Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 14 Issue 03s, 2025

PolySub : AI-Powered Multilingual Subtitle and Dubbing with GenAI

[1]Varad Joshi, [2]Astha Asati, [3]Taufique Sana, [4]Siddharth Gajbhiye, [5]Dr. Hansaraj Wankhede

[1,2,3,4,5] Department of Artificial Intelligence G.H.Raisoni College of Engineering, Nagpur

Email: [1]varad.joshi.aiml@ghrce.raisoni.net, [2]astha.asati.aiml@ghrce.raisoni.net, [3]taufique.sanaullah.aiml@ghrce.raisoni.net, [4]siddharth.gajbhiye.aiml@ghrce.raisoni.net, [5]hansaraj.wankhede@raisoni.net

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Multilingual Subtitles, AI-based Dubbing, Speech Recognition, Text-to-Speech, Video Accessibility

Abstract

Language barriers go beyond communication gaps; they can restrict access to knowledge, limit collaboration, and reduce the global reach of digital content. PolySub addresses this challenge by providing an AI-powered platform for multi- lingual subtitling and dubbing, enabling seamless cross-language video accessibility. The system integrates OpenAI Whisper for transcription, Meta NLLB for translation and Meta MMS (TTS) model for audio generation, producing SRT files, dubbed videos with subtitles and original videos with embedded subtitles as outputs. Designed for scale, PolySub supports more than 150 languages, ensuring accessibility for diverse audiences. Evalua- tion shows strong performance, achieving a BLEU score of 37.0 and a BERTScore of 85.8, reflecting accurate, fluent, and se- mantically consistent output. This paper presents the conceptual framework of PolySub, outlines its architecture, and highlights how automated multilingual pipelines can enhance accessibility, scalability, and global communication across education, media, and professional domains.

Introduction

Automatic subtitle generation has become an essential tool nowadays for making digital content available to global audiences. With the rapid growth of video-based platforms in education, entertainment, and social media, subtitles have become an important aspect of videos. Subtitles improve comprehension, support multilingual viewers, and assist individuals with hearing impairments. Creating accurate and context-aware subtitles, however, remains challenging. AI-based systems often struggle with background noise, overlapping speech, varied accents, and cultural nuances, leading to errors in transcription, timing, and translation that can reduce the overall viewing experience.

Advances in speech recognition, natural language process- ing, and machine translation have enabled fully AI-driven systems that automatically generate subtitles and translations. These systems process audio and video to produce accurate subtitle drafts, handle time stamps, and ensure linguistic and context aware subtitles across multiple languages. Such AI- powered frameworks have proven effective in both educational and entertainment contexts, delivering fast, scalable, and cost- efficient solutions for global content accessibility.

PolySub extends this approach by integrating automated speech recognition, translation models, and dubbing model. It extracts audio, generates synchronized subtitle drafts, translates content into over 150 languages, and produces high-quality AI-generated voiceovers, ensuring accuracy, proper timing, and cultural relevance.

To enhance usability, PolySub provides three

outputs: stan- dard SRT subtitle files, burned-in captions, and multilingual AI-generated audio. A user-friendly interface allows creators to access and download outputs easily. This paper presents PolySub's conceptual framework, reviews existing AI subti- tling and dubbing solutions, and demonstrates how a fully AI-driven approach can deliver high-quality, accurate, and accessible multimedia experiences for global audiences.

Related Work

Recent studies show how artificial intelligence (AI) is changing the way subtitles and dubbing are created for multilingual video content. Striuk and Hordiienko [1] developed an AI-powered subtitle management system that uses deep ASR and NMT models to create synchronized SRT files and burned-in captions. With timing and length constraints built in, and a web interface for editing, their system reduced human subtitling effort for educational videos by 65%. Similarly, Penyameen et al. [2] introduced an end-to-end system that combines multilingual ASR with NMT to produce accurate transcripts and translations (WER 6.5%, BLEU 30), generating SRT files and captions across a variety of videos.

Hybrid systems have also been explored. Kuroiwa et al. [3] built a system that lets AI handle the first draft of subtitles while human editors refine difficult audio, timing, and cultural context. Their experiments on anime videos showed how this balance keeps costs down while improving quality. Favour et al. [4] worked on subtitles in noisy environments, using audio preprocessing and transformer-based ASR trained on noise- augmented data. Their results showed clear improvements in accuracy compared to traditional ASR. Al Sawi and Allam

[5] compared human and AI subtitles in Arabic films and found that while AI handled basic translations, human editors were better at adapting cultural references and avoiding errors, showing why human review is still important.

AI systems for dubbing are also advancing. Suresh et al.

[6] designed a pipeline using Google Speech-to-Text, Google Translate, and gTTS to produce both subtitles and dubbed videos. It reduced manual editing by 70% and achieved strong accuracy (5.8% WER) and user satisfaction (MOS 4.2). Adhikary et al. [7] proposed TRAVID, which combines transcription, translation, and Wav2Lip lip-sync to improve multilingual dubbing in educational videos. Devi et al. [8] developed a Whisper–MarianMT–Tacotron 2 pipeline with Wav2Lip that supports five Indian languages and cut dubbing time by 70%. Broader systems include Kannojia et al.'s [9] generative AI dubbing framework, which works with over 20 languages, and Kim et al.'s [10] VoiceCraft-Dub, which produces expressive, lip-synced speech while preserving the speaker's identity and emotion.

Research has also begun to address long-form video captioning. Wei et al. [11] introduced LongCaption-Agent, a three-stage framework supported by two new datasets (LongCaption-10K and LongCaption-Bench). Their fine-tuned multimodal model generated video captions of more than 1,000 words, matching the quality of much larger systems. Martin et al. [12] developed a subtitle management system for government videos using ASR and NMT. Their system produced industry-standard SRT/WebVTT files and reduced manual work, though they noted the need for wider language support and domain-specific tuning. Together, these studies show clear progress in AI-driven subtitling and dubbing. Automation reduces manual effort and improves speed, but challenges remain in cultural adaptation, noisy audio, and spe- cialized domains. The most effective solutions often combine AI efficiency with human expertise, ensuring both accuracy and quality across education, entertainment, and public con- tent.

System Overview

PolySub is an AI-powered web platform designed to automatically generate multilingual subtitles and dubbing for video content, supporting over 150 languages. It provides a fully automated workflow that processes video input, extracts audio, transcribes speech, translates it into the desired language, and produces synchronized subtitles and dubbed audio without requiring manual intervention.

It works by accepting a video input from the user. The audio is extracted and processed through an automatic speech recognition model to generate a transcript. The system then leverages advanced machine translation models to convert the transcript into the target language while maintaining context and accuracy. Using text-to-speech models, the translated text is converted into natural-sounding speech for dubbing. Finally, the generated subtitles and audio are synchronized with the original video and combined to produce the final output as shown in figure 1.

By integrating AI models for speech recognition, transla- tion, and voice dubbing, PolySub delivers an efficient, scalable, and accessible solution for global audiences. Its multilingual support, automation, and video processing ensure high-quality content localization, enabling content creators to reach wider audiences with minimal effort.

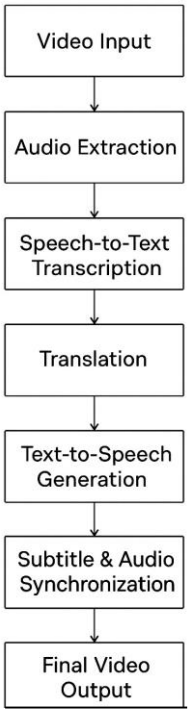

Fig. 1: Workflow of PolySub

Use Cases

PolySub is designed to make multimedia content more accessible and globally understandable by providing AI- powered subtitle generation and dubbing in over 150 languages. The platform automatically extracts audio from videos, generates accurate subtitles, and translates them while maintaining context and meaning. Using advanced AI models, PolySub converts translated subtitles into natural-sounding speech for dubbing, allowing creators to reach international audiences without language barriers.

Content creators, educators, and media platforms can use PolySub to expand their reach, improve engagement, and make content inclusive for viewers with different linguistic backgrounds. By supporting multiple subtitle formats such as .srt and .ass, the platform ensures compatibility with various video players and editing tools. PolySub operates fully automatically, reducing the need for manual intervention, while maintaining high-quality outputs through AI-based models for speech recognition, translation, and text-to-speech synthesis. This enables efficient processing of large volumes of video content.

Additionally, PolySub allows users to preserve original video quality while integrating multilingual subtitles and dub- bing, making it suitable for educational videos, entertainment content, corporate training, and online courses. Together, these features create a seamless and scalable solution for delivering globally accessible multimedia content.

Methodology

PolySub is an AI-driven platform designed to automate subtitle generation and dubbing for multimedia content in over 150 languages. The methodology outlines the workflow from video upload to final multilingual output, describing the role of each component in the system. The process begins when a user uploads a video through the Next.js frontend where user can select srt files, dubbed video and subbed video any of the 3 options. Then, the system first extracts the audio using FFmpeg and Pydub, ensuring that the audio is properly formatted for processing. The extracted audio is then passed to OpenAI Whisper, a robust speech recognition model, which transcribes spoken words into a high-quality text transcript. Whisper handles multiple accents, varying speech speeds, and noisy environments, making transcription accurate and reliable.

Once the transcript is generated, Meta NLLB (No Language Left Behind) is used to translate the text into the target languages. This model supports over 150 languages, enabling global accessibility. After translation, the system formats subtitles into standard formats such as SRT and ASS and aligns them precisely with the original audio. Timing adjustments and text chunking are handled automatically to ensure readability and synchronization. For dubbed audio, the translated text is processed through Meta TTS MMS (Multimodal Speech) to generate natural-sounding speech in the selected language. The TTS output is carefully timed to match the original video, preserving speaker pacing, tone, and clarity. Once both subtitles and dubbed audio are ready, FFmpeg integrates them back into the video file, producing a final output that is ready for streaming or download.

The final output consists of 3 parts: 1. SRT Files 2. Dubbed Videos with subtitles and 3. Original video with embedded subtitles perfectly demonstrated by figure 2. This methodol- ogy ensures scalability, consistency, and accessibility, making multimedia content understandable in an efficient and reliable manner.

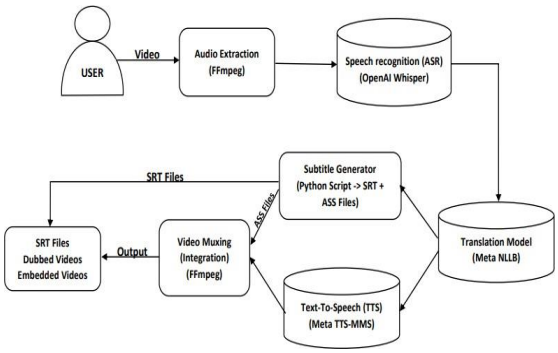

Fig. 2: Architecture of PolySub

A. User Interface

The PolySub frontend provides a clean, intuitive, and accessible interface for users to process videos in multiple languages. Users start by uploading a video through the interface, which serves as the input for the automated transcription, translation, and dubbing pipeline. After uploading, the platform allows users to select one or more languages from a collection of over 150 supported languages, making content localization highly versatile.

Once the languages are selected, users are presented with three output options:

1. Subtitle files in standard formats (SRT)
2. Dubbed videos with subtitles
3. Original videos with embedded subtitles

The output screen displays all generated files correspond- ing to the chosen languages and formats, allowing users to download them directly. Figure 3 shows the user interface and figure 4 presents downloadable output files.

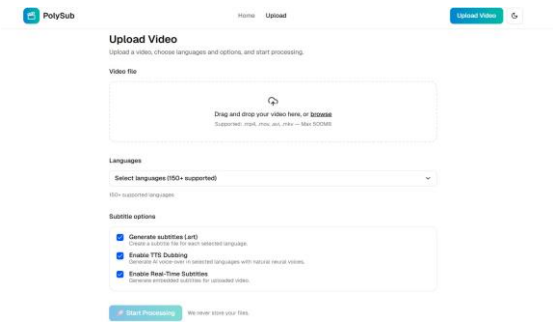

Fig. 3: User Input Interface

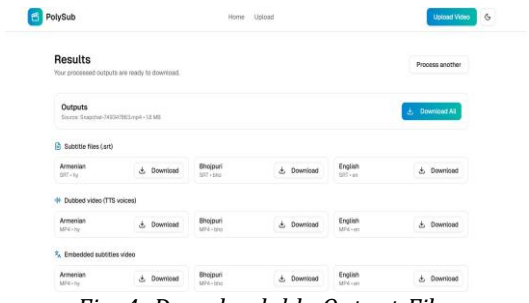

Fig. 4: Downloadable Output Files

Result

PolySub demonstrates high accuracy and quality in automated subtitle generation and dubbing. Using Translations generated with Meta NLLB maintain strong contextual integrity, achieving a BLEU score of 37.0 and a METEOR score of 64.5, reflecting reliable alignment between the translated and reference texts. The BERTScore of 85.8 highlights the semantic accuracy of translations, showing that the generated text effectively preserves meaning. Additionally, the ChrF++ score of 65.3 confirms that the translations maintain solid fluency and character-level consistency across languages. Subtitles are automatically formatted and synchronized to match the video, improving readability and viewer comprehension. Dubbing with Meta MMS produces speech that preserves timing and intonation, allowing viewers to engage with content seamlessly in their preferred language.

Overall, PolySub delivers a fully automated workflow that maintains high transcription and translation quality, reduces manual effort, and ensures accessibility for a global audience. The platform's results indicate that videos processed through PolySub are highly accurate, linguistically reliable, and suit- able for understanding.

Conclusion

PolySub provides an AI-powered solution to make multi- media content accessible across languages and regions. By integrating automated speech recognition, translation, and text- to-speech synthesis, the platform delivers subtitles and dubbed audio in over 150 languages, creating a seamless and inclusive viewing experience. PolySub addresses key challenges in content localization, accessibility, and audience engagement by providing a fully automated, accurate, and scalable workflow. Beyond its technical design, the platform emphasizes global inclusivity, enabling creators, educators, and media platforms to reach diverse audiences effortlessly. As digital media con- tinues to expand worldwide, PolySub highlights the potential of AI to bridge language barriers and make content universally understandable, enhancing both reach and engagement.

Future Scope

Future development of PolySub could focus on enhancing automation, accuracy, and the overall viewing experience. Expanding support for additional regional and low-resource languages would increase accessibility for a truly global audience. Advanced AI improvements, such as

context-aware translation, adaptive speech recognition, and lip-syncing for dubbed videos, could make the audio feel more natural and closely aligned with the original speaker.
Additional enhancements might include customizable subti- tle styling, interactive transcripts, and real-time video stream- ing support to make the platform more versatile and user- friendly. Analytics dashboards could provide creators with insights into viewer engagement, language preferences, and accessibility impact. Collaborative tools, such as community- driven translation, AI-assisted quality checks, and shared editing workflows, could further streamline the localization process. With these advancements, PolySub could become a fully automated platform for high-quality, immersive, and globally accessible multimedia content, enabling creators and educators to deliver accurate and engaging videos to diverse audiences.

References

A. M. Striuk and V. V. Hordiienko, "Research and development of a sub- title management system using artificial intelligence," CEUR Workshop Proceedings, 2025.

K. Penyameen, G. S. S. Rajan, A. A. Ahamed, S. Y. Ram, J. J. Shiny, and A. P. Nayaki, "AI-Based Automated Subtitle Generation System for Multilingual Video Transcription and Embedding," in 2025 3rd Interna- tional Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Feb. 2025, pp. 1096–1101, IEEE.

S. Kuroiwa, C. Oshima, and T. Koita, "Exploring a Hybrid System Com- bining AI and Human Intervention for Subtitle Creation in Entertainment Content," in Proc. World Multi-Conference on Systemics, Cybernetics and Informatics, Sep. 2024, pp. 72–73.

KA. Favour, M. S. Alzaidi, V. Ringu, M. Eltahir, F. Jungmann, and

T. Jorg, "Automatic Subtitle Generation in Noisy Environments Using Robust Speech Recognition Techniques," 2025.

I. Al Sawi and R. Allam, "Exploring challenges in audiovisual transla- tion: A comparative analysis of human-and AI-generated Arabic subtitles in Birdman," PLoS ONE, vol. 19, no. 10, p. e0311020, 2024.

P. Suresh, R. S. Prasad, P. K. Babu, N. Akhil, and B. Avinash, "Multi- lingual Video Content Transformation: Automated Subtitle Translation And Voice Integration," 2025.

P. K. Adhikary, B. Sugandhi, S. Ghimire, S. Pal, and P. Pakray, "Travid: An end-to-end video translation framework," arXiv preprint arXiv:2309.11338, 2023.

S. Devi, R. Sharma, and V. Patel, "Developing a software for dubbing of videos from English to other Indian regional languages," IJARCCE, vol. 14, no. 3, p. 14368, 2025.

R. Kannojia, A. K. Singh, I. Sharma, and S. Gupta, "Gen AI Driven Multilingual Audio Dubbing and Synthesis System for Cross-Language Video Platforms," Results in Engineering, p. 106241, 2025.

S. Kim, J. Choi, P. Peng, J. S. Chung, T. H. Oh, and D. Harwath, "VoiceCraft-Dub: Automated Video Dubbing with Neural Codec Lan- guage Models," arXiv preprint arXiv:2504.02386, 2025.

H. Wei, Z. Tan, Y. Hu, C. W. Chen, and Z. Chen, "LongCaptioning: Unlocking the Power of Long Video Caption Generation in Large Multimodal Models," arXiv preprint arXiv:2502.15393, 2025.

M. S. Mart´ın, J. Heras, and G. Mata, "Automatic Generation of Subtitles for Videos of the Government of La Rioja," in International Conference on Optimization and Learning, Cham: Springer Nature Switzerland, May 2023, pp. 393–402.