IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2025

DOI: 10.17148/IJARCCE.2025.14368

DEVELOPING A SOFTWARE FOR DUBBING OF VIDEOS FROM ENGLISH TO OTHER INDIAN REGIONAL LANGUAGES

Prof. S. S. Bhagat [1], Om Giratkar [2], Tejashree Suryawanshi [3], Shruti Raspayle[4], Vinaykumar Gupta[5]

Assistant Professor, Department of Computer Engineering, TSSM BSCOER, Pune, India[1]

Student, Department of Computer Engineering, TSSM BSCOER, Pune, India[2-5]

Abstract: Our initiative creates a dubbing system powered by AI to convert English audio into various Indian regional languages, thus improving accessibility. By utilizing natural language processing (NLP), speech-to-text (STT), and text-to-speech (TTS) technologies, the system captures spoken language, translates it into text, and produces audio that sounds natural when dubbed [1]. This automated solution streamlines the dubbing process, making it quicker, more affordable, and scalable, which is advantageous for sectors like education, entertainment, and business. It minimizes the need for manual dubbing while enhancing speech fluency and synchronization [2]. We tackle challenges such as timing adjustments and linguistic precision to ensure translations sound natural [3]. Future improvements could feature real-time dubbing capabilities and advanced voice cloning techniques to further enhance quality and maintain speaker consistency [4].

Keywords: AI- Powered Dubbing, Speech-to-Text, Text-to-Speech, Natural Language Processing (NLP), Multilingual accessibility

I. INTRODUCTION

In today's online environment, videos serve as a significant medium for information, entertainment, and education. However, the dominance of English content can pose difficulties for individuals who prefer to consume material in regional languages [5].

To address this issue, our initiative is centered around creating software that can automatically dub videos from English into Indian regional languages. This application leverages Natural Language Processing (NLP), Speech-to-Text (STT), and Text-to-Speech (TTS) technologies to offer an efficient and user-friendly dubbing solution [6].

The procedure begins when a user uploads a video file. The software first extracts the audio track from the video and converts the spoken English into written text through Speech-to-Text (STT) technology [7].

Following this, the text is processed with NLP to translate it into the chosen Indian language. Once the translation is complete, the system employs Text-to-Speech (TTS) technology to produce audio in the target language, striving for a natural sound [8].

At last, this newly generated audio is synchronized with the original video to align with the timing of the speech. Conventional dubbing is often costly and time-consuming, as it requires the involvement of translators, voice actors, and sound engineers [9].

Our AI-driven software accelerates this process, making it more affordable and scalable, thus enabling swift dubbing of large volumes of content. This approach benefits areas such as education, entertainment, governmental communication, and businesses, facilitating access to information in native languages [10].

One of the primary challenges is ensuring that the dubbed speech is both accurate and sounds natural. Given that Indian languages have unique sentence structures and tonal variations, our system is specifically designed to manage these differences adeptly. In summary, this project seeks to enhance the accessibility of video content by overcoming language obstacles.

DOI: 10.17148/IJARCCE.2025.14368

II. LITERATURE REVIEW

Himanshu Sehrawat et al. [1] present a comprehensive method for video translation into multiple languages using deep learning and audio synthesis techniques. CH. Vijaya Kumar et al. [2] propose a survey on sophisticated video dubbing software aimed at bridging linguistic gaps and promoting inclusivity by making video content accessible to a wider audience. Purushottam Sharma et al. [3] focus on translating speech to Indian Sign Language using Natural Language Processing.

Prof. R.K. Nale et al. [4] describe a machine translation system for English educational videos into Indian regional languages, improving accessibility for diverse linguistic audiences. Y V Nagesh Meesala et al. [5] discuss harnessing open innovation for translating global languages into Indian languages using machine translation, speech recognition, and synthesis. Dr. Ankit Sharma [6] explores language translation and subtitling strategies employed by OTT platforms in India, revealing that they use a combination of machine translation and human translators for high-quality content.

III. METHODOLOGY

The design of the research is a vital element of any investigation, as it establishes the structure for data collection, analysis, and interpretation [7]. In this investigation, the research design is meticulously developed to examine the complex ramifications of deepfake technology and its detection. By utilizing a blend of qualitative and quantitative methods, this research aspires to deliver a thorough understanding of the technology's effects, the efficacy of detection techniques, and the views of various stakeholders involved [8].

- Automated Video Dubbing System: The software automates the entire video dubbing process from English to Indian regional languages, eliminating the need for manual dubbing.[9]
- To create a high-quality and accurate dubbing experience by integrating Speech-to-Text (STT), Machine Translation (MT), and Text-to-Speech (TTS) technologies.
- Target Languages: Hindi, Marathi, Tamil, Telugu, Bengali, Kannada, etc.

3.1 Video Processing and Audio Extraction

The input video file is processed, and the audio is extracted using FFmpeg. The extracted audio is converted into a suitable format for further processing.

3.2 Speech Recognition

The extracted audio is transcribed into text using the Whisper model.This ensures high accuracy by effectively handling accents, speech variations, and noisy environments.

3.3 Text Translation

The transcribed text is translated into the desired Indian language using a transformer-based model.The NLTK tokenizer is applied to split text into sentences for precise translation.The translation model maps the source language to the target language using predefined language codes for better accuracy.

3.4 Text to Speech

The translated text is converted into natural-sounding speech using a TTS model. The generated voice is optimized for clarity, pronunciation, and fluency.

3.5 Audio Synchronization

The generated dubbed audio is adjusted to match the original speech timing, ensuring a smooth and natural listening experience.

IV. SYSTEM DESIGN

1.Input **-** The process begins with the input, which is a video file.
2.Processing - The system uploads the video file.
3.Audio Extraction - The audio track is separated from the video.The extracted audio file is stored in a format. FFmpeg is used for extracting the audio.
4.Speech Recognition - The system applies Automatic Speech Recognition (ASR) to convert speech into text.
5.Text Translation- The recognized text is translated into the target languages
6.Text to Speech Conversion- The translated text is synthesized into speech using Text-to-Speech (TTS)
7.Audio Synchronization- The synthesized audio is synchronized with the original video.

time-stretching methods will be applied to match the audio duration with the video.The new dubbed audio is combined with the original video.

8.Output- The final video, now containing the dubbed audio, is generated and stored. The user can download the dubbed video in their preferred format.

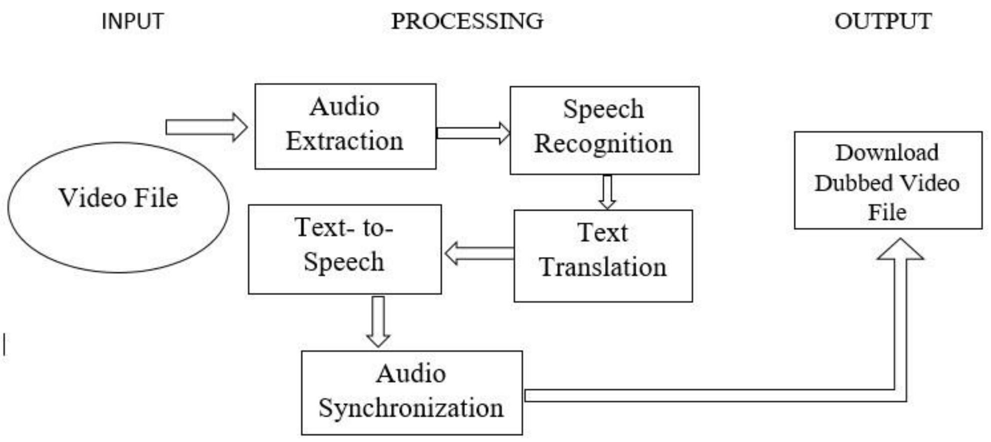

Fig. System Architecture

V. DESIGN AND IMPLEMENTATION

The design of the AI-Based Automated Video Dubbing System follows a modular approach, ensuring seamless integration of multiple technologies for efficient video dubbing. The system consists of three main stages: Input Processing, Language Processing, and Output Generation.

System Design
The system is structured into the following components:

4.1 Input Processing:
The system accepts a video file as input. Audio is extracted using FFmpeg and converted into a suitable format for further processing.

4.2 Language Processing:
Speech Recognition: The extracted audio is transcribed into text using Whisper, a deep learning-based model that ensures high accuracy even in noisy conditions.

4.3 Machine Translation:
The transcribed text is translated into the target Indian language using a transformer-based translation model.

4.3.1 Text Processing:
The translated text is tokenized using NLTK to enhance the accuracy of phrase-based translation.

4.3.2 Output Generation:
Text-to-Speech (TTS) Conversion: The translated text is synthesized into speech using a TTS model, generating a natural-sounding dubbed voice.

4.3.3 Audio Synchronization: The synthesized voice is aligned with the original video timing to ensure smooth synchronization.

4.4 Final Video Output: The dubbed audio is merged back into the video, producing the final dubbed video.

Implementation

Backend: Python (for processing audio, translation, and speech synthesis), Flask Framework

Libraries Used:

- FFmpeg – Audio extraction and processing
- Whisper – Speech recognition
- NLTK – Text tokenization
- Transformers – Machine translation
- GTTS or TTS – Text-to-speech synthesis

Database: MySQL

Frontend: HTML, CSS, JavaScript

VI. CONCLUSION

The AI-Based Automated Video Dubbing System successfully automates the process of translating and dubbing videos into multiple Indian languages, eliminating the need for manual dubbing. By integrating speech recognition, machine translation, and text-to-speech synthesis, the system ensures accurate and natural-sounding dubbed audio. The use of FFmpeg for audio extraction, Whisper for transcription, and transformer-based models for translation enhances the efficiency and precision of the system.

This system significantly improves accessibility by allowing content creators to reach a wider audience without language barriers. The automation of the dubbing process reduces time and costs compared to traditional methods while maintaining high-quality output.

Future enhancements can focus on improving voice modulation, speaker adaptation, and real-time synchronization to make the dubbed content even more seamless and natural. With continuous advancements in AI and deep learning, this system can be further optimized to support more languages and provide an enhanced user experience.

ACKNOWLEDGMENT

We sincerely thank our guide and technical advisors for their invaluable guidance and support in developing this AI-Based Automated Video Dubbing System. Your expertise and insights were crucial in refining the project and overcoming challenges. We also appreciate the collaborative efforts of our peers and the resources provided by our institution. This project would not have been possible without your collective contributions. Thank you all for your unwavering support.

FUTURE SCOPE

1. Support for More Languages - Expand the system to support additional Indian and international languages to make content accessible to a global audience.
2. Enhanced Speech Naturalization- Improve text-to-speech (TTS) models to generate more natural and expressive voices with better emotional tone and intonation.
3. Real-Time Dubbing- Implement real-time dubbing capabilities for live events, webinars, and news broadcasts.
4. Lip-Sync Enhancement- Integrate deep learning-based lip-syncing technology to improve visual synchronization between the dubbed audio and the speaker’s lip movements.
5. Adaptive Voice Cloning- Develop a voice cloning system to match the dubbed voice to the speaker's original tone and style for better user experience.
6. Cloud-Based Deployment- Enable cloud-based processing for scalable and faster dubbing, reducing dependency on local hardware.
7. User Customization- Allow users to select different voice tones, speech speeds, and accents based on their preferences.
8. Integration with Content Platforms- Collaborate with platforms like YouTube, Netflix, and online education portals to provide seamless dubbed content.
9. Improved Noise Reduction- Enhance background noise reduction in speech recognition to improve accuracy in noisy environments.

DOI: 10.17148/IJARCCE.2025.14368

REFERENCES

[1]. Afouras, T., Chung, J. S., and Zisserman, A. (2018). "LRS3-TED: A large-scale dataset for visual speech recognition.

[2]. Bigioi, D., and Corcoran, P. (2023). "Multilingual video dubbing—a technology review and current challenges. Frontiers in Signal Processing.

[3]. Chen, L., Liu, Z., and Wang, X. (2018). "GAN-based lip movement generation from driving speech and reference lip frames.

[4]. Dhariwal, P., and Nichol, A. (2021). "Improved denoising diffusion probabilistic models." Advances in Neural Information Processing Systems.

[5]. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., and Goldman, D. B. (2019). "Text-based editing of talking-head video." ACM Transactions on Graphics (TOG).

[6]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., and Ozair, S. (2014). "Generative adversarial nets." Advances in Neural Information Processing Systems.

[7]. Lu, Y., Chai, J., and Cao, X. (2021). "Live speech portraits: Real-time photorealistic talking-head animation." ACM Transactions on Graphics (TOG).

[8]. Mittal, G., and Wang, B. (2020). "Animating faces using disentangled audio representations.

[9]. Narvekar, N. D., and Karam, L. J. (2011). "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)." IEEE Transactions on Image Processing.

[10] Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). "Synthesizing Obama: Learning lip sync audio." ACM Transactions on Graphics (TOG).

[11]. Ren, Y., et al. (2019). "FastSpeech: Fast, Robust and Controllable Text to Speech." Neural Information Processing Systems (NeurIPS).

[12]. Chung, J. S., Jamaludin, A., & Zisserman, A. (2017). "You said that? Synthesizing talking faces from audio." IEEE Transactions on Pattern Analysis and Machine Intelligence.