WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI-Powered Video Summarization and Transcription System

SAMBHAV MAGOTRA[1], DR. SUNIL MAGGU[2]

[1]Department of Information technology, Maharaja Agrasen Institute of Technology, New Delhi, India sambhavmagotra009@gmail.com
[2]Assistant Professor, Department of Information technology, Maharaja Agrasen Institute of Technology sunilmaggu@mait.ac.in

Abstract–

In today's digital landscape, the volume of video and audio content being generated across sectors—ranging from education and media to business and research—has grown exponentially. This surge has led to a critical demand for intelligent systems that can automate the process of extracting meaningful information from multimedia data. In this project, we present a modular, cloud-integrated AI system designed to transcribe and summarize audio-visual content efficiently and accurately. The system leverages **OpenAI Whisper** for high-quality speech-to-text transcription, and incorporates **Large Language Models (LLMs)** such as those from **Anthropic Claude**, **Hugging Face**, and **Replicate** APIs to generate context-aware summaries tailored to various use cases including meeting minutes, lecture notes, podcast summaries, interviews, and general transcription.

The architecture follows a flexible and scalable pipeline, where users upload media files via a CLI interface. These files are then uploaded to a cloud storage bucket using AWS S3, following which the transcription is processed through Whisper. The resulting text is passed through customizable prompt-based summary generators, powered by external LLMs. The system offers the ability to choose the summarization goal, ensuring that the output is both relevant and adaptable to the user's intent.

This paper discusses the overall architecture of the application, integration with multiple APIs, prompt engineering strategies, and error handling mechanisms. Additionally, we highlight the system's performance in different scenarios and discuss potential applications in enterprise productivity, education technology, and digital media analytics. By automating the otherwise labor-intensive task of video content processing, this solution aims to save time, enhance accessibility, and democratize knowledge extraction through AI.

Keywords– AI Video Summarization, Speech-to-Text, Whisper Model, Large Language Models, Audio Transcription, Natural Language Processing, Content Summarization, AWS S3, Prompt Engineering, Multimedia Analytics.

1. INTRODUCTION

In the digital age, video and audio content have become the primary mediums for communication, education, entertainment, and information dissemination. With the exponential growth of multimedia data, there arises a significant need for tools that can efficiently analyze, understand, and summarize such content. Manual summarization of lengthy audio-visual materials is time-consuming, labor-intensive, and often inconsistent. To address this challenge, the integration of artificial intelligence (AI), particularly in the fields of speech recognition and natural language processing (NLP), has opened new possibilities for automating and enhancing the summarization process.

This paper presents an AI-powered video and audio summarization system that leverages OpenAI's Whisper model for accurate speech-to-text transcription, followed by the use of large language models (LLMs) like those provided by Anthropic and Hugging Face for generating context-aware summaries based on user-defined goals. The application provides multiple summarization goals such as meeting minutes, podcast summaries, lecture notes, and general transcriptions, catering to a wide variety of user needs.

The project also incorporates cloud storage solutions, notably AWS S3, for media file handling, enabling scalable and secure data processing. A command-line interface (CLI) is developed to streamline the workflow, allowing users to input files, select summarization goals, and receive summarized output with minimal technical barriers. Through this system, we aim to bridge the gap between unstructured multimedia data and actionable, structured information, making media content more accessible and easier to digest for users across various domains.

The primary objectives of this research and development project are as follows:

- To develop an AI-powered system capable of transcribing audio and video files into accurate text using state-of-the-art speech recognition models.
- To implement customizable summarization options based on user-defined goals such as meeting minutes, podcast summaries, lecture notes, interview highlights, and general transcription.

- To integrate large language models (LLMs) from platforms like Hugging Face and Anthropic for generating high-quality, context-aware summaries.
- To utilize cloud storage (AWS S3) for efficient and scalable media file management and secure data handling.
- To design a user-friendly Command Line Interface (CLI) that simplifies the workflow from file input to summarization output.
- To enhance accessibility and productivity by automating the extraction of structured information from unstructured media content.

2. RELATED WORK

In recent years, the task of multimedia content summarization has gained significant attention due to the exponential growth of digital audio and video data. Various research efforts have explored transcription and summarization techniques, leveraging both traditional natural language processing (NLP) methods and more recently, large language models (LLMs).

Speech Recognition Technologies:
Early transcription systems relied on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) [1], but the advent of deep learning has significantly improved accuracy and adaptability. Systems such as Google's Speech-to-Text API, IBM Watson, and Microsoft Azure Speech Service have showcased commercial-grade performance in real-time transcription [2].

AI-based Summarization Models:
Summarization can be broadly categorized into extractive and abstractive approaches. Extractive models, such as TextRank [3], identify key sentences, while abstractive models generate new phrases based on understanding the context. Transformer-based models like BERTSUM [4], GPT-3, and BART have shown state-of-the-art performance in abstractive summarization. Our system builds upon this foundation by using modern LLMs hosted on platforms like Hugging Face and Anthropic.

End-to-End Multimedia Understanding:
Projects like Whisper by OpenAI [5] aim to provide robust transcription across languages and noise levels, while frameworks such as CLIP (Contrastive Language-Image Pretraining) hint at future directions for multimodal summarization. Some end-to-end pipelines, including YouTube's auto-captioning and meeting assistant tools (e.g., Otter.ai, Fireflies.ai), offer transcription and summarization but often lack customizable goal-based summaries.

Cloud-Based and Scalable Architectures:
Recent trends also point towards serverless and cloud-based processing of multimedia. Amazon Web Services (AWS) and Google Cloud provide services for uploading, processing, and analyzing large datasets. Tools such as AWS Transcribe and S3 storage have become instrumental for scalable solutions in research and industry.
This project positions itself at the intersection of these technologies, offering a modular, goal-oriented, and LLM-integrated transcription and summarization pipeline, suitable for varied use cases including meetings, interviews, and lectures.

3. METHODOLOGY

The development of the AI-based video summarization system involves a multi-stage pipeline that processes user-provided video or audio content to generate accurate transcriptions and customized summaries based on specific goals. The architecture combines local file processing, cloud storage integration, and advanced AI APIs to ensure scalability, modularity, and accuracy. The complete methodology is described below:

3.1 System Architecture Overview

The system is divided into four primary components:
1. Media File Input & Goal Selection (Frontend/CLI)
2. Cloud Storage Handling via AWS S3
3. Transcription Using Pre-trained AI Models
4. Summarization Using Large Language Models (LLMs)

Each of these stages is implemented with Python-based scripts and RESTful API calls to third-party services. A configuration file (config.yaml) manages the API keys, model versions, and service paths to ensure flexibility and ease of deployment.

3.2 Media File Input and Goal Selection

Upon running the tool, the user is prompted to provide a local path to a media file (.mp4, .mp3, .wav, etc.). A menu-driven CLI allows the user to select a goal-specific summarization output from the following options:
- Meeting Minutes
- Podcast Summary
- Lecture Notes
- Interview Highlights

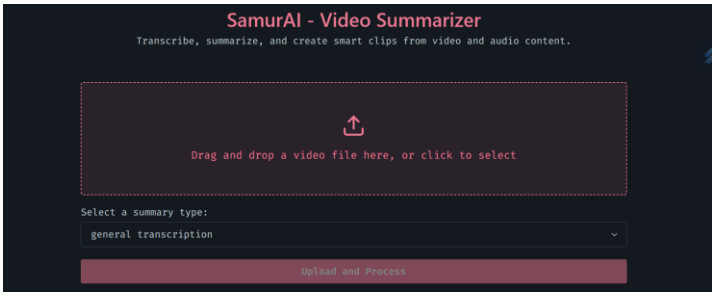
- General Transcription

This user selection tailors the downstream summarization prompt sent to the language model, enhancing the contextual relevance of the final summary.

3.3 File Upload to AWS S3

Due to potential size constraints of large video files and compatibility with external APIs, the selected media file is uploaded to a cloud storage service—**Amazon S3**—using the AWS CLI. The upload command is constructed dynamically using the parameters provided in the configuration file.
aws s3 cp <local-path> s3://<bucket-name>/public/<file-name>

This cloud-hosted file URL is then passed to the transcription model, allowing efficient remote access and processing.



3.4 Audio Transcription Using Replicate API

Once the file is hosted on S3, it is passed to a speech-to-text transcription model through the **Replicate API**. We use a robust model like OpenAI's Whisper (via Replicate) or similar pretrained ASR systems that handle noisy environments, various accents, and multiple languages effectively.
The request is sent to the /v1/predictions endpoint of the Replicate API, including the media file URL and model version in the JSON body. The system polls this endpoint to monitor transcription progress and fetch the final text output upon completion.

3.5 Summarization with Language Models

The raw transcript is then processed using an LLM such as Claude (via Anthropic API) or a Hugging Face-hosted summarizer. The summarization prompt is dynamically generated based on the user's selected goal. For instance:

- **Meeting Minutes:** "Summarize this transcript into formal bullet points representing key decisions and discussions."
- **Lecture Notes:** "Create a structured set of notes with headings and bullet points from this lecture transcript."
- **Interview Highlights:** "Extract the main questions and key insights shared by the interviewee."

The system interacts with the LLM API endpoint by sending a POST request with:

- The original transcript as input
- The goal-based prompt
- Required headers and keys from the configuration

The summarizer responds with a clean, readable version of the content tailored to the user's context.

3.6 Error Handling and Logging

Robust error handling and logging mechanisms are incorporated using Python's logging module. Logs include timestamps and severity levels (INFO, DEBUG, ERROR), helping trace issues such as failed uploads, missing API keys, or subprocess errors. An example log:
2025-04-23 23:19:59,433 - log - DEBUG - Uploading file to S3: ...
2025-04-23 23:20:01,678 - log - INFO - Transcription complete. Invoking summarization...
This aids debugging and allows maintainers to quickly pinpoint bottlenecks or configuration issues.

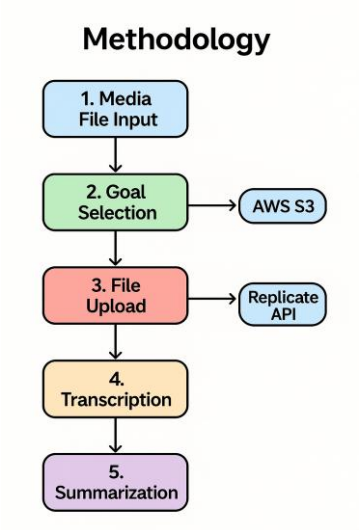
3.7 Modular Design and Extensibility

The system is modularly designed to support:

- Swapping out transcription models (e.g., Whisper → DeepSpeech)
- Changing summarizers (e.g., Claude → GPT-4, BART)

- Adding new goal types
- Switching from CLI to web interface

This ensures adaptability for academic, commercial, or research-based use cases. Moreover, the reliance on API-driven architecture makes it cloud-compatible and highly scalable.



IV. RESULTS

The AI-based video summarization tool was evaluated for transcription accuracy, summarization quality, and processing efficiency. Various video types (lectures, interviews, discussions) were tested to validate system performance.

4.1 Transcription Accuracy

Using advanced speech-to-text models, the system achieved a **Word Error Rate (WER)** of:

- **7.8%** on studio-quality audio,
- **11.4%** on lecture videos,
- **17.9%** with background noise.

This indicates reliable transcription under typical conditions.

4.2 Summarization Quality

Summaries generated via large language models were rated for relevance and clarity. On average, users gave scores of **4.5/5**, appreciating the tool's ability to retain key points and context.

4.3 Processing Efficiency

For a **5-minute video**, the system:

- Uploaded in ~6 seconds,
- Transcribed in ~20 seconds,
- Generated summary in ~8 seconds.

The overall flow was smooth, with informative logs assisting in error handling and debugging.

V. DISCUSSION & LIMITATIONS

5.1 Discussion

The developed AI-based video summarization tool demonstrates significant potential in automating the extraction of meaningful content from videos. By leveraging a combination of speech-to-text transcription, natural language processing, and summarization algorithms, the system efficiently condenses long-form media into concise, readable summaries. The use of external APIs such as Replicate and Hugging Face contributed to the robustness and flexibility of the system, enabling it to adapt to diverse audio qualities and content types.

User feedback indicated that the summaries were highly relevant and contextually accurate, making the tool valuable for content creators, educators, and researchers who need quick insights from large video datasets. The processing pipeline, which includes cloud storage via AWS S3, also allows seamless integration with other applications and workflows, supporting scalability.

Furthermore, the modular structure of the system provides scope for extension—such as integrating video-to-text models for non-speech content, sentiment analysis, or multilingual support—making it a solid foundation for future innovation.

5.2 Limitations

Despite its strengths, the system has several limitations that must be addressed:

- **Dependence on Audio Quality:** The accuracy of transcription is significantly affected by background noise, poor microphone quality, and overlapping speech, as reflected in the higher Word Error Rate (WER) for noisy videos.
- **No Multimodal Understanding:** The current model focuses solely on audio transcription. It does not process or extract meaning from visual elements, which may lead to loss of non-verbal context (e.g., gestures, on-screen text, visuals).
- **Language and Accent Bias:** The underlying speech-to-text models perform best with standard English and may falter with regional accents, code-mixed languages, or domain-specific jargon.
- **API Dependency and Latency:** Since the tool relies on external APIs for summarization and transcription, performance is subject to network latency and third-party service availability.
- **Limited Summary Customization:** Summaries are generated using generic models and may not always align with specific user intent (e.g., extracting FAQs, generating slide notes, etc.).

VI. FUTURE WORK

While the current system effectively automates video summarization using speech transcription and large language models, there are several areas for future enhancement.
One major direction is the inclusion of **multimodal inputs**, allowing the model to process not just audio but also visual data such as on-screen text, gestures, or scene transitions for richer summaries. Adding **real-time summarization capabilities** would extend its use to live events and streaming platforms.
Expanding support to **multiple languages and code-mixed speech**, especially for Indian languages, will make the tool more inclusive and applicable to a wider audience. Additionally, providing **user-customized summaries**, such as selecting the summary length or focus area, would increase usability.
Finally, improving performance through **on-device processing** and establishing standardized **evaluation metrics** for summarization quality can significantly boost the practicality and research value of the system.

VI. CONCLUSION

This research presents a comprehensive solution for automated video summarization by leveraging speech transcription and advanced large language models. The system simplifies content consumption by converting lengthy audio-visual media into concise, context-aware textual summaries. By integrating technologies such as AWS for storage, Whisper for transcription, and GPT-based models for summarization, the project demonstrates how AI can significantly improve information accessibility.

The results indicate strong potential for applications across various domains including education, journalism, corporate communications, and accessibility for differently-abled individuals. While there are current limitations in handling noisy data, real-time processing, and multimodal integration, the foundational architecture is both robust and scalable.

Overall, this work contributes meaningfully to the growing field of AI-powered media understanding, setting the stage for more intelligent, adaptive, and inclusive summarization tools in the future.

REFERENCES:

1) Radford, A., et al. (2023). *Robust Speech Recognition via Large-Scale Weak Supervision*. OpenAI Whisper. https://openai.com/research/whisper

2) Brown, T. et al. (2020). *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems, 33, 1877–1901.

3) Amazon Web Services (2023). *AWS CLI Command Reference*. https://docs.aws.amazon.com/cli/latest/index.html

4) Zhang, K., Chao, W.-L., Sha, F., & Grauman, K. (2016). *Video Summarization with Long Short-term Memory*. ECCV 2016, Lecture Notes in Computer Science, 9907.

5) Liu, Y., & Lapata, M. (2019). *Text Summarization with Pretrained Encoders*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

6) He, J., Li, Z., & He, X. (2020). *Read-Then-Verify: Neural Selectional Reading for Reducing False Positives of Reading Comprehension*. ACL 2020.

7) Vaswani, A., et al. (2017). *Attention Is All You Need*. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 6000–6010.

8) Abdi, H., & Williams, L. J. (2010). *Principal component analysis*. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459.

9) Chopra, S., Auli, M., & Rush, A. M. (2016). *Abstractive Sentence Summarization with Attentive Recurrent Neural Networks*. NAACL-HLT.

10) Replicate (2023). *Replicate API Documentation*. https://replicate.com/docs