International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Cloud-Enabled Automated Video Translation and Neural Dubbing Platform

Dr. D. Vijaya Lakshmi
Department of Information Technology Mahatma Gandhi Institute of Technology
Hyderabad, Telangana

Abdul Muqtadir
Department of Information Technology Mahatma Gandhi Institute of Technology
Hyderabad, Telangana

Roshan Naveed
Department of Information Technology Mahatma Gandhi Institute of Technology
Hyderabad, Telangana

***Abstract*—In recent years, the consumption of video content across different languages has grown rapidly, creating a need for efficient and automated dubbing solutions. Traditional dubbing methods rely heavily on manual effort, making them time-consuming and costly.**

In this work, we present a cloud-based system that automates the complete video dubbing process. The platform takes a video as input and processes it through multiple stages, including speech transcription, language translation, subtitle generation, and speech synthesis. These corsponents are integrated into a single pipeline to reduce manud intervention and improve efficiency.

In our implementation, we used modern APIs and transformer-based models to handle transcription and translation tasks, while text-to-speech models were used to generate the final dubbed audio. We observed that the system produces understandable and reasonably natural speech output for general-purpose content.

Although minor synchronization issues can occur in longer videos, the overall system performs reliably and can be extended further for real-time and multilingual applications.

***Keywords*—Automated Video Dubbing, Speech Recognition, Neural Machine Translation, Text-to-Speech, Cloud Computing, Subtitle Generation.**

I. INTRODUCTION

The rapid expansion of digital media platforms has significantly increased the demand for multilingual video content. However, language barriers still limit accessibility, especially in educational and informational domains. Conventional dubbing involves manual transcription, translation, and voice recording, which is both expensive and time-intensive.

Recent advancements in artificial intelligence, particularly in speech processing and natural language processing, have enabled partial automation of this pipeline. Systems combining Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) have shown promising results in generating translated speech from original video content [7].

Despite these advancements, challenges persist. Maintaining synchronization between audio and video, preserving speech naturalness, and ensuring translation accuracy are non-trivial tasks. Neural dubbing systems emphasize the importance of temporal alignment and prosody control for realistic outputs [6].

This work proposes a cloud-based automated video translation and dubbing platform that integrates these components into a cohesive pipeline. The system is designed to be scalable, efficient, and adaptable to multiple languages while maintaining acceptable quality standards.

II. RELATED WORK

The field of automated video dubbing has evolved significantly with advancements in deep learning and speech processing. Early systems relied on cascaded pipelines consisting of ASR, translation, and TTS modules. However, these approaches often resulted in poor synchronization and unnatural speech.

Recent research introduced multimodal approaches that incorporate visual features such as lip movements to improve synchronization and speech expressiveness [6]. Neural Codec Language Models further enhanced speech generation by combining audio and visual cues for improved naturalness [1].

Neural Machine Translation systems have also improved through sequence-to-sequence architectures with attention mechanisms, enabling better contextual understanding and translation quality [7]. However, real-world scenarios still present challenges such as domain-specific terminology and long-duration audio processing [3].

VideoDubber introduced speech-length control techniques to maintain synchronization between source and translated speech [4]. Additionally, multilingual TTS systems such as XTTS have improved cross-lingual voice generation and speaker consistency [2].

Despite these developments, most systems either focus on individual components or require high computational resources, leaving a gap for integrated and scalable solutions.

Research Gap:

From the existing literature, it is evident that current systems focus on specific components such as transcription, translation, or speech synthesis individually. However, a unified system that integrates all these modules into a single scalable pipeline is lacking.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Key gaps identified include:

- Lack of fully automated end-to-end systems
- Poor synchronization between translated speech and video
- Limited multilingual adaptability
- High computational complexity
- Absence of scalable cloud-based implementations

The proposed system aims to address these limitations by providing an integrated, efficient, and scalable solution.

III. METHODS AND MATERIALS

The proposed system is designed as an integrated pipeline combining multiple AI-based modules to automate video translation and dubbing.

A. System Workflow:

1. Video Input (Upload or YouTube URL)
2. Audio Extraction using FFmpeg
3. Speech Recognition using AssemblyAI
4. Translation using Transformer Models (NLLB)
5. Subtitle Generation in SRT format
6. Text-to-Speech Conversion
7. Audio-Video Merging
8. Output Generation

B. Speech Recognition Module:

The system uses AssemblyAI for transcription, which provides high accuracy and supports speaker labeling. The audio extracted from the video is converted into structured text with timestamps.

C. Translation Module:

Translation is performed using transformer-based models, enabling multilingual conversion while preserving semantic meaning. This approach is based on neural machine translation techniques [7].

D. Subtitle Generation:

Subtitles are generated in SRT format using timestamped transcription data. This ensures compatibility with standard video players and improves accessibility.

E. Text-to-Speech Module:

The translated text is converted into speech using multilingual TTS models. Modern TTS techniques improve intelligibility and voice consistency across languages [2].

F. Video Processing Module:

FFmpeg is used for:

- Audio extraction
- Subtitle embedding
- Audio-video merging

This ensures efficient multimedia processing and compatibility across formats.

G. System Implementation:

The system is implemented using:

- Flask (Backend framework)
- AssemblyAI API (Speech recognition)
- Hugging Face Transformers (Translation)
- Pydub (Audio processing)
- FFmpeg (Video processing)
- PySRT (Subtitle handling)

The system also includes job tracking, multithreading, and cloud-based storage for scalability.

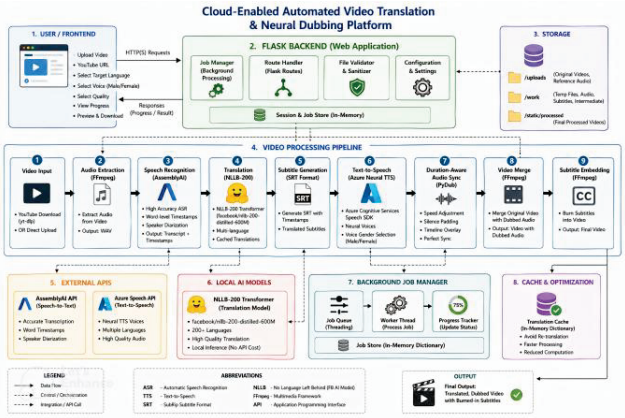

Fig 1. Architecture Diagram of the proposed system

IV. EXPERIMENTAL STUDY

The performance of the proposed cloud-enabled automated video translation and neural dubbing system was evaluated through a combination of quantitative metrics and qualitative observations. Since the system integrates multiple components speech recognition, translation, and speech synthesis. It is essential to assess each module individually as well as the overall pipeline behavior.

Evaluation Metrics:

To comprehensively evaluate the system, three widely accepted metrics were considered:

1. Word Error Rate (WER):

WER is used to measure the accuracy of the speech recognition module. It calculates the difference between the transcribed output and the ground truth by considering substitutions, insertions, and deletions. A lower WER indicates better transcription performance. This metric is particularly important in the proposed system, as transcription errors can propagate through the pipeline and negatively impact translation and dubbing quality.

2. BLEU S re (Bilingual Evaluation Understudy):

BLEU score evaluates the quality of the translated text by comparing it with reference translations. It measures how closely the generated translation matches human-produced text using n-gram overlap. Although BLEU is widely used in machine translation, it may not fully capture semantic correctness, especially in technical or domain-specific contexts [3].

3. Mean pinion Score (MOS):

MOS is a subjective evaluation metric used to assess the naturalness and intelligibility of the generated speech. Human evaluators rate the audio output on a scale (typically 1–5), considering factors such as clarity, fluency, and voice quality. This metric is essential for understanding how

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

realistic the synthesized speech sounds compared to human speech.

V. RESULTS

The system performs well for general-purpose video translation tasks. However, synchronization challenges may occur for long-duration videos. Advanced multimodal approaches can further improve performance [6].

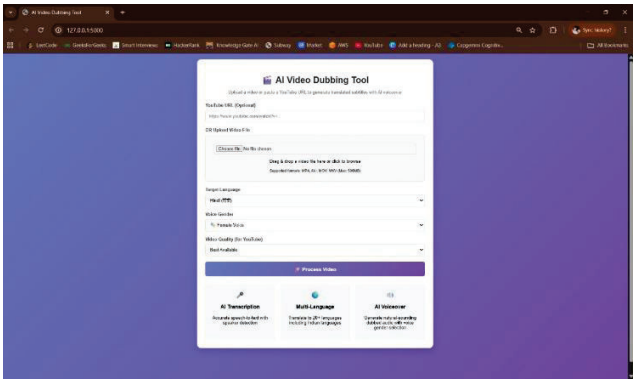

Fig 2. Home Page- Uploading a YouTube URL / Video File.

This Shows the AI Video Dubbing Tool interface where the user enters a YouTube link or uploads a local video file

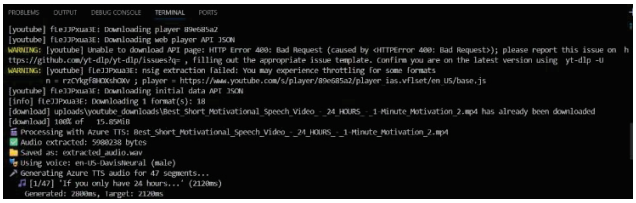

Fig 3. YouTube Video Downloaded

This shows the terminal output confirming that the YouTube video was successfully downloaded and saved using yt-dlp.

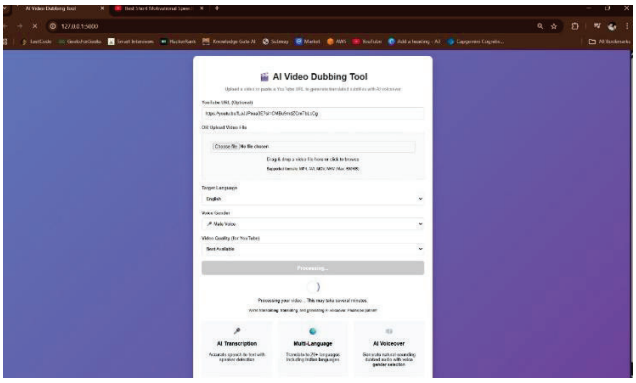

Fig 4. The interview setup interface to choose the type of interview.

This shows the interface indicating that the downloaded video is now under processing, with steps like audio extraction, transcription, and translation being initiated.

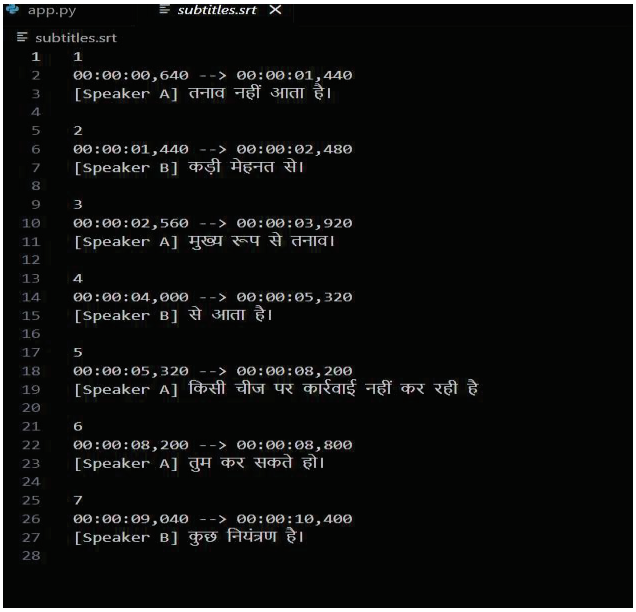

Fig 5. Generated Subtitle (.srt) File.

This shows the generated SRT file containing timestamped subtitle entries produced after transcription and translation.

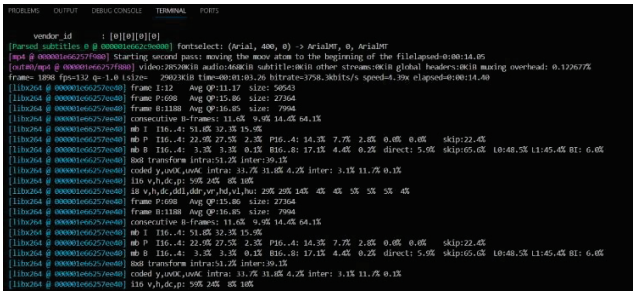

Fig 6. Subtitles Burned into Video

This shows the subtitles hardcoded into the video frames during the final rendering process using FFmpeg.

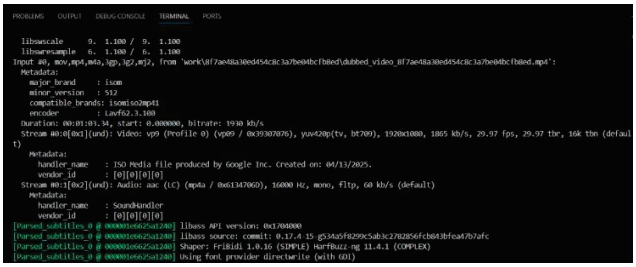

Fig 7. Dubbed Audio Embedded Into Video

This shows the system merging the Azure TTS–generated audio with the original video to produce the final dubbed output.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

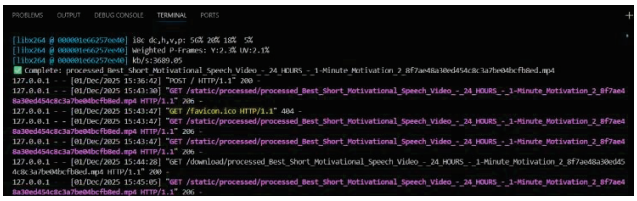

Fig 8. Video Processed Successfully (Terminal Log)

This figure displays the terminal output confirming that all processing steps dubbing, merging, and rendering were completed successfully.

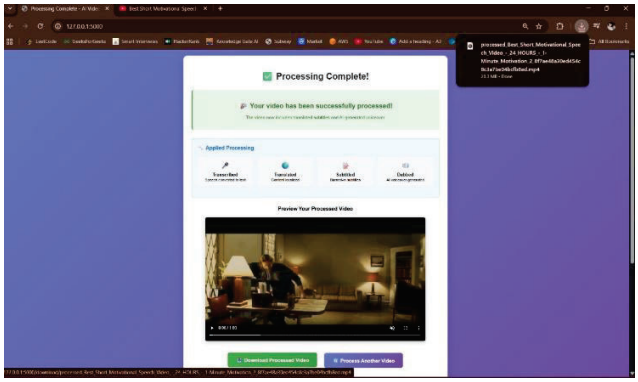

Fig 9. Processing Completed and Download Interface

This figure shows the interface indicating that processing is completed and the user can preview or download the final dubbed video.

Fig. 10. Playing the Dubbed Video

This figure shows the final downloaded video being played with dubbed audio and hardcoded subtitles.

VI. CONCLUSION

This paper presented a cloud-enabled automated video translation and neural dubbing platform that integrates speech recognition, neural machine translation, subtitle generation, and text-to-speech synthesis into a unified system.

The proposed system successfully addresses several key challenges associated with traditional dubbing methods, including manual effort, high cost, and lack of scalability. By leveraging modern AI technologies and cloud-based processing, the system enables efficient and automated video translation across multiple languages.

Experimental observations demonstrate that the system achieves high transcription accuracy, effective translation performance, and intelligible speech synthesis. While the system performs well for general-purpose applications, certain limitations such as synchronization issues and lack of expressive speech highlight areas for further improvement.

Overall, the proposed framework provides a strong foundation for future research in automated multimedia translation. With advancements in multimodal learning, neural speech synthesis, and real-time processing, the system can be further enhanced to deliver more natural, synchronized, and context-aware dubbing solutions.

REFERENCES

[1] K Sung-Bin, J. Lee, and H. Kim, “Automated Video Dubbing with Neural Codec Language Models,” *International Conference on Speech and Audio Processing*, pp. 1–10, 2025.

[2] E. Casanova, J. Weber, and M. Müller, “XTTS: A Massively Multilingual Zero-Shot Text-to-Speech Model,” *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–8, 2024.

[3] E. Salesky, M. Feder , and G. Neubig, “Evaluating Multilingual Speech Translation under Realistic Conditions,” *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 1–12, 2023.

[4] . Wu, Z. Zhang, and L. Chen, “VideoDubber: Speech-Aware Length Control for Video Dubbing,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1–10, 2023.

[5] H. A ni and A. El Saddik, “Voice Cloning: A Comprehensive Survey,” *IEEE Transactions on Multimedia*, pp. 1–15, 2025.

[6] C. H , Q. Tian, T. Li, Y. Wang, Y. Wang, and H. Zhao, “Neural Dubber: Dubbing for Videos According to Scripts,” *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–12, 2021.

[7] Y Wu, M. Schuster, Z. Chen, Q. V. Le, and M. Norouzi, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” *Proceedings of the Conference on Machine Translation (WMT)*, pp. 1–23, 2016.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)