

e-ISSN: 2583-9241

Volume 09 Issue 03
Sep-Dec, 2026

***Corresponding**

Author: S. Y. Inamdar,
Assistant Professor,
Department of Computer
Science and Engineering,
AGTI's Dr.Daulatrao
Aher College of
Engineering, Karad,
Maharashtra, India

AI-Based Live Translation and Dubbing Systems for Multilingual Video Content

¹S. Y. Inamdar, ²Dhanashree Bhoite, ³Ananya
Patil, ⁴Janhavi Ghare

¹Assistant Professor, ^{2,3,4} Student, Department of
Computer Science and Engineering,
AGTI's Dr.Daulatrao Aher College of Engineering,
Karad, Maharashtra, India

ABSTRACT

The language diversity in India is a big challenge when it comes to access to digital video content in various geographical areas. News, educational materials, and multimedia materials are also usually readable only in a particular language, which restricts their access to a greater audience. The recent developments in the field of Artificial Intelligence allowed automated speech recognition, machine translation, text-to-speech synthesis, and synchronization of lips, which means that multilingual video translation is now possible. In this review paper, we will discuss the current AI-based systems and techniques that are in use to provide live translation and dubbing of video materials. The paper evaluates speech-to-text solutions, neural machine translation systems, voice synthesis systems, subtitle generation systems, and lip-sync systems. A comparative analysis of the current solutions reveals its advantages and constraints especially with regards to Indian languages. The review outlines the major gaps in the research, which are absence of integrated systems, and offline limitations as well as difficulties in processing in real time. On the basis of these observations, the current paper addresses the necessity of an interdisciplinary AI-based structure to enhance access, efficiency, and scalability of multilingual video translators systems.

Keywords:-Artificial Intelligence, Speech-to-Text, Machine Translation, Text-to-Speech, Video Dubbing, Subtitle Generation, Multilingual Systems.

1. INTRODUCTION

The country of India is diverse in terms of language and dialect that creates a huge problem in digital media communication and information accessibility. Video-based information has emerged as the most popular in terms of communication in education, entertainment, and social sites. But language barriers limit the usefulness of such contents to the audiences not acquainted with the original language. Conventional ways of manual dubbing and subtitle production are time consuming, costly as well as entails the potential of human skilled resources.

As an embodiment of the fast development of the Artificial Intelligence, the automated speech recognition, translation, and voice synthesis are becoming a subject of significant interest. The speech-to-text AI-based system can be used to transform spoken text to textual form, whereas neural machine-translation models may be used to translate text between two different languages effectively. Text-to-speech systems go to further improve these systems, by yielding spoken text in the target language. The recent advances in lip-sync and video processing methods are meant to enhance the look realism of the dubbed video by aligning the audio of the speech of the speaker with the lips. Irrespective of such developments, the majority of the available platforms are based on individual features, including translation or dubbing, and are not combined smoothly with each other. Also, encouragement of Indian regional language, offline processing and real-time performance is less.

The presented review paper offers an overview of existing AI-based live translation and dubbing methods, examines their methodology, as well as highlights research gaps. The research seeks to offer information on the creation of multilingual video translation systems

that are more accessible and accessible to integrate into various language settings.

2. KEY CONTRIBUTIONS

This review describes AI-based multilingual video translation systems, including analysis of the main parts, including ASR, NMT, TTS, subtitle generation, and lip synchronization, and in particular problems in Indian languages.

A comparative analysis indicates such limitations as latency, synchronization problems, and poor support of low-resource languages. The paper pinpoints deficiencies in complete integrated systems and recommends future advancements in real-time execution, multilingual databases, emotion recognizant TTS, lip-sync fidelity, and off-line execution to improve its accessibility.

3. LITERATURE REVIEW

As it has been mentioned, recent artificial intelligence breakthroughs made a considerable step forward in multilingual video translation and dubbing. The initial studies were oriented at the enhancement of multilingual speech-to-text accuracy. Wu et al. presented a speech-conscious machine translation strategy to control the output volume to achieve superior audiovisual consistency [1], and Bigioi et al. gave an overview of the multilingual dubbing technologies and summarized the limitations of audiovisual consistency [2].

As neural networks developed, scientists resolved the issue of duration mismatch in dubbing. Subramanian et al. suggested a length-conscious speech translation model, which is phoneme- based and in attempt to enhance temporal alignment [3]. Choi et al. proposed Dub-S2ST, a speech-to-speech translation system which does not rely on text and retains the speaker identity and duration, [4] and Cui et al. improved

the synchronization by adopting fine-grained segment- level duration alignment [5].

Visual cues have also been used in recent studies to enhance lip synchronization. Wang et al. introduced SyncVoice, a vision-based TTS system to enhance the consistency of lips and audio [6], and Won

et al. created an end-to-end multilingual dubbing system based on big language models that uses duration-based translation [7]. More feasible practices are also evident, such as Kannoja introduced a real-time multi-lingual dubbing generator system [8]

Table 1:-Summary of Literature Review

Sr. No.	Author & Year	Title	Key Contribution	Limitations
1	Wu et al., 2022	VideoDubber: Speech-Aware MT	Length-controlled translation for dubbing	Limited datasets
2	Bigioi et al., 2023	Multilingual Video Dubbing Review	Overview of dubbing and lip-sync technologies	Broad survey only
3	Subramanian et al., 2025	Length-Aware Speech Translation	Phoneme-based duration control	Limited language pairs
4	Choi et al., 2025	Dub-S2ST	Textless speech-to-speech dubbing	High model complexity
5	Cui et al., 2025	Duration Alignment Optimization	Segment-level synchronization improvement	Computational cost
6	Wang et al., 2025	SyncVoice	Vision-augmented text-to-speech synthesis	Vision dependency
7	Won et al., 2025	LLM-Based Auto Dubbing	End-to-end multilingual dubbing	LLM latency
8	Kannoja, 2025	Gen-AI Multilingual Dubbing	Web-based real-time dubbing	Limited evaluation

4. COMPARATIVE ANALYSIS

Current video translation and dubbing software is primarily devoted to a single element of speech recognition, translation, or subtitle generation. The conventional methods are manual dubbing and simple automated software that are time consuming and expensive. New AI systems are neural machine translators and text-to-speech, which enhance the accuracy of translation and naturalness of voices. Nevertheless, audio-video synchronization remains a challenge in many systems particularly

when the original speech and translated speech are not of equal length. The existing solutions are mostly in the form of partially integrated pipelines, and they need dedicated tools to translate, dub and generate subtitles. They are also limited in the use of the cloud dependency and ability to work offline. The language support of the Indian regional ones is not developed enough because of the variety of accents and the absence of training data, which must be noted and underlines the necessity of a single and accessible AI- based system.

Table 2:-Comparative Analysis.

Aspect	Traditional Systems	AI-Based Systems
Translation Method	Manual / Statistical	Neural Translation
Audio Dubbing	Manual	AI-based TTS
Subtitle Generation	Manual	Automatic
Lip Synchronization	Not available	Limited
System Integration	Separate tools	Partial integration
Offline Support	Not supported	Limited
Indian Language Support	Very limited	Moderate

5. RESEARCH GAPS

Although AI-based video translation and dubbing have gained a lot of development, there are multiple research gaps. The first and the most obvious limitation is the lack of an end-to-end system integrating speech recognition, translation, dubbing, generation of subtitles and audiovisual synchronisation in one framework. The solutions that are in place use individual modules, which result in inefficiencies and synchronization problems.

The encouragement of regional languages spoken by Indians is also weak especially of low resource languages, various accents and dialects. The absence of big and high quality data also has an impact on the model accuracy and generalization. The real-time processing is still challenging because of the high computational needs, and most of the systems are heavily reliant on cloud computing, which limits the ability to operate offline.

Also, time error between the speech and video being translated can influence the timing of lips and the timing of subtitle. Such ethical issues as the abuse of voice cloning and the restrained feelings in voice synthesis are also in need of additional research. These gaps need to be addressed in order to come up with

scalable and reliable multilingual video translation systems.

6. PROBLEM STATEMENT

The multilingual accessibility has become a growing demand due to the rapid growth of digital video content in the field of education, media, and common communication. In places with linguistic diversity such as India, majority of video content is provided in dominant languages, thus posing obstacles to access information, and the digital divide is continued to widen.

Though AI technologies, including speech recognition, machine translation, text-to-speech, and lip synchronization, have made it possible to translate videos automatically, the current state of the art is characterized by the lack of support of the Indian regional languages as a single module that can be used to achieve the desired effect.

The reasons include the low-resource datasets, speech accent variability, dialect diversity, and code-mixed speech. Otherwise, real-time processing and offline deployment are limited by computational and accuracy of synchronization. Thus, it is necessary to have a single, scalable, and linguistically adaptable system that could result in accurate and efficient multilingual video translation.

7. PROPOSED SYSTEM

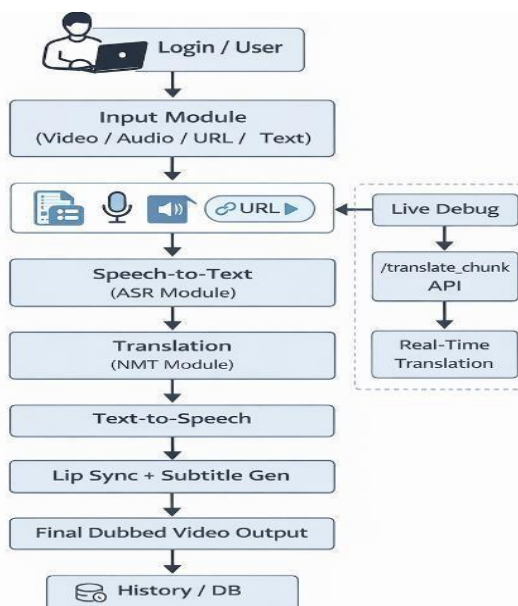


Fig.1:-Flowchart of Proposed System

The suggested AI Live Translation and Dubbing Tool is a multilingual implementation that will process the video, audio, media based on URLs, text input, and live stream of translation. This system starts with the authentication of the user and then selection of the input. In the case of multimedia inputs, a speech is captured and transformed into a text with the help of an Automatic Speech Recognition (ASR) module. Neural Machine Translation (NMT) is used to translate the extracted text into the target language. The text that has been translated will be turned into dubbed audio with the help of Text-to-Speech (TTS) synthesis. Audio-video synchronization makes sure that there should be appropriate alignment between the original video and the synthesized audio. The creation of subtitles is automatic and is incorporated into the final video. Live debugging module permits a translation in real time in chunks via an API interface. All the products are placed in a database to be accessed and downloaded in future.

8. METHODOLOGY

The system is based on a modular processing pipeline. The input can be a video or audio, URL, text or live mode (after user login the input is given to the system). Video and audio input In the case of video and audio, the audio stream is removed and the speech-to-text unit is used to get text as the source. The remainder of the text obtained is translated by a neural translation model. A TTS engine is then used to synthesize the translated text into audio to produce dubbing. A step on audio-video synchronization is used to modify the length of the audio that was produced and align it with the original video. At the same time, the subtitles are created in the form of SRT and placed within the video. In the case of live mode, the text chunks are handled with a translation API to generate real-time output and not handle the entire media. Lastly, all the processed outputs and metadata are put in the system database to be scalable and reusable.

9. RESULT AND DISCUSSION

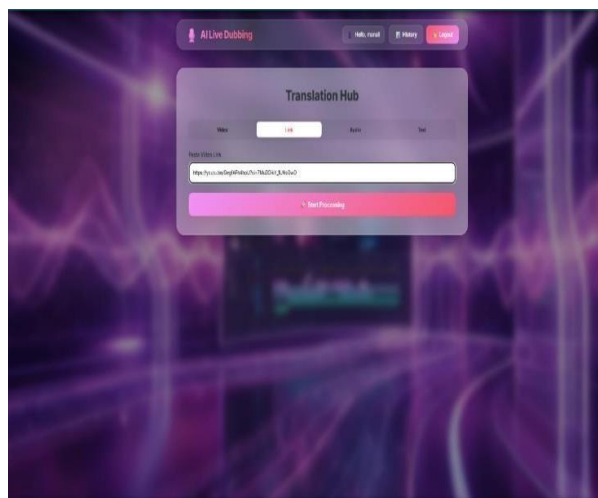


Fig.2:-Input Interface

The AI Live Translation and Dubbing system was tested on the indicative multimedia Indian language contents and different forms of inputs were taken, such as video links, uploaded videos and audio files. Translation Hub interface allowed users to input data and take a step towards processing smoothly. The system was able to retrieve audio of the input medium and translate the audio content to text with the use of the speech-to-text module.



Fig.3:-Sample Input Video Frame

Neural machine translation was applied to extract text and translate it to the English language without losing its semantic meaning. Auto generation of subtitles produced matched time stamps on SRT files. Text-to-speech synthesis provided the clear audio at the English language with the original video. The end result proved to have a good synchronization of audio, subtitles, and video frame, which improved on the viewing experience.

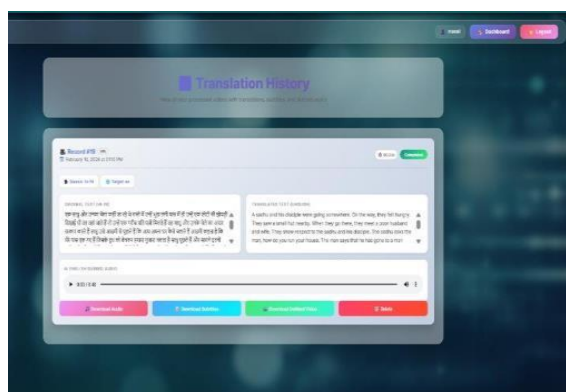


Fig.4:-Translation History Page

Live AI Debugger interface exemplifies speech recognition and translation in the real-time and with a minimal latency. The system manages to capture live video image, create instant- translated subtitles, and it presents processing logs that could

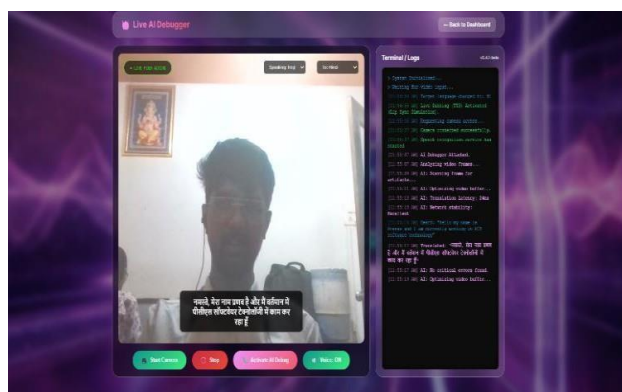


Fig.5:-Live AI Debugger interface

monitor the work of the system. All the data was stored in the Translation History module where it contained original text, translated text, dubs, subtitle files, and the completed videos.

10. PERFORMANCE ANALYSIS

1. Video Duration vs Processing Time

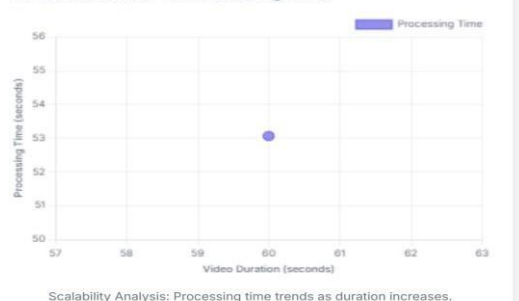


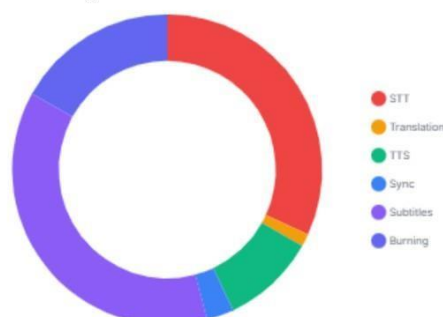
Fig.6:-Video Duration vs Processing Time Distribution

The scalability analysis shows that the overall processing time is proportional to the duration of input video and, thus, there are no fluctuations in the behavior of the computational process and the linear increase in the performance. There were no observed sharp latency spikes, which is evidence of predictable system behavior. Time distribution analysis done in modules shows that the most computationally demanding processes are the Speech-to-Text (STT) and the subtitle generation processes, and then the audio-video synchronization and the subtitle embedding processes. On the contrary, neural translation and text-to-speech modules have a relatively lower latency. The obtained results confirm the effectiveness of the suggested pipeline with a reasonable processing delay to be used in practice.

11. CONCLUSION

The review paper has discussed recent developments in AI-based video translation and dubbing systems and how they can help to break language barriers in multimedia materials. Some of the technologies that were discussed in the study include speech-to-text, machine translation, text-to-speech synthesis, subtitle generation, and

2. Processing Time Distribution



Stage Breakdown: Average time spent in each pipeline stage.

Fig.7:-Module-wise Processing Time

audiovisual synchronization. The analysis revealed that, existing systems have partial solutions, but in most cases fail to integrate smoothly, go offline, and deliver good support to the Indian regional languages. The analyzed methodologies indicate the increasing opportunities of artificial intelligence to streamline the video localization process. All in all, the results demonstrate the necessity of a comprehensive and scalable model that can effectively translate, dub and subtitle the content of the multilingual video in order to enhance its accessibility in the educational, media, and communications fields.

REFERENCES

1. Y. Wu et al., "VideoDubber: Machine Translation with Speech-Aware Length Control for Video Dubbing," arXiv, 2022. <https://arxiv.org/abs/2211.16934>
2. D. Bigioi et al., "Multilingual Video Dubbing – A Technology Review," Frontiers in Signal Processing, 2023A <https://www.frontiersin.org/articles/10.3389/frsip.2023.1230755>
3. A. S. Subramanian et al., "Length-Aware Speech Translation

- for Video Dubbing,” Interspeech 2025.https://www.isca-archive.org/interspeech_2025/subramanian25_interspeech.pdf
4. J. Choi et al., “Dub-S2ST: Textless Speech-to-Speech Translation for Seamless Dubbing,” arXiv, 2025. <https://arxiv.org/abs/2505.20899>
 5. C.Cui et al., “Fine-grained Video Dubbing Duration Alignment,” arXiv, 2025. <https://arxiv.org/abs/2508.08550>
 6. K. Wang et al., “SyncVoice: Vision-Augmented TTS for Video Dubbing,” arXiv, 2025. <https://arxiv.org/abs/2512.05126>
 7. H.-S. Won et al., “End-to-End Multilingual Automatic Dubbing via Duration-based Translation with LLMs,” EMNLP, 2025. <https://aclanthology.org/2025.emnlp-demos.37>
 8. R. Kannoja, “Gen AI Driven Multilingual Audio Dubbing,” ScienceDirect, 2025. <https://www.sciencedirect.com/science/article/pii/S2590123025023138>