International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 625-629, May-2026

https://doi.org/10.47001/IRJIET/2026.105084

Automatic Video Subtitle Generation System through AI

[1]M.Mamatha, [2]Y Pavan Narashimha Rao, [3]Ch.Manoj Babu, [4]E.Vikram

[1]Assistant Professor, Dept. of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

[2,3,4]Dept. of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

***Abstract* - The rapid growth of digital video content on online platforms has created a significant demand for automated subtitle generation systems to improve accessibility, content understanding, and multilingual communication. Subtitles play an important role in helping hearing-impaired individuals, non-native speakers, and viewers in noisy environments understand video content effectively. However, manual subtitle creation is time-consuming, labor-intensive, and costly, especially for multilingual videos. To address this problem, this project presents an AI-powered multilingual subtitle generation system that automatically generates and embeds subtitles for video content in Telugu and English. The proposed system uses speech recognition and natural language processing technologies to convert spoken audio from video into text subtitles. The system integrates the speech recognition model OpenAI Whisper for automatic speech transcription and translation, and the multimedia processing tool FFmpeg for audio extraction and subtitle embedding. The entire system is deployed as a web application using Flask, allowing users to upload videos and download subtitled videos through a browser interface. The system workflow begins with video upload through the web interface. The uploaded video is processed using FFmpeg to extract the audio stream in a 16 kHz mono PCM format suitable for speech recognition processing. The extracted audio is then converted into log-mel spectrogram features and processed using a transformer-based encoder-decoder architecture of the Whisper model. The model generates time-stamped text segments directly from speech audio, eliminating the need for separate alignment tools.**

Keywords: Automatic Speech Recognition, Transformer Encoder-Decoder, ML model, Log-Mel Spectrogram, Subtitle Generation, FFmpeg, SubRip Text, Multilingual Video Processing, Web Application.

I. INTRODUCTION

In recent years, the use of digital video content has increased significantly due to the growth of online education platforms, social media, video streaming services, and digital communication systems. Videos are widely used for education, entertainment, training, business communication, and information sharing. However, many videos do not include subtitles, which makes it difficult for hearing-impaired individuals, non-native speakers, and viewers in noisy environments to understand the content. Subtitles help improve accessibility, comprehension, and user engagement. Creating subtitles manually is a time-consuming and labor-intensive process, especially for multilingual videos, as it requires transcription, timestamp synchronization, translation, and subtitle embedding.

To overcome these challenges, automated subtitle generation systems using Artificial Intelligence and Automatic Speech Recognition (ASR) have become increasingly important. Modern speech recognition systems use deep learning and transformer-based architectures to convert spoken language into text with high accuracy. In this project, an AI-powered multilingual subtitle generation system is developed to automatically generate subtitles for video content in Telugu and English. The system uses OpenAI Whisper for speech recognition and translation, which supports multilingual transcription and speech-to-text conversion.

The system extracts audio from video using FFmpeg, processes the audio through the speech recognition model to generate time-stamped text, and converts the generated text into subtitle format such as SRT files. These subtitles are then embedded back into the original video. The entire system is implemented as a web-based application using Flask, allowing users to upload videos and download subtitled videos through a browser interface without installing additional software.

The main objective of this project is to develop an automated system that reduces manual effort in subtitle creation, improves accessibility for multilingual video content, and provides a simple and efficient subtitle generation platform. The system integrates Artificial Intelligence, Speech Processing, Video Processing, and Web Development technologies to create a complete automated subtitle generation pipeline. This project demonstrates the practical application of AI in multimedia processing and digital

© 2026 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 625-629, May-2026
https://doi.org/10.47001/IRJIET/2026.105084

accessibility, particularly for Telugu and English video content.

System Architecture: Automatic Video Subtitle Generation System

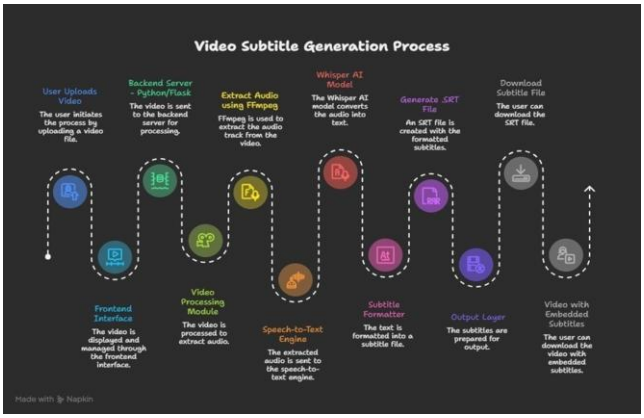


Figure 1: System architecture of the proposed AI-based multilingual subtitle generation framework showing the five-layer end-to-end pipeline from video upload and audio processing to speech recognition, subtitle generation, and web application deployment

II. LITERATURE REVIEW

Automatic subtitle generation has become an important research area due to the rapid growth of multimedia content on digital platforms such as online education systems, streaming services, and social media. Subtitles improve accessibility for hearing-impaired users, help non-native speakers understand video content, and improve video indexing and search. Earlier subtitle generation methods were manual, which required transcription, timestamp synchronization, translation, and subtitle embedding, making the process time-consuming and labor-intensive. To overcome these limitations, researchers developed Automatic Speech Recognition (ASR) based subtitle generation systems.

Early automatic subtitle generation systems were based on traditional speech recognition techniques such as Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). These systems converted speech into text but required separate modules for timestamp alignment and subtitle synchronization. Although these systems reduced manual work, they had limitations in accuracy, especially in noisy environments and multilingual speech recognition.

With the advancement of Machine Learning and Deep Learning, modern subtitle generation systems started using neural network-based speech recognition models. Deep learning models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and transformer-based architectures significantly improved speech recognition accuracy. These models can process large amounts of audio data and learn complex speech patterns, accents, and language structures, making automatic subtitle generation more accurate and efficient.

Recent research focuses on end-to-end speech recognition models that directly convert speech into text with timestamps. One of the most advanced models used for speech recognition and subtitle generation is OpenAI Whisper, which is trained on large multilingual audio datasets and supports multiple languages including Telugu and English. Whisper can perform speech recognition, language detection, and speech translation, making it suitable for multilingual subtitle generation systems.

Many subtitle generation systems also use multimedia processing tools such as FFmpeg for audio extraction from video and subtitle embedding into video files. FFmpeg is widely used in research and industry for video and audio processing because it supports multiple formats and efficient media processing.

Recent trends in subtitle generation research include real-time subtitle generation, multilingual subtitle translation, cloud-based subtitle services, and AI-based subtitle editing systems. These advancements show that automatic subtitle generation systems are becoming more accurate, faster, and more accessible.

Overall, the literature review shows that automatic subtitle generation systems have evolved from traditional speech recognition systems to modern AI-based deep learning systems. The integration of speech recognition models, multimedia processing tools, and web applications has made it possible to develop fully automated subtitle generation systems. The proposed project builds on these technologies to develop an AI-powered multilingual subtitle generation system for Telugu and English video content, providing an automated and accessible solution for subtitle generation.

III. DATASET AND METHODOLOGY

3.1 Dataset Description

The dataset used in this project is based on multilingual speech data used to train the speech recognition model OpenAI Whisper. The Whisper model is trained on approximately 680,000 hours of multilingual audio data collected from various sources such as audiobooks, podcasts, interviews, lectures, and online videos. The dataset contains audio recordings along with corresponding text transcripts in multiple languages including English and Telugu, which helps the model perform speech recognition and translation tasks.

In this project, the model is not trained from scratch; instead, the pre-trained Whisper model is used for subtitle

© 2026 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 625-629, May-2026

https://doi.org/10.47001/IRJIET/2026.105084

generation. For testing the system, sample video files such as lectures, interviews, and educational videos in Telugu and English are used as the testing dataset. The video files are processed using FFmpeg to extract audio, which is then converted into 16 kHz mono audio format and given to the model for transcription and subtitle generation.

The dataset therefore consists of multilingual speech audio used for model training (Whisper dataset) and video files used for testing subtitle generation. This dataset helps the system generate accurate subtitles for Telugu and English video content.

3.2 Data Preprocessing

Data preprocessing is an important step in the subtitle generation system because the speech recognition model requires audio input in a specific format. In this project, the uploaded video file is first processed using FFmpeg to extract the audio stream from the video file. The extracted audio is then converted into a 16 kHz mono PCM audio format, which is required for processing by the speech recognition model OpenAI Whisper. Standardizing the audio format ensures compatibility with the model and improves speech recognition accuracy.

After audio extraction and conversion, the audio signal is segmented into smaller time intervals so that long audio files can be processed efficiently. The segmented audio is then converted into log-mel spectrogram features, which represent audio signals in the frequency domain and are commonly used in speech recognition systems. The preprocessing stage may also include audio normalization and basic noise handling to improve audio quality. These preprocessed audio features are then provided as input to the speech recognition model for transcription and subtitle generation. Proper data preprocessing improves transcription accuracy and overall subtitle generation performance.

3.3 Feature Engineering

A critical contribution of this work is the design of engineered audio features that effectively represent speech characteristics for accurate subtitle generation. These features are derived from raw audio signals and processed for input into the speech recognition model.

- **log_mel_spectrogram:** Core feature representing audio in time–frequency domain, obtained by converting waveform into mel-scaled spectrogram and applying logarithmic scaling, capturing speech intensity and frequency variations.
- **frame_segmentation:** Audio is divided into short overlapping frames (typically milliseconds) to preserve temporal speech patterns and enable fine-grained analysis of speech signals.
- **fft_features:** Fast Fourier Transform (FFT) is applied to each audio frame to convert time-domain signals into frequency-domain components, enabling extraction of spectral information.
- **acoustic_embeddings:** High-level feature representations generated internally by the encoder of OpenAI Whisper, capturing phonetic, linguistic, and contextual speech information for accurate transcription.

3.4 Control Flowchart

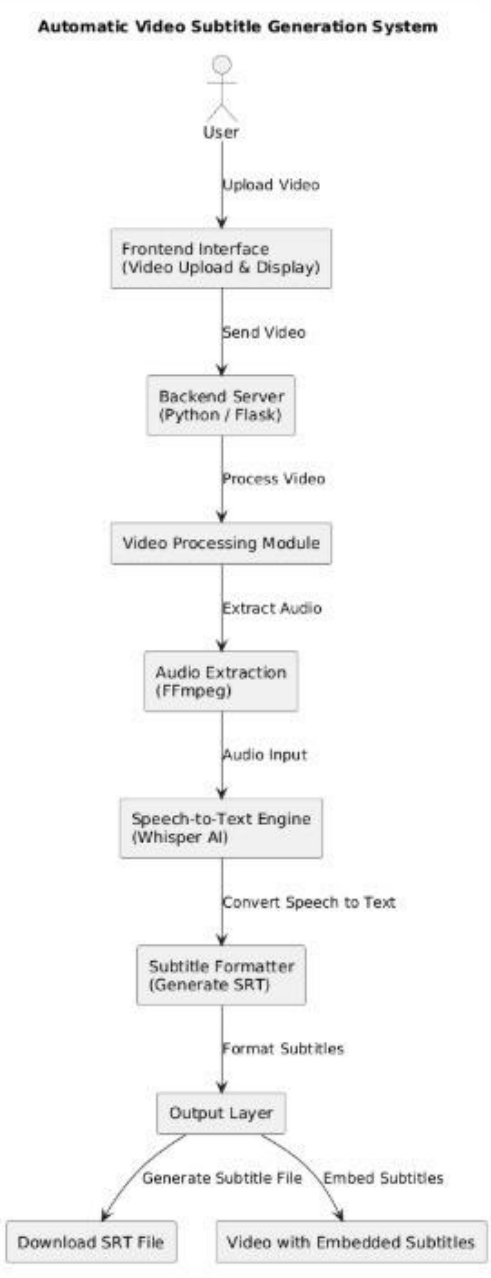


Figure 2: Flowchart of the proposed AI-based subtitle generation system illustrating the complete workflow from user video upload, backend processing, audio extraction, speech-to-text conversion, subtitle formatting, and output generation

© 2026 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 625-629, May-2026

https://doi.org/10.47001/IRJIET/2026.105084

IV. PROPOSED SYSTEM ARCHITECTURE

4.1 Video Input and Processing Module

The system begins with the video input module, where users upload video files through a web interface. The uploaded video is processed by the backend server, which manages the workflow and sends the video to the video processing module. The video processing module prepares the video for audio extraction and ensures compatibility with the audio processing system. This module handles video formats such as MP4, AVI, and MKV and prepares the video for further processing.

4.2 Audio Extraction and Preprocessing

In this stage, audio is extracted from the video using FFmpeg, which separates the audio track from the video file and converts it into a standard format such as 16 kHz mono PCM audio. The extracted audio is then preprocessed to improve speech recognition accuracy. Audio preprocessing includes noise reduction, normalization, segmentation into smaller audio frames, and conversion into log-mel spectrogram features. These features represent speech signals in the frequency domain and are used as input for the speech recognition model.

4.3 Speech Recognition and Transcription

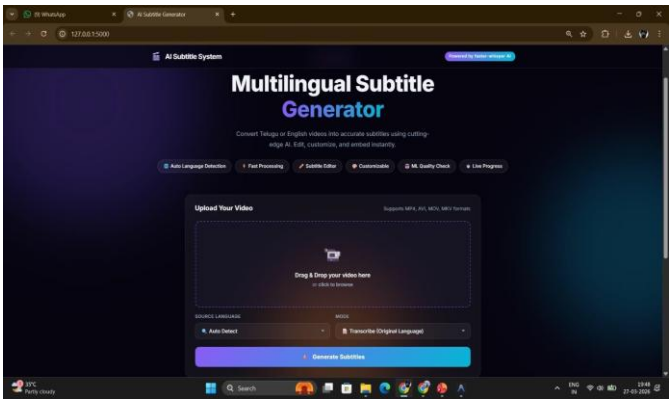
The speech recognition module uses OpenAI Whisper to convert speech audio into text. The Whisper model processes audio features using a transformer-based encoder-decoder architecture and generates text along with timestamps. The model supports multilingual transcription and translation, allowing the system to generate subtitles in Telugu and English. The output of this module is time-stamped text segments that are used for subtitle generation.

4.4 Subtitle Generation and Formatting

The subtitle generation module converts the time-stamped text into subtitle format such as SRT (SubRip Subtitle) files. This module formats subtitle text, assigns timestamps, and organizes subtitles into proper subtitle structure. The generated subtitle file can be downloaded directly or used for embedding subtitles into the video. Subtitle formatting ensures proper synchronization between speech and subtitles.

4.5 Web Application Deployment and Output

The complete system is deployed as a web application using a backend server built with Flask. The web application allows users to upload videos, process subtitle generation, and download the subtitle file or the video with embedded subtitles. The output module provides two output options: downloading the subtitle file (SRT) or generating a video with embedded subtitles. The system provides an automated, user-friendly interface for subtitle generation and video processing.



V. RESULTS AND DISCUSSION

5.1 System Performance Evaluation

The system performance was evaluated based on transcription accuracy, subtitle synchronization, and subtitle quality. The system was tested on various Telugu and English videos such as lectures, interviews, and presentations. The speech recognition model OpenAI Whisper generated accurate subtitles for clear audio videos, and the subtitles were properly synchronized with the video speech.

5.2 Processing Time Analysis

The processing time depends on video duration and system hardware. Audio extraction using FFmpeg was fast, while speech recognition took most of the processing time. Longer videos required more time because audio was processed in segments.

5.3 Subtitle Generation Results

The system successfully generated subtitles in SRT format and also produced videos with embedded subtitles. The generated subtitles were readable, properly formatted, and synchronized with the video. The system supports both subtitle file download and subtitled video output.

5.4 Discussion

The results show that the system successfully automates subtitle generation and reduces manual effort. The system works well for Telugu and English videos, but transcription accuracy may decrease for noisy audio. Overall, the system provides an efficient and automated solution for subtitle generation.

© 2026 IRJIET All Rights Reserved

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 625-629, May-2026
https://doi.org/10.47001/IRJIET/2026.105084

VI. CONCLUSION

The AI-based multilingual subtitle generation system developed in this project provides an automated solution for generating subtitles from video content using speech recognition and video processing technologies. The system integrates video upload, audio extraction, speech-to-text conversion, subtitle generation, and subtitle embedding into a single automated workflow. The system uses OpenAI Whisper for speech recognition, FFmpeg for audio extraction and subtitle embedding, and a Flask web application for user interaction.

The system reduces manual effort in subtitle creation and improves accessibility for multilingual video content, especially for Telugu and English videos. The generated subtitles are synchronized and available in SRT format or embedded directly into the video. The results show that the system works effectively for clear audio videos and provides an efficient and automated subtitle generation solution. Future improvements may include real-time subtitle generation and support for additional languages.

REFERENCES

[1] Fastelli *et al.*, "Speech-to-Text Captioning and Subtitling in Schools," *Audio Research Journal*, 2025.
[2] J. Poncelet *et al.*, "Leveraging Broadcast Media Subtitle Transcripts for ASR and Subtitling," *arXiv*, 2025.
[3] K. Sindhu *et al.*, "AI Powered Real-Time Video Caption Recommendation System," *IJCRT*, 2025.
[4] N. Nguyen *et al.*, "Whisper Based Speech-to-Text Captioning Performance Study," 2024.
[5] R. Veroz-Gonzalez *et al.*, "Automatic Closed Captions in Academic Video Presentations," 2024.
[6] S. Anand *et al.*, "Real-Time Subtitle Generation for Live Videos Using AI and Machine Learning," 2023.
[7] S. Polepaka *et al.*, "Automated Caption Generation for Video Call with Language Translation," *E3S Web of Conferences*, 2023.
[8] S. Papi *et al.*, "Direct Speech Translation for Automatic Subtitling," *TACL*, 2023.
[9] S. Polepaka *et al.*, "Automated Caption Generation for Video Call," *E3S Conference Proceedings*, 2023.
[10] Y. Ming *et al.*, "Visuals to Text: A Comprehensive Review on Automatic Image Captioning," 2022.
[11] M. Amirian *et al.*, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review," 2020.
[12] K. R. Aiswarya, "Automatic Multiple Language Subtitle Generation for Videos," *IRJET*, 2020.
[13] P. Sharma *et al.*, "Automatic Generation of Subtitle in Videos," *IJCSE*, 2019.
[14] A.Hannun *et al.*, "Deep Learning Based Speech Recognition Caption Systems," *arXiv*, 2019.
[15] N. Radha and R. Pradeep, "Automated Subtitle Generation," *IJAERV*, 2015.

Citation of this Article:

M.Mamatha, Y Pavan Narashimha Rao, Ch.Manoj Babu, & E.Vikram. (2026). Automatic Video Subtitle Generation System through AI. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(5), 625-629. Article DOI https://doi.org/10.47001/IRJIET/2026.105084

****

© 2026 IRJIET All Rights Reserved