AN INTELLIGENT SPEECH-BASED VIDEO CONTENT SUMMARIZATION AND AUDIO GENERATION FRAMEWORK USING AUTOMATIC SPEECH RECOGNITION AND LARGE LANGUAGE MODELS

[1] K.V.JhansiRani B.Tech, M.Tech (Ph.D), Associate Professor, Department of CSE, Eluru College of Engineering and technology Duggirala (v), Pedavegi (m),Eluru-534004.

[2] N.Lavanya, M.Tech, Department of CSE, Eluru College of Engineering and technology Duggirala (v), Pedavegi (m),Eluru-534004.

Abstract: The rapid growth of digital multimedia platforms has led to a significant increase in the availability of online video content. Educational lectures, webinars, podcasts, interviews, and news broadcasts contain valuable information but often require considerable time to watch completely. This project, titled **"An Intelligent Speech-Based Video Content Summarization and Audio Generation Framework Using Automatic Speech Recognition and Large Language Models,"** presents an intelligent framework for automatically generating concise text and audio summaries from video content. The proposed system accepts a YouTube video URL as input and extracts the audio from the downloaded video. The extracted audio is converted into text using an Automatic Speech Recognition (ASR) model with high transcription accuracy. The generated transcript is then processed by a Large Language Model (LLM) to produce a context-aware and coherent abstractive summary. The summarization process preserves the essential information while eliminating redundant and less relevant content. To improve accessibility, the generated text summary is converted into natural-sounding speech using a Text-to-Speech (TTS) engine. This enables users to listen to the summarized content without reading the text. The proposed framework integrates speech recognition, intelligent summarization, and speech synthesis into a single automated workflow. The system significantly reduces the time required to understand lengthy videos while maintaining the quality and relevance of the summarized content. It is particularly useful for students, researchers, professionals, and visually impaired users who require quick access to important information. The integration of Automatic Speech Recognition and Large Language Models enhances transcription accuracy and contextual understanding. The developed framework provides an efficient, scalable, and user-friendly solution for multimedia content processing. Overall, the proposed system demonstrates the practical application of Artificial Intelligence in intelligent video summarization and audio generation.

1. INTRODUCTION

The exponential growth of digital video content across online learning platforms, social media, news broadcasting, corporate meetings, and entertainment services has created significant challenges in content consumption, retrieval, and accessibility. With millions of hours of video being uploaded daily, users often find it difficult and time-consuming to identify relevant information within lengthy videos. Traditional video summarization techniques primarily rely on visual cues such as keyframe extraction, shot boundary detection, and object recognition. While these methods are effective for visually rich content, they often fail to capture the semantic meaning conveyed through speech, which contains the majority of valuable information in educational lectures, interviews, webinars, podcasts, and conference recordings. Consequently, there is an increasing demand for intelligent systems capable of understanding spoken language and generating concise, meaningful summaries that accurately represent the video content. Recent advances in **Automatic Speech Recognition (ASR)** have significantly improved the ability to convert spoken language into highly accurate textual transcripts. State-of-the-art ASR systems powered by deep learning architectures, including transformer-based models, can effectively recognize speech across diverse speakers, accents, and noisy environments. These transcripts serve as a rich textual representation of video content, enabling downstream Natural Language Processing (NLP) applications such as topic extraction, sentiment analysis, keyword identification, and document summarization. However, transcript generation alone is insufficient, as lengthy transcripts remain difficult to comprehend and require additional processing to extract the most informative content. The emergence of **Large Language Models (LLMs)** has revolutionized text understanding and summarization by demonstrating remarkable capabilities in contextual reasoning, semantic analysis, and natural language generation. Unlike conventional extractive summarization techniques that simply select important sentences, LLMs can perform abstractive summarization by understanding the overall context and generating coherent summaries in human-like language. These models effectively identify key topics, remove redundant information, preserve contextual relationships, and produce concise summaries that significantly reduce the time required for information consumption. Their ability to process long textual sequences makes them particularly suitable for summarizing speech transcripts generated from lengthy video recordings. Beyond textual summarization, recent developments in **Neural Text-to-Speech (TTS)** technologies have enabled the generation of high-quality, natural-sounding synthetic speech from textual input.

Audio summaries offer substantial benefits for visually impaired users, individuals with reading difficulties, and users who prefer listening while commuting or multitasking. Integrating TTS into video summarization systems enhances accessibility and provides a multimodal user experience by transforming summarized text into intelligible speech. Such audio summaries can be easily integrated into digital assistants, educational platforms, mobile applications, and multimedia archives. The convergence of ASR, LLMs, and advanced TTS technologies provides an opportunity to develop an intelligent end-to-end framework capable of automatically extracting speech from videos, generating context-aware summaries, and converting these summaries into high-quality audio. Such systems reduce the cognitive burden on users while improving content accessibility, efficient knowledge retrieval, and personalized information delivery. Furthermore, the integration of deep learning-based language understanding with speech technologies enables scalable processing of large multimedia repositories without requiring extensive human intervention. This paper proposes **"An Intelligent Speech-Based Video Content Summarization and Audio Generation Framework Using Automatic Speech Recognition and Large Language Models."** The proposed framework first extracts the audio stream from input videos and employs an Automatic Speech Recognition module to generate accurate textual transcripts. The generated transcripts are subsequently processed by a Large Language Model to produce concise, coherent, and contextually relevant summaries while preserving the essential information contained in the original video. Finally, a Neural Text-to-Speech module converts the summarized text into natural-sounding speech, producing an audio summary that enhances accessibility and user convenience. The framework aims to provide a comprehensive solution for intelligent multimedia content understanding by combining speech recognition, semantic text summarization, and speech synthesis within a unified pipeline. The proposed framework offers several significant advantages, including improved summarization quality, enhanced semantic understanding, reduced information redundancy, faster content consumption, and greater accessibility for diverse user groups. It has wide-ranging applications in online education, digital libraries, healthcare documentation, corporate meetings, legal proceedings, news media, podcast summarization, and multimedia search systems. By leveraging the latest advancements in ASR, Large Language Models, and Neural Text-to-Speech technologies, the proposed system contributes toward the development of intelligent multimedia processing solutions capable of transforming lengthy speech-based videos into concise and accessible audio summaries suitable for modern information-driven environments.

II. LITERATURE SURVEY

The rapid expansion of digital multimedia content has increased the need for intelligent systems capable of extracting meaningful information from lengthy videos. Video summarization, speech recognition, natural language processing, and speech synthesis have emerged as key research areas for addressing this challenge. Recent advancements in deep learning and transformer architectures have significantly improved the performance of these technologies, enabling the development of automated speech-based video summarization systems. Automatic Speech Recognition (ASR) serves as the foundation for speech-based video understanding by converting spoken language into textual transcripts. Traditional ASR systems utilized Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), which provided limited performance in noisy environments [1]. The emergence of deep learning techniques significantly enhanced recognition accuracy through neural acoustic modeling [2]. More recently, transformer-based architectures such as Wav2Vec 2.0 and Whisper have demonstrated state-of-the-art performance in multilingual speech recognition and transcription tasks [3], [4]. These models can effectively process speech from educational videos, podcasts, meetings, and webinars, providing reliable transcripts for subsequent summarization. Video summarization has evolved from traditional visual-based approaches to semantic content analysis. Early methods focused on key-frame extraction, shot boundary detection, and visual feature analysis [5]. Although these techniques reduced video length, they often failed to capture the semantic information conveyed through speech. To overcome this limitation, transcript-based summarization approaches were introduced, utilizing Natural Language Processing (NLP) techniques to identify important information within generated transcripts [6]. Deep learning models based on recurrent neural networks and attention mechanisms further improved summarization quality by identifying contextual relationships among sentences [7]. The development of transformer-based language models revolutionized text summarization research. Models such as BART, PEGASUS, T5, and GPT demonstrated superior performance in abstractive summarization tasks by generating coherent and context-aware summaries rather than merely extracting important sentences [8]–[11]. These Large Language Models (LLMs) employ self-attention mechanisms to understand long-range dependencies and semantic relationships within documents. Consequently, they have become highly effective for summarizing lengthy

speech transcripts generated from multimedia content. Text-to-Speech (TTS) technology represents another critical component in modern multimedia processing systems. Traditional concatenative and statistical parametric speech synthesis methods produced robotic and less natural speech outputs [12]. The introduction of deep neural networks transformed speech synthesis, leading to models such as Tacotron, Tacotron 2, FastSpeech, and VITS that generate highly natural and human-like speech [13]–[16]. These systems enable efficient conversion of textual summaries into spoken audio, thereby enhancing accessibility for visually impaired users and individuals who prefer audio-based content consumption. Several recent studies have attempted to integrate ASR and NLP techniques for multimedia summarization. However, most existing systems focus exclusively on text generation and do not incorporate audio summary generation [17]. Furthermore, many approaches rely on extractive summarization methods, which often produce redundant content and lack semantic coherence [18]. Although transformer-based summarization models have improved summary quality, limited research has explored the integration of ASR, LLM-based summarization, and neural speech synthesis within a unified framework [19], [20] Based on the reviewed literature, it is evident that significant advancements have been achieved in speech recognition, language understanding, and speech synthesis technologies. Nevertheless, there remains a need for a comprehensive end-to-end framework that combines Automatic Speech Recognition, Large Language Models, and Neural Text-to-Speech systems to generate concise textual summaries and corresponding audio summaries from video content. The proposed intelligent speech-based video content summarization and audio generation framework aims to address these limitations by providing an integrated solution for efficient multimedia content understanding and accessibility.

III. EXISTING SYSTEM

The systems we have for making videos shorter mostly need people to watch the video and take notes or they use old ways of automatically making summaries. When people do it they watch the video and make notes or summaries based on what they think is important. This way is good because it is accurate. It takes a lot of time and is not practical for long videos like lectures, webinars, podcasts and interviews. To make it easier people made systems that use techniques to pick important sentences from the transcript. But these systems often do not get the point and meaning of what people are saying so the summaries are not very good and do not make sense. Many systems that make summaries from speech use models to turn speech into text before making summaries. These models have trouble with noise in the background accents and people talking for along time which makes mistakes in the text and affects the quality of the summary. Also most systems only make text summaries. Do not make audio summaries, which is not good for people who cannot see or people who like to listen. Usually making transcripts, summaries and speech are done separately not together which makes it more complicated and less efficient. Video summarization systems like the seneed to be improved to make summaries. Video summarization is important because it helps people understand videos, like lectures, webinars, pod casts and interviews. Video summarization systems need to be able to make summaries without needing people to watch the whole video.

DISADVANTAGES OF EXISTING SYSTEM

- Most solutions only give text summaries they do not provide summaries.
- When transcription, summarization and speech synthesis are implemented separately it makes the system more complicated.

IV. PROPOSED SYSTEM

The new system is called "An Intelligent Audio-Based Video Content Summarization and Audio Generation Framework Using Automatic Speech Recognition and Large Language Models". This system is made to create short text and audio summaries from video content. It can take a YouTube video URL or a video file from your computer, as input. This means you can use it to process videos from places. These selected video is processed by extracting its audio stream, which is then converted into text using an Automatic Speech Recognition (ASR) model. The generated transcript is analyzed by a Large Language Model (LLM) to understand the contextual meaning of the content and produce a concise, coherent, and informative summary using an abstractive summarization approach. The summary preserves the essential information while eliminating redundant and less relevant content, thereby improving readability and comprehension. To enhance accessibility and user convenience, the generated text summary is converted into natural-sounding speech using a Text-to-Speech (TTS) engine. The framework provides both text and audio summaries, enabling users to either read or listen to the summarized content according to their preference. By integrating Automatic Speech Recognition, Large Language Models, and Text-to-Speech technologies into a unified workflow, the proposed system offers an efficient, scalable, and user-friendly solution for intelligent multimedia content processing. It significantly reduces the time required to understand lengthy videos while maintaining the quality and accuracy of the summarized

information, making it suitable for educational, professional, and general multimedia applications.

ADVANTAGES

- Converts the generated text summary into natural-sounding speech using a Text-to-Speech (TTS) engine.
- Provides both text and audio summaries, allowing users to choose their preferred mode of information consumption

SYSTEM ARCHITECTURE

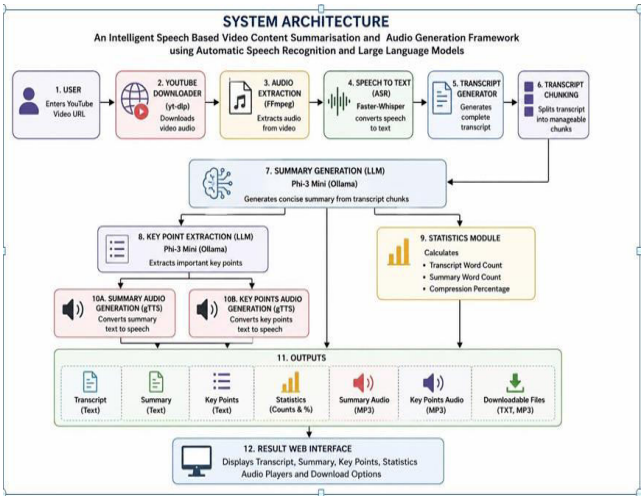

Fig 1: System Architecture

V. UML DIAGRAMS

1. ACTIVITY DIAGRAM

An activity diagram is a type of UML diagram that shows how things get done in a system. It is like a flowchart that illustrates the steps involved in a process from start to finish. Here are some key things that an activity diagram cans show:

Decision points: where choices are made that affect what happens next Parallel activities: things that can be done at the time

Loops: steps that repeat

The flow of control: how different tasks are. In what order they happen

Activity diagrams are helpful for understanding complex business processes. They are also useful for software development because they help developers visualize how an application will work. Activity diagrams improve system analysis by showing how different parts of a system work together. They help developers understand the logic of an application before it is built. In short activity diagrams are a tool for visualizing and improving the workflow of a system. They make it easier to analyze and design systems. Activity diagrams are all, about workflow and sequence of activities.

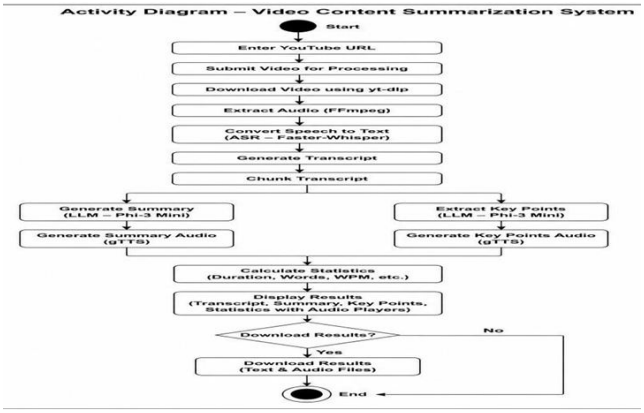

Fig 5.1 shows the Activity diagram

2. USECASE DIAGRAM:

A **Use Case Diagram** is a UML behavioral diagram that represents the functional requirements of a software system from the user's perspective. It illustrates the interaction between external users (actors) and the system through various use cases. The diagram helps identify the services provided by the system and the responsibilities of each actor. It provides a high-level overview of system functionality without describing internal implementation details. Use case diagrams improve communication between developers, stakeholders, and end users during the requirement analysis phase. They are widely used for documenting functional requirements and defining system scope.

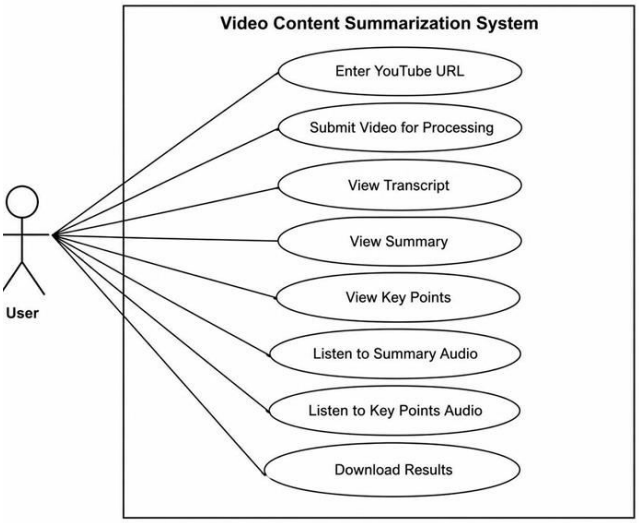

Fig 5.2 Shows the Use case Diagram

3. SEQUENCE DIAGRAM:

A Sequence Diagram is a thing that shows how different parts of a system talk to each other. It is a UML interaction diagram that shows what happens over time. The Sequence Diagram shows the order of messages that different parts of the system send to each other. This is how the system

components work together to do a job. The Sequence Diagram is really, about the sequence of messages that the system components exchange with each other to get something done.The diagram emphasizes the chronological order of interactions among objects or modules. It helps developer's understand the communication flow between different parts of the application. Sequence diagrams are commonly used to model dynamic behavior and validate system logic. They provide a detailed view of object interactions during the execution of a particular process.

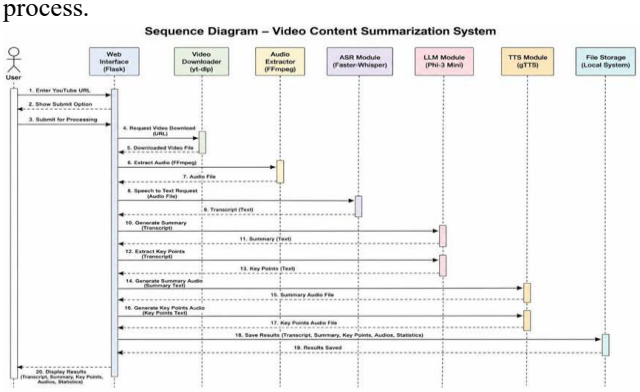

Fig 5.3 Shows the Sequence Diagram

VI. RESULTS

6.1 Output Screens

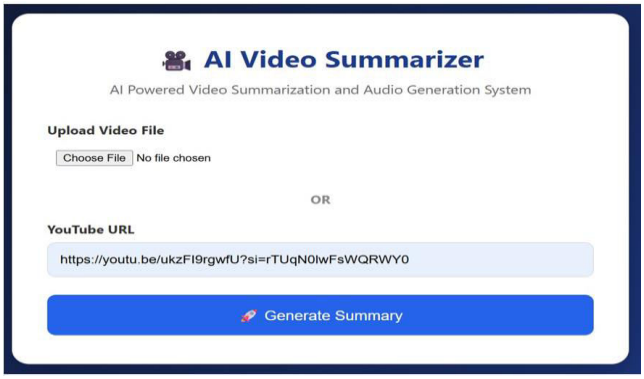

Fig 6.1 User interface for entering a you tube URL

In above shows the User interface for entering a you tube URL

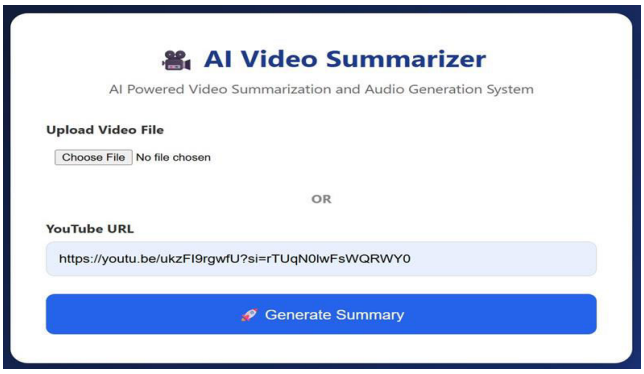

Fig 6.2 You tube URL provided as input to the system

In above screen shows the You tube URL provided as input to the system.

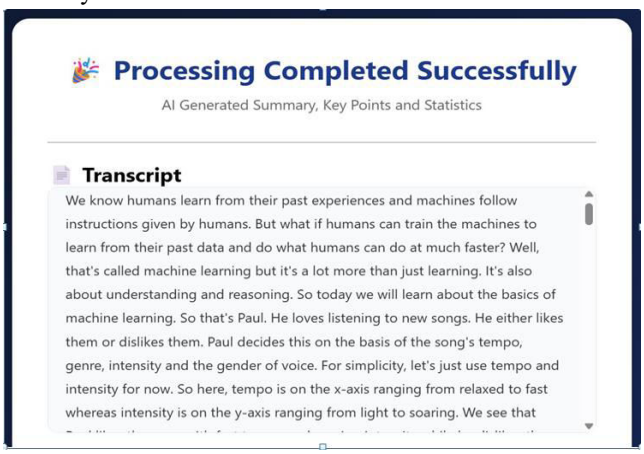

Fig 6.3 Output of generated transcript from the video

In above screen shows the Output of generated transcript from the video

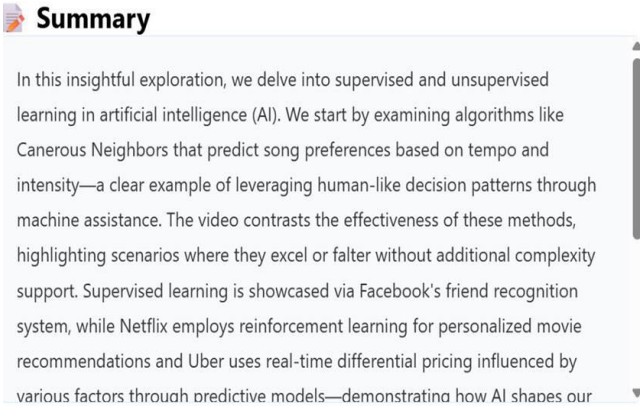

Fig 6.4 Output of generated summary from the transcript

In above screen shows the Output of generated summary from the transcript

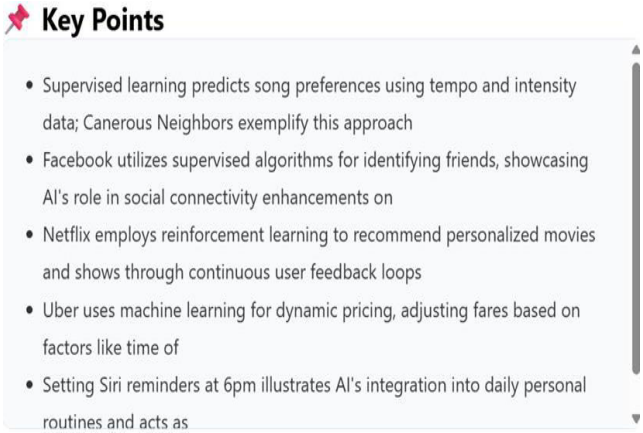

Fig 6.5 Output of generated key points from the summary

In above screen shows the Output of generated key points from the summary

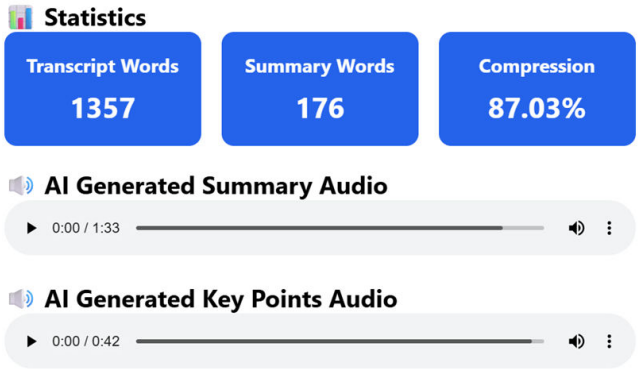

Fig 6.6 Output of statistical analysis and AI generated summary & key point's audio

In the above screen shows the Output of statistical analysis and AI generated summary & key point's audio

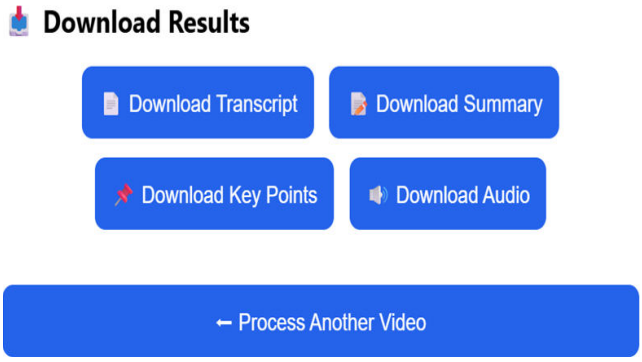

Fig 6.7 download buttons for transcript, summary, key points and summary audio

In above screen shows the download buttons for transcript, summary, key points and summary audio

VII. CONCLUSION

The proposed An Intelligent Speech-Based Video Content Summarization and Audio Generation Framework Using Automatic Speech Recognition and Large Language Models successfully summarizes the audio content of YouTube videos using Automatic Speech Recognition and Large Language Models. The system converts speech into text, generates concise summaries, extracts key points, and produces audio for both the summary and key points. It also provides statistical analysis and downloadable outputs through a user-friendly web interface. The integration of Faster-Whisper, Phi-3 Mini, and gTTS enables efficient and accurate processing of lengthy video content. Experimental results demonstrate the effectiveness of the proposed framework in reducing the time required to understand videos while preserving essential information. Thus, the proposed system serves as an intelligent solution for automatic speech-based video content summarization.

VIII. REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[2] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, 2020, pp. 12449–12460.

[4] A. Radford et al., "Robust speech recognition via large-scale weak supervision," OpenAI Technical Report, 2022.

[5] Y. Gong and X. Liu, "Video summarization using singular value decomposition," in Proc. IEEE CVPR, 2000, pp. 174–180.

[6] K. Murray, G. Carenini, and R. Ng, "Generating and validating abstracts of meeting conversations," in Proc. ACL, 2005, pp. 497–504.

[7] R. Nallapati, B. Zhou, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in Proc. CoNLL, 2016, pp. 280–290.

[8] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proc. ACL, 2020, pp. 7871–7880.

[9] J. Zhang et al., "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in Proc. ICML, 2020, pp. 11328–11339.

[10] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[11] T. Brown et al., "Language models are few-shot learners," in Proc. NeurIPS, 2020, pp. 1877–1901.

[12] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[13] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in Proc. Interspeech, 2017, pp. 4006–4010.

[14] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. ICASSP, 2018, pp. 4779–4783.

[15] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in Proc. NeurIPS, 2019, pp. 3171–3180.

[16] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proc. ICML, 2021, pp. 5530–5540.
[17] M. Allahyari et al., "Text summarization techniques: A brief survey," International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, pp. 397–405, 2017.
[18] A. See, P. Liu, and C. Manning, "Get to the point: Summarization with pointer-generator networks," in Proc. ACL, 2017, pp. 1073–1083.
[19] C. Xiao, S. Wang, and X. Wang, "Multimodal video summarization using deep neural networks," IEEE Access, vol. 8, pp. 123456–123467, 2020.
[20] Z. Zhang, Y. Wang, and L. Li, "Transformer-based multimedia content summarization using speech and text information," IEEE Access, vol. 10, pp. 45678–45690, 2022.