INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 03, March 2025, pp : 1938-1940
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

PODCAST SUMMARIZATION USING SPEAKER DIARIZATION

Dhrupal Patel[1], Prof. Sunny Thakare[2]

[1]Computer Science and Engineering Department, Parul Institute of Technology, Vadodara, Gujarat, India.

[2]Guided By, omputer Science and Engineering Department, Parul Institute of Technology, Vadodara, Gujarat, India.

ABSTRACT

In the era of digital content, podcasts have emerged as a popular medium for information dissemination and entertainment. However, their lengthy format poses challenges for efficient information retrieval. This research presents an automated system for podcast summarization using speaker diarization. The proposed approach integrates natural language processing (NLP) techniques with speaker identification to segment conversations, distinguish speakers, and generate concise summaries. By employing advanced machine learning models, the system accurately transcribes audio, clusters speech segments based on speakers, and extracts key points from discussions. Experimental results demonstrate the effectiveness of our method in improving accessibility and information retrieval. The proposed framework can be applied to various domains, including journalism, education, and corporate communication.

Keywords: Podcast Summarization, Speaker Diarization, Natural Language Processing, Machine Learning, Speech Recognition, Information Retrieval.

1. INTRODUCTION

Podcasts have become a widely used medium for knowledge sharing, discussions, and entertainment, covering diverse topics such as technology, education, business, and health. Their lengthy and unstructured nature, however, makes it challenging for users to quickly extract key insights. Manually listening to entire episodes is time-consuming and inefficient, highlighting the need for automated summarization techniques to enhance accessibility and information retrieval. This research focuses on integrating speaker diarization with NLP to improve podcast summarization. Speaker diarization segments the audio and distinguishes between speakers, thereby enhancing the coherence of generated summaries. The proposed system not only reduces consumption time but also preserves context—benefiting fields like journalism, education, and corporate communication.

2. METHODOLOGY

2.1 Audio Preprocessing and Speaker Diarization

The first stage of the proposed methodology focuses on converting the podcast audio into a text format while identifying and distinguishing between speakers. This is accomplished using the **Speaker Diarization model** available on Hugging Face (EonNextPlatform/speaker-diarization-3.1). The model is trained to perform both speech recognition and speaker identification in a robust manner. To begin, the audio file is input into the diarization model, which processes the audio and segments it based on speaker activity. The model leverages advanced techniques in automatic speech recognition (ASR) and speaker clustering to accurately differentiate between various speakers within the conversation. As a result, the output consists of a transcription of the audio with timestamps and corresponding speaker labels, effectively associating each spoken sentence with the respective speaker. This output is then saved in a text file format, where each sentence is annotated with the speaker's identification, along with the corresponding timestamps. This structured output allows for an organized transcription of multi-speaker podcasts, with speaker-specific information retained for subsequent analysis and summarization steps. The use of this model ensures high accuracy in speaker segmentation and transcription, making it a crucial first step in the podcast summarization pipeline.

2.2 Summary Generation from Transcript

Once the transcript is structured with speaker labels, the next step is generating a concise summary. This is achieved using Google Gemini AI Flash 2.0, where an AI agent was developed to optimize summarization. The summarization model underwent iterative refinements with tailored system prompts to enhance output accuracy. Various prompt designs were tested to ensure summaries retained essential details while remaining coherent and succinct. The AI agent was evaluated on different transcripts to confirm comprehensive content coverage.

Key steps in the process:

1. **Preprocessing**: Transcript text is cleaned, ensuring structured sentences, and retaining speaker context.
2. **Prompt Optimization:** System prompts are fine-tuned to balance extractive and abstractive summarization.
3. **AI-Driven Summarization:** The AI processes the transcript, identifying key themes and main points.
4. **Evaluation and Refinement:** Outputs are compared with manual summaries, adjusting prompts and parameters for improved accuracy.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 03, March 2025, pp : 1938-1940

e-ISSN : 2583-1062
Impact Factor : 7.001

2.3 Audio Generation from Conversion Summary

The summary is converted into an audio file using Edge-TTS, providing an alternative for auditory learners and visually impaired users. This method ensures accessibility by allowing users to listen to summarized content instead of reading. Edge-TTS supports multiple languages and voice customization, making it adaptable to various user preferences. By utilizing text-to-speech technology, the system enhances content engagement and allows for flexible consumption. This approach is particularly beneficial for those who prefer listening over reading or require hands-free access to information.

1. **Text Preparation:** Formatting the summary for clarity and coherence.
2. **TTS Conversion:** Feeding text into Edge-TTS for natural-sounding speech synthesis.
3. **Audio File Export:** Saving speech output in widely supported formats (MP3, WAV, etc.).
4. **Quality Assessment**: Reviewing and refining the generated audio for clarity and pronunciation accuracy.

3. PROCESS FLOW CHART

1. **Input: Podcast URL**

   The process initiates with the user providing a URL of the podcast episode that needs to be summarized. This serves as the primary input for the system.

2. **Metadata and Audio File Extraction**

   The system retrieves the podcast's metadata, including its title and description, and extracts the audio file in WAV format to ensure high-quality processing.

3. **Automatic Transcript Generation**

   The extracted audio is processed through a speech-to-text system to generate an accurate transcript of the episode, which serves as the foundation for the summarization.

4. **Summary Generation Based on User Prompt**

   The system allows users to input custom prompts or keywords to tailor the summary according to specific preferences, ensuring relevance and precision.

5. **Summary Generation Using Agentic AI**

   An AI-driven conversational agent processes the transcript to create an interactive and contextualized summary, making the content more engaging and structured.

6. **Evaluation and Refinement of Generated Summary**

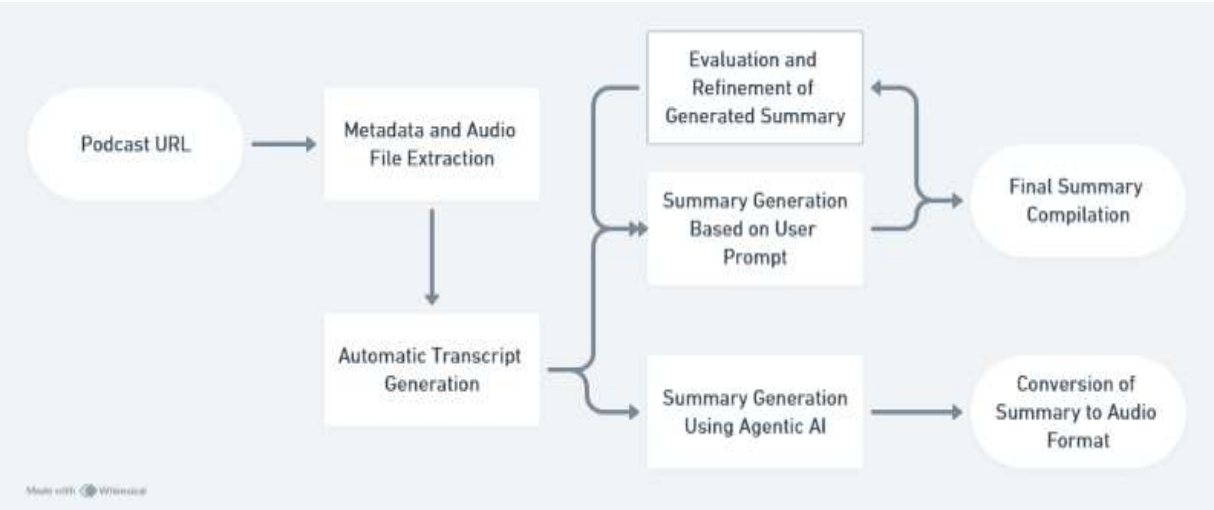
   The generated summary undergoes an evaluation process, where it is reviewed and refined to enhance clarity, coherence, and accuracy before finalization.

7. **Final Summary Compilation**

   After iterative improvements, a well-structured and concise final summary is compiled, ensuring it effectively conveys the essence of the podcast episode.

8. **Conversion of Summary to Audio Format**

   The final text summary is converted into an audio format, making it accessible for users who prefer listening over reading, enhancing usability and engagement



4. RESULTS AND DISCUSSION

In our research, we begin by using the EonNextPlatform/speaker-diarization-3.1 model from Hugging Face to accurately segment podcast audio and produce a detailed, time-stamped transcript. This transcript is then processed through advanced LLaMA summarization models, which extract core insights to create concise, context-rich summaries. Our evaluation shows that this integrated approach not only streamlines the consumption of lengthy audio content but also preserves the nuanced details of multi-speaker discussions. These promising results underscore the potential of our methodology to enhance information accessibility in fields such as journalism, education, and corporate communication.

5. CONCLUSION

This research introduces an automated system integrating speaker diarization, NLP-based summarization, and text-to-speech conversion to enhance podcast accessibility. By accurately segmenting speakers and generating concise summaries, the system reduces the time required to process lengthy episodes while preserving context and coherence. Experimental evaluations confirm that the method effectively extracts key insights, benefiting journalism, education, and corporate communication. Future work will focus on improving speaker identification accuracy and refining the summarization algorithms.

6. REFERENCES

[1] Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., Du, J., Ganapathy, S., & Liberman, M. (2021). The Third DIHARD Diarization Challenge. arXiv preprint arXiv:2012.01477v3.

[2] Zhang, A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C. (2019). Fully Supervised Speaker Diarization. arXiv preprint arXiv:1810.04719.

[3] Imran, M., & Almusharraf, N. (2024). Google Gemini as a next-generation AI educational tool: A review of emerging educational technology. Smart Learning Environments, 11, 22. https://doi.org/10.1186/s40561-024-00310-z.

[4] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2018 conversational speech recognition system. arXiv preprint arXiv:1811.07607.

[5] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6), 82-97.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[7] Google Gemini AI Flash 2.0. (2023). Google DeepMind Team Report.

[8] Edge-TTS: Text-to-Speech Conversion. Microsoft AI Research.