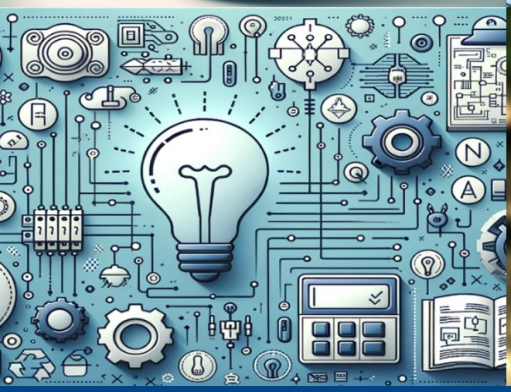




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(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.2026.0904380

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)

AI Multilingual Speech Translation Using Voice Cloning Techniques

Susmitha A[1], Dr. Mohamed Divan Masood[2]

Student, Dept. of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India[1]

Professor, Dept. of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India[2]

ABSTRACT: An AI-based multilingual speech translation system is presented in this paper with the goal of removing linguistic barriers in multimedia content. The suggested system automatically translates spoken videos into several target languages from beginning to end while maintaining context. Signal processing methods are used to extract audio features, and a transformer based model is employed for multilingual translation and precise speech-to-text transcription. A text-to-speech engine produces natural sounding voice output and beam search decoding enhances the quality of translations. Time alignment techniques are used to synchronise subtitles, guaranteeing precise display with speech segments. To create a fully localized output, the original video is combined with the translated audio an subtitles. Effective translation performance and dependable transcription accuracy are demonstrated by experimental evaluation. Applications in media localization, education, and multilingual communication platforms can all benefits from the system's scalable solution.

KEYWORDS: Multilingual Speech Translation, Neural Machine Translation, Transformer Architecture, Text-to-Speech Synthesis, Subtitle Synchronization, Artificial Intelligence, End-to-End Systems

I. INTRODUCTION

Language barriers still restrict access to information and learning resources, even in the age of digital transformation. Conventional translation methods necessitate manual intervention and separate tools for speech recognition, translation, and subtitle generation, leading to fragmented workflows and inefficiencies. Recent developments in deep learning have made significant progress in speech processing and natural language understanding possible. Transformer-based architectures have shown superior contextual modelling capabilities in sequence-to-sequence tasks like machine translation and speech recognition, and neural text-to-speech systems have improved the quality of speech synthesis. A unified AI-based multilingual speech translation system that can automate the entire process from input video to translated output is proposed in this study. The system seeks to offer scalable deployment for real-world applications and smooth multilingual communication by incorporating sophisticated algorithms into a modular yet automated framework.

II. LITERATURE REVIEW

Early translation systems had little flexibility and were based on rules. Probabilistic modelling was introduced bystatistical machine translation, but it lacked contextual awareness. By employing encoder-decoder architectures, neural machine translation increased translation accuracy. By using self-attention mechanisms, Transformer models greatly improved contextual representation.

Reliance on distinct ASR and MT modules was lessened by end-to-end speech translation frameworks. In sequence generation tasks, Beam Search increased the accuracy of decoding. In contrast to conventional concatenative systems, neural text-to-speech models enhanced naturalness. Waveform reconstruction from spectrogram representations was made possible by Griffin-Lim. Accurate synchronization was improved by subtitle alignment strategies like Dynamic Time Warping. Cross-lingual research was made easier by publicly available multilingual datasets like Common Voice and OPUS. Despite these developments, many systems still do not fully automate and integrate voice synthesis, subtitle

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)

creation, translation, and speech recognition into a single framework.classifiers have been implemented to recognize surrounding objects and deliver spoken descriptions in real time. While these systems demonstrate the effectiveness of AI-driven assistive technologies, most existing solutions are limited to a single functionality, either text recognition or object detection. Additionally, many applications lack comprehensive voice command integration, emergency assistance features, and a modular architecture that supports scalability and future expansion. These limitations highlight the need for a unified and fully integrated mobile-based assistive framework. The proposed research addresses these gaps by combining voice interaction, OCR, object detection, audio feedback, and SOS functionality within a single, scalable mobile application designed specifically for visually impaired users.

III. RESEARCH GAP

Few studies concentrate on integrating all modules into a fully automated end-to-end system, despite the fact that numerous studies have addressed individual speech translation components. Subtitle synchronisation and multilingual scalability are lacking in many current systems. Furthermore, little research has been done on the integration of voice synthesis into speech translation frameworks. By integrating synchronisation, translation, synthesis, and recognition into a single architecture, the suggested system overcomes these drawbacks.services into a single, unified mobile platform to enhance independence, safety, and usability.

IV. PROPOSED SYSTEM

In order to automatically translate spoken video content into several target languages while maintaining contextual meaning and speech continuity, the suggested system offers an AI-based end-to-end multilingual speech translation framework. The suggested architecture unifies speech recognition, neural machine translation, voice synthesis, and subtitle synchronisation into a single automated pipeline, in contrast to conventional modular systems that necessitate manual intervention between stages. To extract the audio signal, the input video is first processed. To improve signal quality, the audio is preprocessed using techniques like noise reduction and normalisation. Time-domain audio signals are transformed into frequency-domain features appropriate for speech modelling using the Fast Fourier Transform (FFT). The speech recognition module transcribes audio features into text using a Transformer-based sequence-to-sequence model. The model can better capture long-range contextual dependencies thanks to self-attention mechanisms, which also increase transcription accuracy. The speech recognition module transcribes audio features into text using a Transformer-based sequence-to-sequence model. The model can better capture long-range contextual dependencies thanks to self-attention mechanisms, which also increase transcription accuracy. A text-to-speech synthesis engine is then used to turn the translated text into speech. To create comprehensible and realistic-sounding audio output, the speech waveform is reconstructed. Synchronized subtitles are created by matching translated text with speech segments using time-alignment mechanisms to improve accessibility. The final localized output is created by combining the original video with the translated audio and subtitles using multimedia processing tools. With its scalable Python-based architecture and support for numerous regional and international languages, the suggested system allows for smooth automation from input video to translated output.

V. SYSTEM ARCHITECTURE

The modular yet integrated architecture of the suggested AI-based multilingual speech translation system is intended to accomplish total end-to-end automation. Audio preprocessing, speech recognition, multilingual translation, voice synthesis, and subtitle synchronization with output integration are the five main functional components of the architecture. A user uploads a video or audio file to the input module, which is where the system starts. After being extracted, the audio stream is sent to the preprocessing module. In this stage, noise removal and normalization are performed to improve signal clarity. Fast Fourier Transform (FFT) is applied to convert time-domain signals into frequency-domain features, enabling efficient feature representation for speech modeling. To translate processed audio features into text transcription, the speech recognition module uses a Transformer-based sequence-to-sequence architecture. By identifying contextual dependencies in thespeech signal, the attention mechanism increases the accuracy of transcription in a variety of languages. The multilingual translation module then uses a neural machine translation framework to process the transcribed text. In order to select the most likely translated output while preserving semantic consistency, beam search decoding is used during sequence generation. The translated text is then converted into audio by the text-to-speech synthesis module. Intelligible voice output in the chosen target language is produced by speech waveform generation. Time-alignment mechanisms that match translated text segments with corresponding

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904380

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)

speech timestamps are used to perform subtitle synchronization in order to improve user accessibility. Lastly, the output integration module creates the final localized output by combining the original video with the translated audio and subtitles. The system is appropriate for multilingual communication applications in the real world because of its modular design, which guarantees scalability, flexibility, and effective integration of extra languages or sophisticated models.

VI. METHODOLOGY

The suggested multilingual speech translation system uses a structured pipeline that combines multimedia integration, neural sequence modeling, and signal processing.

a. Audio Processing

The audio stream is extracted by processing the input video. To improve clarity, the audio signal is normalized and noise-reduced. Time-domain audio signals are transformed into frequency-domain representations using the Fast Fourier Transform (FFT). The speech recognition model receives structured input from these spectral features.

b. Speech Recognition

For automatic speech recognition, a Transformer-based sequence-to-sequence model is used. Audio features are processed by the encoder, and text transcription is produced by the decoder. The attention mechanism improves transcription accuracy across languages by enabling contextual understanding of speech patterns.

c. Multilingual Translation

The transcribed text is passed to a multilingual neural machine translation model. The model generates translated sentences using contextual sequence modeling. Beam Search decoding is applied to select the most probable translation sequence.

d. Voice Synthesis

The translated text is converted into speech using a text-to-speech synthesis engine. Spectrogram representations are generated and converted into waveform audio signals to produce intelligible speech output.

e. Subtitle Synchronization

Segments of subtitles are produced using translated text and synchronised with speech timestamps. Accurate synchronisation between spoken audio and displayed subtitles is guaranteed by time-alignment techniques.

ALGORITHM

1) Audio Feature Extraction Algorithm– The preprocessing module transforms time-domain speech signals into frequency-domain representations using the Fast Fourier Transform (FFT).

$(k) = N^{\Sigma^{\rightarrow}}_{n=0\rightarrow x}(n)^{-j2\ \Pi kn/N}$ (1) Effective spectral feature extraction for speech recognition is made possible by this transformation.

2) Speech Recognition Algorithm – For transcription, a Transformer-based sequence-to-sequence model is employed. The definition of the attention mechanism is:

$Attention(\ ,K,V) = softmax(Q\ K^{\overline{T}}/\sqrt{d_k})\ V$ (2) This enhances transcription accuracy and makes contextual dependency modelling possible.

3) Translation Algorithm – Neural sequence modelling is used in the multilingual translation module. The best translation sequence is chosen by beam search decoding:

$$P(Y|X) = \prod_{t=1}^{T} P(Yt|y<t,X) \quad (3)$$

The number of candidate sequences kept during decoding is determined by beam width.

4) Voice Synthesis Algorithm – Using spectrogram features, the text-to-speech module reconstructs the waveform. The Griffin-Lim algorithm produces audible speech output by iteratively estimating phase information.

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)

5) Subtitle Alignment Algorithm – By reducing the cumulative alignment distance, Dynamic Time Warping (DTW) aligns speech timestamps with subtitle segments: DTW(i,j) = d(i,j) + min(DTW(i-1,j),DTW(i,j-1),DTW(i-1,j-1)) (4)

VII. IMPLEMENTATION

Python was used as the main development environment for the proposed multilingual speech translation system because of its broad support for multimedia integration, artificial intelligence, and natural language processing. Using the PyTorch framework and Transformer-based architectures made possible by sophisticated NLP libraries, the system incorporates deep learning models for speech recognition and multilingual translation. To improve feature representation, audio streams are preprocessed using normalisation and frequency-domain transformation techniques after being extracted from input videos. A multilingual neural machine translation framework with optimised decoding strategies generates the translated output after the speech recognition module uses sequence-to-sequence modelling to convert audio signals into text. To create understandable speech in the target language, the translated text is further processed using a neural text-to-speech engine. To guarantee synchronisation with speech segments, timestamp alignment mechanisms are used to automatically generate subtitle files. To create a fully localised output, multimedia processing tools are used tocombine translated audio and subtitles with the original video stream. Future developments can incorporate more languages, optimised models, and real-time processing improvements thanks to the modular implementation's scalability.

VIII. EXPERIMENTAL RESULTS

A multilingual speech dataset with a variety of audio samples in supported languages, such as English, Tamil, Hindi, Telugu, and Malayalam, was used to experimentally assess the suggested multilingual speech translation system. Subtitle synchronisation, translation quality, transcription accuracy, and overall system efficiency were the main areas of assessment. Word Error Rate (WER), which measures transcription errors such as substitution, insertion, and deletion, was used to gauge speech recognition performance. The Transformer-based architecture's successful contextual modelling and precise speech-to-text conversion were demonstrated by the system's low WER values across test samples. The BLEU score, which measures the n-gram similarity between generated translations and reference texts, was used to evaluate the translation quality. The multilingual translation module's competitive BLEU scores attested to its grammatical accuracy and semantic preservation across various language pairs. When compared to greedy decoding techniques, beam search decoding helped to improve translation fluency. The voice synthesis module's intelligibility, clarity, and consistency of pronunciation were assessed qualitatively. With proper prosody and timing in the target language, the generated speech output was intelligible and clear. Through timestamp alignment testing, the accuracy of subtitle synchronization was confirmed, showing that there was little delay between spoken output and subtitle display. In order to assess the efficiency of end-to-end processing, system latency was also measured. According to the results, the integrated architecture offers scalable performance and automated multilingual translation with less manual intervention. The robustness, dependability, and practicality of the suggested system for multilingual video localization tasks in the real world are confirmed by experimental results.

IX. DISCUSSION

Experimental results show that the suggested multilingual speech translation system successfully combines voice synthesis, subtitle synchronization, and speech recognition neural translation into a single end-to-end framework. The transformer-based architecture's ability to capture contextual dependencies in multilingual speech signals is confirmed by the low word error rate seen during transcription. Strong semantic preservation and grammatical consistency across supported language pairs were demonstrated by the translation module's competitive BLEU scores. When compared to simple decoding techniques, the use of Beam Search improved the generated translations' fluency and coherence. The voice synthesis component improved accessibility for non-native speakers by producing understandable speech outputs. By ensuring precise correspondence between audio and displayed text, time-based synchronization for subtitle alignment enhanced the user experience. Additionally, modular architecture showed scalability. permitting the addition of new languages and model enhancements without requiring a structural overhaul.

However, background noise, speaker accents, and audio quality can all affect how well a system performs. When using limited computational resources or longer video inputs, processing latency may increase. Notwithstanding these

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904380

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)

drawbacks, the framework as a whole is reliable and useful for automated communication systems and multilingual video localization.

ADVANTAGES

Compared to conventional translation frameworks, the suggested multilingual speech translation system has a number of noteworthy advantages. First of all, it offers total end-to-end automation, doing away with the necessity for human intervention in the stages of voice synthesis, translation, speech recognition, and subtitle creation. This integrated architecture lowers operational complexity and increases processing efficiency. Second, the system makes use of Transformer-based neural models, which improve transcription and translation accuracy across multiple languages and improve contextual understanding. By producing output sequences that are coherent and semantically correct, Beam Search decoding further improves the quality of translation.Third, voice synthesis improves accessibility for users who prefer audio output by enabling speech-to-speech translation. Furthermore, synchronized subtitle generation improves usability for multilingual audiences and those with hearing impairments. Lastly, the system is flexible for media localization, education, and international communication applications due to its modular and scalable architecture, which makes it simple to incorporate new languages, updated models, and real-time processing improvements.

X. LIMITATIONS

The suggested multilingual speech translation system has some drawbacks despite its efficiency. First, background noise, speaker accents, and pronunciation variations can all impact speech recognition performance, increasing the likelihood of transcription errors. Second, because there is less training data available for low-resource languages, translation accuracy may differ. Complex idiomatic expressions and domain-specific terminology may still present difficulties, despite the fact that transformer-based models enhance contextual understanding.Furthermore, the text-to-speech model determines the quality of synthesized speech, which may not accurately retain speaker identity or emotional tone. Real-time deployment on low-end hardware devices may be limited by the system's high computational resource requirements, especially for deep learning model inference. Additionally, longer videos may result in higher processing latency.These drawbacks point to areas that could use further development in real-time system improvement, dataset expansion, and model optimization.

XI. FUTURE ENHANCEMENT

Real-time deployment on low-end hardware devices may be limited by the system's high computational resource requirements, especially for deep learning model inference.Additionally, longer videos may result in higher processing latency.

These drawbacks point to areas that could use further development in real-time system improvement, dataset expansion, and model optimization. The system's high computational resource requirements, particularly for deep learning model inference, may limit real-time deployment on low-end hardware devices. Additionally, processing latency may increase with longer videos.These shortcomings highlight areas that require more work in the areas of dataset expansion, model optimization, and real-time system improvement.

XII. CONCLUSION

In order to get around language barriers in multimedia content, an AI-based end-to-end multilingual speech translation system was presented in this paper. The suggested framework combines voice synthesis, neural machine translation, speech recognition, and subtitle synchronization into a single automated pipeline. The system accomplishes precise transcription and context-aware translation in a number of languages by utilizing Transformer-based sequence modeling and optimized decoding techniques. The user experience and accessibility are improved by the addition of synchronized subtitle generation and text-to-speech synthesis. Reliable performance in terms of subtitle alignment, translation quality, and transcription accuracy was shown by experimental evaluation. Scalability and adaptability for upcoming enhancements are guaranteed by the modular architecture. All things considered, the suggested system advances intelligent automated translation technologies by offering a viable and effective solution for multilingual video localization, e-learning platforms, and international media communication.

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)

REFERENCES

[1] Speech Captioning System, "Deep Learning-Based Speech Captioning System for Document Accessibility," 2024.
[2] Q. Gao et al., "VI-OCR: Visually Impaired Optical Character Recognition," Scientific Reports, 2025.
[3] S. Saleem, M. A. Khan, and A. Abbas, "Smart assistive system for visually impaired using OCR and speech synthesis," IEEE Access, vol. 9, pp. 123456–123468, 2021.
[4] E. Salesky, Y. Jia, J. Mahadeokar, and K. Kirchhoff, "Direct speech-to-speech translation with discrete units," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2021.
[5] B. Zhang, D. Zhang, Z. Chen, et al., "Neural speech translation with pre-trained large language models," in Proceedings of INTERSPEECH 2023, 2023.
[6] A. Radford et al., "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2023.
[7] A. Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
[8] M. Ott et al., "Scaling Neural Machine Translation," in Proceedings of the Workshop on Neural Machine Translation, ACL, 2018.Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," in Proceedings of INTERSPEECH, 2017.
[9] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2,
pp. 236–243, 1984.

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