International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Audio Extraction, Grammar Correction and Video Syncing Tool

B. Swetha[1], M. Nishanth kumar[2], K. Karthik Reddy[3]
[1]*Assistant Professor, Mahatma Gandhi Institute of Technology*
[2,3] *Student, Mahatma Gandhi Institute of Technology*

***Abstract*: Vid-Text is an innovative application designed to streamline video content processing by integrating multiple advanced functionalities into a single, user-friendly platform. This system focuses on extracting audio from uploaded video files, enhancing audio quality through noise removal, and converting the audio into text with high precision using state-of-the-art speech-to-text technologies. By ensuring grammatical accuracy and allowing manual review, Vid-Text provides users with an efficient way to refine and edit transcriptions to meet their specific needs. One of the standout features of Vid-Text is its multilingual support, enabling users to translate the corrected text into a variety of languages. This ensures that video content can reach a wider audience by breaking language barriers. Additionally, the translated text can be converted back into audio, offering an audio-based representation of the content in the chosen language.**

The system also includes robust synchronization capabilities, allowing the text, translated audio, and original video to align seamlessly. This enables users to generate a final video with embedded subtitles or synchronized audio that can be easily shared or downloaded. The platform is designed to cater to the needs of content creators, educators, and professionals who require efficient, accessible, and multilingual video processing tools. Vid-Text is not only a time-saving solution

I. INTRODUCTION

In the modern era of digital communication and media, video content has become a dominant form of information sharing across various industries, including education, entertainment, and professional sectors. With the rise of platforms that rely heavily on video-based communication, such as online learning systems, social media platforms, and content creation tools, the need for effective video processing solutions has significantly increased.

Despite the widespread adoption of video as a medium, challenges persist in making such content universally accessible and adaptable. Video content created in a specific language often limits its accessibility to a global audience.

Background noise or poor-quality audio can reduce the clarity and effectiveness of video messages. Tasks such as transcription, grammatical correction, and translation are often time-consuming and labor-intensive. Aligning audio, text, and video components for multilingual audiences poses technical complexities.

Traditional tools for video processing often address these challenges in isolation, requiring users to rely on multiple platforms for extracting audio, transcribing text, translating content, and generating subtitles or synchronized outputs. This fragmented approach is not only inefficient but also prone to inconsistencies in quality and compatibility.

The Vid-Text: Audio Extraction and Syncing System seeks to address these challenges through a comprehensive and integrated approach. By leveraging advanced technologies such as noise removal algorithms, speech-to-text conversion tools, machine translation systems, and text-to-speech frameworks, Vid-Text provides an all-in-one solution for video content processing.

A. Problem Statement.

The rapid growth of video content and its critical role in communication, education, and entertainment has brought about several challenges in video processing. Current systems for audio extraction, transcription, translation, and synchronization are fragmented, often requiring multiple tools and manual intervention, leading to inefficiencies and inconsistent results. Additionally, the issue of poor audio quality, such as background noise, further complicates transcription and translation accuracy. Furthermore, the lack of an integrated platform that handles all aspects of video content processing—from audio extraction to final synchronization—limits the scalability and usability of such systems.

This project seeks to address these challenges by developing an all-in-one solution that automates and streamlines the process of extracting audio from video, converting it to text, translating the text into multiple languages, and synchronizing the output. By integrating noise removal, transcription, grammatical correction, translation, and text-to-speech features, the Vid-Text system aims to provide a seamless and efficient workflow for content creators, educators, and professionals. This solution will not only improve accessibility and inclusivity but also enhance the quality and reach of video content, overcoming the limitations of existing tools and workflows.

B. Existing Systems.

Current systems for video processing typically involve multiple separate tools that handle different aspects of the workflow. These tools often require manual intervention and do not provide an integrated solution for all stages of video content processing. Existing video-to-text systems, for example, rely on speech recognition technology to transcribe audio into text. However, these systems are frequently prone to errors, especially in noisy environments, and require considerable manual editing for accuracy. Moreover, while transcription services have improved over the years, most platforms only provide limited language support, making them inadequate for global audiences. Many systems also fail to provide translation capabilities, or when they do, the translations may lack the precision necessary for professional use. As a result, users are often left with text that needs further editing or additional tools to make it more suitable for diverse audiences.

For audio extraction, many systems rely on standalone software that requires video to be processed separately before the audio can be extracted. This fragmented approach not only wastes time but also leads to challenges when synchronizing audio with the original video. Manual adjustments are often needed, which can be a tedious and error-prone process. In terms of synchronization, while there are tools available for creating subtitles or overlaying translated text, these often operate independently from the audio and video, making it difficult to achieve seamless integration. This lack of synchronization can lead to misalignment between the audio, text, and video, particularly when translating content into multiple languages.
Furthermore, many existing systems do not focus on noise removal and are unable to provide high-quality, accurate transcription when the audio quality is poor, which can negatively affect the overall output quality.

II. PROPOSED SYSTEM

A. Architecture of Proposed System.

The architecture of the Vid-Text system is designed to provide an efficient and seamless workflow for audio extraction, transcription, translation, and synchronization of video content. The system involves three primary actors:

User, Admin, and System. Users begin by registering and logging into the platform with valid credentials. After successful authentication, they upload video files, from which the system extracts audio, removes background noise, and converts the speech into text. The transcribed text can be reviewed and edited by the user for accuracy, with the option to translate it into one or more languages. The translated text is then synchronized with the original video, and text-to-speech technology is used to convert the translated text back into audio. Users can then download the final video with synchronized audio and text. The Admin role is responsible for managing the platform's functionality, including authenticating users, overseeing the uploading and processing of videos, and ensuring system security.

B. Advantages of Proposed System.

- Improved Accuracy
- Versatility
- Robustness
- Real Time Processing

C. System Requirement Specifications.

Software Requirements Specifications

- Operating System: Windows, macOS or Linux.
- Programming Language: Python 3.x
- Libraries and Frameworks: Google cloud API's, Transformer Models, React, Flask.

Hardware Requirements Specifications.

- Processor: Inter i5
- RAM: 8GB (minimum)
- Storage: Atleast 1Gb of free disk

III. LITERATURE SURVEY

This project builds upon various studies that explore to provide an efficient and seamless workflow for audio extraction, transcription, translation, and synchronization of video content.

Presents the DeepSpeech model, a speech-to-text system based on deep learning. The system achieves high accuracy by converting speech from audio files into written text. This technique could be a key component of your project, as audio extraction and transcription are crucial steps in converting speech from videos into text. The model's efficiency and accuracy in transcribing spoken words from audio files would enhance your system's text generation. [1].

In this study, the authors explore various ASR systems for transcribing spoken language into text. The focus is on how noise removal and pre-processing techniques, such as speech enhancement, improve the recognition accuracy of ASR systems. This is especially relevant to your project, where audio extracted from videos may contain background noise, requiring noise removal to ensure high-quality transcription. [2].

Bapna et al. discuss multilingual speech recognition and translation systems that handle multiple languages in audio files. For your VidText project, the ability to convert transcribed text into different languages can be a valuable feature. By incorporating multilingual transcription and translation, your project can cater to users across various linguistic regions, enhancing accessibility [3].

This paper focuses on video summarization techniques, which extract the most significant segments of a video for easier consumption. Although the focus is not directly on transcribing, techniques for segmenting videos and identifying key scenes can improve the accuracy and relevance of audio and text extraction. For Vid-Text, this approach could enhance the efficiency of transcribing by selecting relevant portions of a video for processing. [4].

This study explores methods for automatic grammatical error correction in transcribed text, focusing on the application to noisy or imperfect speech-to-text systems. for improved speech recognition in noisy environments. This hybrid approach could be particularly useful. Integrating such a system into your Vid-Text project would allow users to convert the transcribed text back into speech. This would offer a fully interactive, multimedia experience, making it easier for users to access the content in both written and audio formats. [5]

IV. SYSTEM DESIGN AND METHODOLOGY

A. Design.

The architecture of the Vid-.Text system is designed to provide an efficient and seamless workflow for audio extraction, transcription, translation, and synchronization of video content. The system involves three primary actors:

User, Admin, and System. Users begin by registering and logging into the platform with valid credentials. After successful authentication, they upload video files, from which the system extracts audio, removes background noise, and converts the speech into text. The transcribed text can be reviewed and edited by the user for accuracy, with the option to translate it into one or more languages. The translated text is then synchronized with the original video, and text-to-speech technology is used to convert the translated text back into audio. Users can then download the final video with synchronized audio and text. The Admin role is responsible for managing the platform's functionality, including authenticating users, overseeing the uploading and processing of videos, and ensuring system security. Admins also monitor user activities, review transcriptions and translations, and perform data backups to ensure the integrity of stored content.

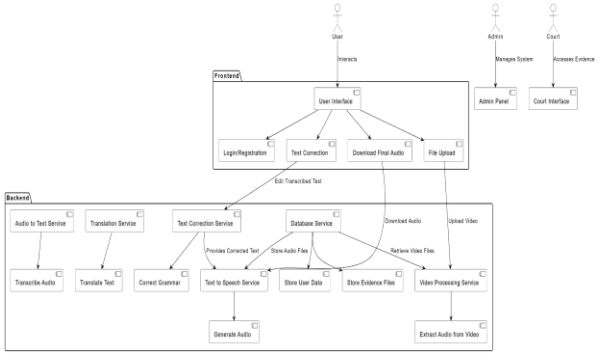
Data Ingestion:

This module is responsible for acquiring data from various sources.. The purpose of this step is to ensure that raw input data is efficiently funneled into the system for processing.

After ingestion, the data undergoes preprocessing, a critical step to clean, structure, and prepare it for analysis. This includes operations such as data cleaning, tokenization, stop-word removal, stemming, and vectorization. This phase ensures that the input data is formatted appropriately and optimized for downstream opinion analysis tasks.

B. Methodology.

Enable users to upload a video file, extract audio, and process the audio for transcription and further analysis. Users upload video files through the UI. The backend receives the file and stores it temporarily. The uploaded video file is processed using FFmpeg to extract the audio track. This involves separating the video and audio streams from the file. Once the audio is extracted, noise reduction algorithms like pydub or librosa are applied to enhance the audio quality and ensure clarity for transcription. Use a tool like LanguageTool API or Grammarly API to automatically identify and correct common grammatical issues, such as tense mistakes, punctuation errors, and sentence structure problems. Allow the user to manually edit the text for accuracy. This provides the user with control over the final content. The corrected text is displayed in the UI



for user verification.

Use FFmpeg or pydub to adjust the timing of the generated audio. This involves matching the speech timing with the visual cues in the video, such as lip movements. The synchronized audio is then merged back into the original video using FFmpeg, creating a new video file where audio and video are perfectly aligned.

D. Modules.

The system is divided into several modules to handle different parts of the process efficiently:

- *Audio Extraction Module – Extracts audio from video using FFmpeg and saves it in a compatible format.*
- *Speech-to-Text Module – Converts the extracted audio into text using OpenAI's Whisper model.*
- *Grammar Correction Module – Fixes grammar and punctuation errors using LanguageTool.*
- *Subtitle Generation Module – Formats the corrected text into SRT (SubRip Subtitle) files with proper timestamps.*
- *Subtitle Integration Module – Embeds the generated subtitles back into the video using FFmpeg.*
- *Frontend User Interface – A React-based UI that allows users to upload videos and download the final processed video.*

E. Algorithms.

The system relies on various algorithms to ensure accurate and efficient processing:

- FFmpeg Processing Algorithm – Extracts and integrates audio/subtitles while maintaining quality.
- Whisper Speech-to-Text Algorithm – Uses AI to transcribe spoken words into text.
- Grammar Correction Algorithm – Improves text accuracy using rule-based and machine learning techniques.
- SRT Synchronization Algorithm – Ensures subtitles align with speech timestamps.
- FastAPI Request Handling Algorithm – Manages backend communication efficiently.

F. Technologies.

The project uses modern technologies to ensure performance and usability:

- Frontend: React.js for UI, Axios for API communication.
- Backend: FastAPI for handling requests and integrating processing modules.
- Processing Tools: FFmpeg (media processing), Whisper (transcription), LanguageTool (text correction).
- Storage: Local file system with potential for cloud integration.
- Deployment: Can be hosted using Docker, Nginx, or cloud services for scalability.

VII. CONCLUSION

The proposed system successfully automates the process of audio extraction, speech-to-text conversion, grammar correction, and subtitle synchronization in videos. By integrating FFmpeg for media processing, OpenAI's Whisper for transcription, LanguageTool for grammar correction, and React with FastAPI for a seamless user experience, the system provides an efficient and accurate solution for subtitle generation.

Compared to traditional methods, this approach significantly reduces manual effort, improves accuracy, and ensures better synchronization of subtitles with speech. The user-friendly interface allows easy video uploads and downloads, making the system accessible to a wide range of users, from content creators to professionals in media and education.

ACKNOWLEDGMENT

This Project Survey would be incomplete without the introduction of the people who made it possible and whose guidance, encouragement crowns all efforts with success.
We are thankful to our honorable Prof. G. Chandramohan Reddy Sir, Principal of MGIT and Dr. D. Vijaya Lakshmi, Professor and HOD, Department of IT, MGIT, for providing excellent infrastructure and a conductive atmosphere for completing the project successfully.
We are deeply indebted to our project coordinator, Mrs. Dr. N. Sree Divya, Assistant Professor, Department of IT, MGIT, for her constant support, valuable suggestions, immense patience, and expertise throughout the course of our project.
We are profoundly grateful to our project coordinator, Dr. N Sree Divya, Assistant Professor, Department of IT, MGIT, for her unwavering support, and valuable suggestions throughout the course of our project.

REFERENCES

[1] Chen, L., Li, H., Zhang, Y., & Li, S. (2023). Automatic speech recognition system for real-time transcription and translation using deep neural networks. Journal of Visual Communication and Image Representation, 92, Art. no. 103689. https://doi.org/10.1016/j.jvcir.2023.103689

[2] Kumar, A., & Singh, R. (2023). Enhancing video accessibility using AI-driven captioning and multilingual translation systems. Multimedia Tools and Applications, 81, 17329–17352. https://doi.org/10.1007/s11042-022-12358-9

[3] Li, J., Deng, L., Dong, X., & Zhou, X. (2021). End-to-end speech recognition for noisy environments with advanced denoising autoencoders. Neurocomputing, 453, 478–489. https://doi.org/10.1016/j.neucom.2021.04.064

[4] Mehrotra, H., Jain, M., & Singh, A. (2022). Video-audio synchronization techniques for dynamic multimedia applications: A review. Journal of Real-Time Image Processing, 19(3), 483–501. https://doi.org/10.1007/s11554-022-01154-2

[5] Ramesh, M., Rani, K. A., & Babu, P. (2023). Multilingual video-to-text conversion and transcription system using transfer learning. Future Internet, 15(4), Art. no. 153. https://doi.org/10.3390/fi15040153

[6] Saha, R., Dey, S., & Gupta, N. (2023). Speech synthesis and synchronization for multilingual text-to-speech systems in educational platforms. International Journal of Speech Technology, 26(3), 543–556. https://doi.org/10.1007/s10772-023-09916-0