



Automated Video Transcription and Multilingual Subtitle Generator using AI

K.Pavani¹, S.Aswini²

**#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology,
Vijayawada.**

#2 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract: The rapid growth of digital video content has created a significant demand for automated systems capable of converting speech into understandable multilingual subtitles. Traditional subtitle creation methods are manual, time-consuming, and inefficient for large-scale video processing. To address this challenge, this project proposes an AI-based Automated Video Transcription and Multilingual Subtitle Generation System that converts video speech into text and generates synchronized subtitles in multiple languages automatically. The proposed system uses FFmpeg for audio extraction, the Whisper deep learning model for speech recognition, and Neural Machine Translation techniques for multilingual subtitle generation. The system supports multiple languages including English, Telugu, Hindi, and Tamil, improving accessibility for users from diverse linguistic backgrounds. Extracted speech is processed, translated, and synchronized with video playback in real time through an interactive graphical user interface developed using Python and Tkinter.

The system is designed to provide high transcription accuracy, efficient subtitle synchronization, and reduced manual effort compared to traditional subtitle generation methods. Experimental analysis demonstrates that the proposed model delivers effective speech recognition and multilingual translation performance even for long-duration video content. The project can be further extended for live video translation, mobile applications, and real-time streaming platforms, making it suitable for educational, entertainment, communication, and media applications

1. INTRODUCTION

In recent years, digital video content has increased rapidly across platforms such as online education, entertainment, social media, news broadcasting, and video conferencing systems. Understanding video content in different languages has become a major challenge for global audiences due to language barriers. Subtitles play an important role in improving accessibility and communication; however, traditional subtitle generation is mostly performed manually, which is time-consuming, labor-intensive, and costly for large-scale video processing.



Advancements in Artificial Intelligence (AI), Deep Learning, Speech Recognition, and Natural Language Processing (NLP) have enabled the development of intelligent systems capable of automatically converting speech into text and translating it into multiple languages. Modern Automatic Speech Recognition (ASR) models such as Whisper provide highly accurate transcription by processing multilingual audio data using transformer-based deep learning architectures. Similarly, Neural Machine Translation (NMT) systems help translate text into different languages while maintaining contextual meaning and sentence structure.

This project presents an AI-Based Automated Video Transcription and Multilingual Subtitle Generation System that automatically extracts speech from video files, converts audio into text, translates the generated text into multiple languages, and displays synchronized subtitles during video playback. The system uses FFmpeg for audio extraction, Whisper AI for speech recognition, and translation APIs for multilingual subtitle generation. The generated subtitles are synchronized with video timing to improve user experience and accessibility.

The proposed system supports multiple languages including English, Telugu, Hindi, and Tamil, making it useful for regional and global users. A user-friendly graphical interface is developed using Python and Tkinter to allow users to upload videos, process subtitles, and display translated output efficiently. The automation of subtitle generation significantly reduces manual effort and improves processing speed compared to traditional methods.

The project is highly beneficial in areas such as online learning platforms, media broadcasting, multilingual communication systems, video streaming services, and accessibility support for hearing-impaired users. By integrating speech recognition, machine translation, subtitle synchronization, and video processing into a single platform, the system provides an efficient and intelligent solution for multilingual video understanding.

2. LITERATURE SURVEY

[1] R Robust Speech Recognition via Large-Scale Weak Supervision

We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.



[2] Attention is all you need

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

[3] End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results

We replace the Hidden Markov Model (HMM) which is traditionally used in continuous speech recognition with a bi-directional recurrent neural network encoder coupled to a recurrent neural network decoder that directly emits a stream of phonemes. The alignment between the input and output sequences is established using an attention mechanism: the decoder emits each symbol based on a context created with a subset of input symbols elected by the attention mechanism. We report initial results demonstrating that this new approach achieves phoneme error rates that are comparable to the state-of-the-art HMM-based decoders, on the TIMIT dataset.



3. METHODOLOGY

i) Proposed Work:

The proposed system is an AI-based video speech translation system that automatically converts speech from video into text and generates subtitles in multiple languages. It supports local Indian languages such as Telugu, Hindi, and Tamil, along with English, making it useful for a wide range of users. The system processes video by extracting audio, performing speech recognition, translating the text, and displaying synchronized subtitles in real-time through a user-friendly interface.

ii) System Architecture:

The AI-Based Video Speech Translation and Subtitle Generation System is an intelligent multimedia processing framework that automatically converts spoken content in a video into multilingual subtitles. The process begins with an input video from which the audio is extracted using FFmpeg. The extracted audio is then processed by the Whisper Speech Recognition Model, which converts speech into accurate transcribed text. This transcribed text is passed to a Translation Module that utilizes the Google Translator API to translate the content into multiple languages such as English, Telugu, Hindi, and Tamil. The translated text, along with the original transcript, is then sent to a Subtitle Synchronization Module, where subtitle timings are aligned with the corresponding speech segments to ensure accurate display during video playback. Finally, the synchronized multilingual subtitles are integrated with the original video and presented through the Output Display Module, producing a video with translated subtitles and timestamped text. This system enhances accessibility, enables cross-language communication, reduces manual subtitle creation efforts, and is highly useful for educational platforms, online video streaming services, conferences, media broadcasting, and content localization applications.

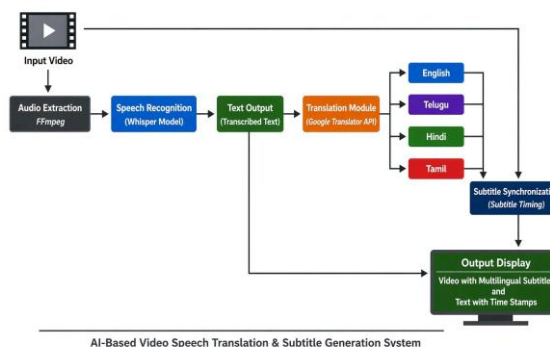


Fig1 Proposed system

iii) Modules:

1. Video Upload Module

This module allows users to upload video files into the system through a graphical user interface. It accepts different video formats and sends the selected video for further processing.

2. Audio Extraction Module

The uploaded video is processed using FFmpeg to extract audio from the video file. The extracted audio is converted into a suitable format for speech recognition processing.

3. Audio Preprocessing Module

In this module, the extracted audio is cleaned and normalized to improve speech recognition accuracy. Noise reduction and audio formatting operations are performed before transcription.

4. Speech Recognition Module

This module uses the Whisper deep learning model to convert speech from the audio into text. It processes audio segments and generates accurate transcriptions with timestamp information.

5. Text Processing Module

The generated text is cleaned and formatted by removing unnecessary characters, extra spaces, and



unwanted symbols. This improves readability and translation quality.

6. Translation Module

The processed text is translated into multiple languages such as Telugu, Hindi, Tamil, and English using Neural Machine Translation techniques and translation APIs.

7. Subtitle Generation Module

This module converts translated text into subtitle format with proper timing information. Subtitle files are generated for displaying multilingual subtitles during video playback.

8. Subtitle Synchronization Module

The synchronization module matches subtitle timestamps with video playback time to ensure subtitles appear accurately in real time.

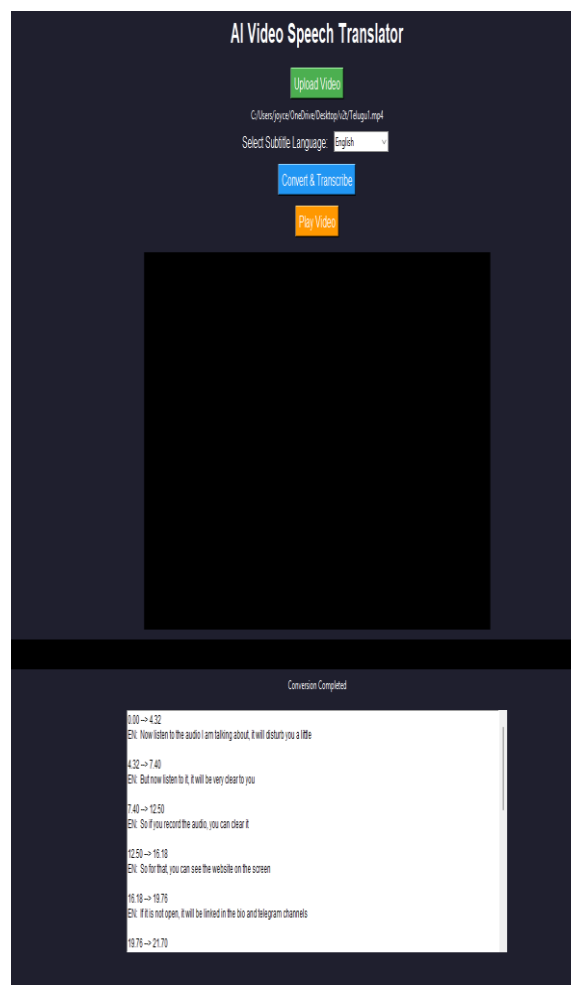
9. Video Playback Module

This module uses VLC integration to play videos along with synchronized multilingual subtitles. Users can watch videos while viewing translated subtitles.

10. Output Display Module

The final processed video with multilingual subtitles is displayed to the user. The system provides an easy-to-use interface for viewing and accessing generated subtitles.

4. EXPERIMENTAL RESULTS





AI Video Speech Translator

Upload Video

C:\Users\joyce\OneDrive\Desktop\vid1\Telugu1.mp4

Select Subtitle Language: English

Convert & Transcribe

Play Video



So if you record the audio, you can clear it

Conversion Completed

0:00 → 4:32

EN: Now listen to the audio I am talking about, it will disturb you a little

4:32 → 7:40

EN: But now listen to it, it will be very clear to you

7:40 → 12:50

EN: So if you record the audio, you can clear it

12:50 → 16:18

EN: So for that, you can see the website on the screen

16:18 → 19:76

EN: If it is not open, it will be linked in the bio and telegram channels

19:76 → 21:70

AI Video Speech Translator

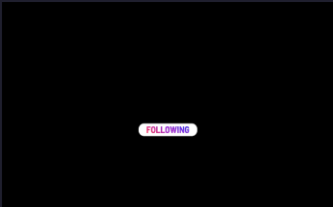
Upload Video

C:\Users\joyce\OneDrive\Desktop\vid1\Telugu1.mp4

Select Subtitle Language: English

Convert & Transcribe

Play Video



Like it if you like it, for interesting reasons, just follow our page

Conversion Completed

0:00 → 4:32

EN: Now listen to the audio I am talking about, it will disturb you a little

4:32 → 7:40

EN: But now listen to it, it will be very clear to you

7:40 → 12:50

EN: So if you record the audio, you can clear it

12:50 → 16:18

EN: So for that, you can see the website on the screen

16:18 → 19:76

EN: If it is not open, it will be linked in the bio and telegram channels

19:76 → 21:70

AI Video Speech Translator

Upload Video

C:\Users\joyce\OneDrive\Desktop\vid1\Telugu1.mp4

Select Subtitle Language: English

Convert & Transcribe

Play Video



इसलिए यदि आप ऑडियो रिकॉर्ड करते हैं, तो आप इसे साफ़ कर सकते हैं

Conversion Completed

0:00 → 4:32

EN: Now listen to the audio I am talking about, it will disturb you a little

4:32 → 7:40

EN: But now listen to it, it will be very clear to you

7:40 → 12:50

EN: So if you record the audio, you can clear it

12:50 → 16:18

EN: So for that, you can see the website on the screen

16:18 → 19:76

EN: If it is not open, it will be linked in the bio and telegram channels

19:76 → 21:70

www.ajaccm.com

893

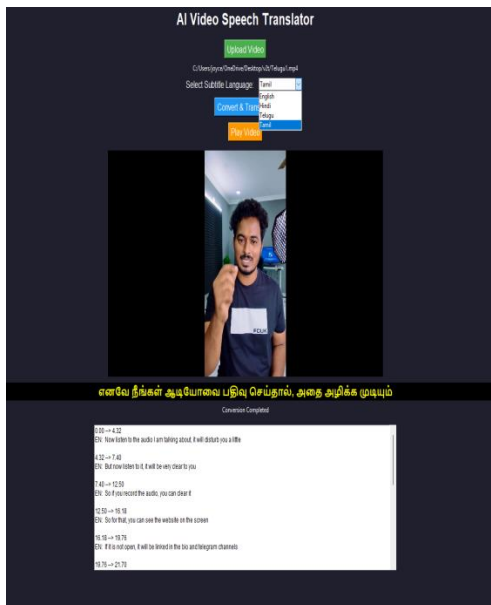
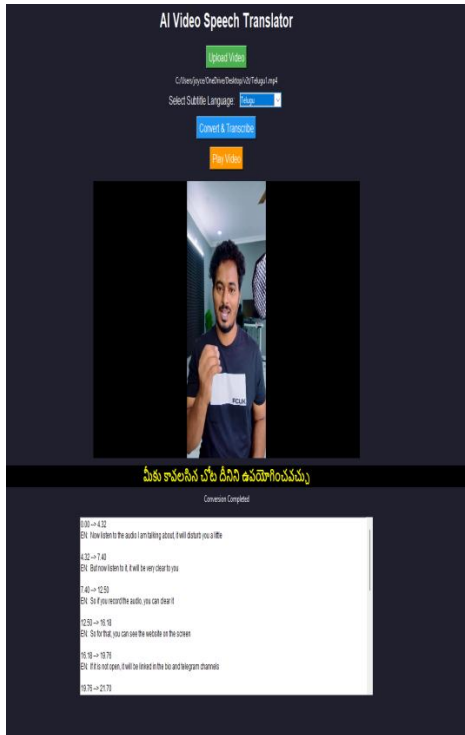


Fig 2:Results

5. CONCLUSION

The proposed AI-Based Automated Video Transcription and Multilingual Subtitle Generation System successfully converts video speech into accurate text and generates synchronized subtitles in multiple languages. The integration of FFmpeg, Whisper speech recognition, and Neural Machine Translation techniques enables efficient audio extraction, transcription, translation, and subtitle synchronization within a single platform. The system reduces manual effort, improves subtitle generation speed, and enhances accessibility for users from different linguistic backgrounds. Support for regional languages such as Telugu, Hindi, and Tamil further increases the usefulness of the system for educational, media, and communication applications. Experimental analysis shows that the proposed system provides reliable performance, better transcription accuracy, and effective multilingual subtitle generation, making it a practical and intelligent solution for modern video processing requirements

REFERENCES

- A. Radford, J. W. Kim, C. Hallacy, et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” OpenAI, 2022.
- S. B. Pratap, A. Tjandra, B. Shi, et al., “Scaling Speech Recognition Models to Large Datasets,” arXiv preprint arXiv:2005.08100, 2020.



A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is All You Need,” Advances in Neural Information Processing Systems (NeurIPS), 2017.

Google, “Google Cloud Translation API Documentation,” Available: <https://cloud.google.com/translate/docs>

FFmpeg Developers, “FFmpeg Multimedia Framework,” Available: <https://ffmpeg.org/>

Python Software Foundation, “Python Language Documentation,” Available: <https://www.python.org/doc/>

VideoLAN Organization, “VLC Media Player Documentation,” Available: <https://www.videolan.org/vlc/>

J. Chorowski, D. Bahdanau, K. Cho, et al., “End-to-End Continuous Speech Recognition Using Attention-Based Recurrent Neural Networks,” ICASSP, 2015.

D. Bahdanau, K. Cho, Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” ICLR, 2015.

Author Profiles



Mrs.K.Pavani Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with MCA and M. Tech in Computer Science .She has 10 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada,NTR District. Her area of interest includes AI ML, etc.



Ms.S.Aswini is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in **B.S.C** from SVL KRANTHI DEGREE College Avanigadda Her area of interest are AI and DBMS