Research Article

Revolutionizing Meeting Analysis: An AI-Powered Approach to Automatic Transcription and Summarization

Sibin Thomas*

*Corresponding author:** Sibin Thomas, Tech Lead, USA, E-mail: sibin_thomas15@hotmail.com

A B S T R A C T

This research proposes a novel approach to automatic meeting summarization using generative AI and LLMs1. The system leverages advanced techniques to transcribe audio recordings, extract key points, and generate informative summaries. By automating this process, organizations can significantly improve meeting efficiency, productivity, and decision-making. The proposed system offers a wide range of applications beyond audio summarization, including document summarization, content creation and question answering. This research demonstrates the potential of generative AI and LLMs to revolutionize the way organizations conduct and manage meetings.

Keywords: Meeting summarization, Generative AI, LLMs, Natural language processing, Automatic transcription, Audio analysis, AI-powered meeting notes, Speech-to-text

1. Introduction

The rapid growth of remote work and virtual meetings has created a pressing need for efficient tools to capture, analyze, and summarize meeting content[2]. Traditional methods, such as manual note-taking, are time-consuming, error-prone, and often fail to capture the nuances of complex discussions. As a result, organizations struggle to efficiently review, analyze and act upon meeting outcomes.

To address these challenges, this research proposes a novel approach to automatic meeting summarization using generative AI and LLMs. By leveraging the power of these technologies, the proposed system aims to overcome the limitations of manual methods and provide a more accurate, efficient, and informative solution for meeting analysis.

The system operates by first transcribing audio recordings of meetings into text using state-of-the-art speech-to-text models[3].

Subsequently, a generative AI model is employed to summarize the transcribed text, extracting key points, identifying actionable items, and maintaining contextual understanding. The model is trained on a large dataset of meeting transcripts and summaries to ensure the quality and accuracy of the generated summaries.

In the following sections, we will delve into the technical details of the proposed system, including its architecture, training methodology, and evaluation metrics. We will also discuss the ethical considerations associated with the development and deployment of such a system. Finally, we will conclude by highlighting the potential impact of this research on organizations and the future of meeting analysis.

2. A High-Level Architecture for Video Calling Applications

A video calling application, such as Google Meet or Zoom, requires a sophisticated infrastructure to facilitate real-time audio and video communication across vast distances[4]. At its core, a video calling application consists of several interconnected components that work together to provide a seamless user experience.

Client Application: This is the interface that users interact with. It handles tasks like capturing audio and video from the user's device, processing the media data, and rendering the received media streams[5]. The client application also manages the user interface, including controls for muting, unmuting, sharing screens and more.

Signaling Server: The signaling server acts as a central hub for communication between clients[6]. It manages session establishment, media negotiation and the exchange of control messages. When a user initiates a call, the signaling server coordinates the process of connecting the user to other participants, ensuring compatibility between their devices and applications.

Media Server: The media server is responsible for transporting and processing media data between clients[7]. It handles tasks like mixing multiple video streams, adding effects, and ensuring quality of service. In some cases, the media server may also act as a relay, forwarding media data between clients that cannot establish a direct connection due to network restrictions.

Database: A database stores information about users, the past of sessions and other useful information[8]. In this way, the app can keep track of missed calls, handle user accounts and offer personalized features.

Network Infrastructure: A very important part of video talking is the network infrastructure that supports it[9]. This includes the internet, routers, switches and other network equipment that makes it easier for clients and hosts to send and receive data.

Example: Take a look at a simple video call situation to see how these parts work together. The user's client app talks to the signaling server when they join a meeting. The signaling server then talks about media parameters and sets up a link between the participants' devices while exchanging information with the other participants. The media server starts sending audio and video data between the clients as soon as the link is made[10]. It makes sure that the streams are in sync and of high quality.

3. Problem statement

Traditional ways of summarizing meetings often depend on taking notes by hand, which takes a lot of time, can lead to mistakes, and leaves out important information[11]. This limitation makes it harder to quickly go over, think about, and act on what was said in meetings.

The hard part is getting good recordings, understandings, and summaries of meetings with many people, changing audio and video quality, and the subtleties of human language[12]. Current methods often have trouble pulling out key points, finding things that can be done and keeping context in mind.

4. Generative AI and LLMs for Audio Summarization in Calling Applications

The proposed system consists of several interconnected components:

Audio Ingestion and Storage: Audio files from meetings are uploaded to a cloud-based storage service. To optimize upload efficiency, large files are divided into smaller chunks and uploaded in parallel using a multi-part upload mechanism. Transcription: Once the audio file is stored, a transcription process is initiated[3]. The audio is converted into text using a state-of-the-art speech-to-text model. The transcribed text is then saved as a JSON file in a separate cloud storage location.

Text Summarization: A generative AI model is employed to summarize the transcribed text[13]. The model is configured with specific parameters to control the length of the summary, the level of detail and the balance between factual accuracy and creativity. The generated summary is stored in a database.

5. Model Training and Prompts

The generative AI model is trained on a large dataset of meeting transcripts and summaries[14]. To ensure the quality of the generated summaries, the model is provided with prompts that specify the desired structure and content. These prompts include instructions for identifying the meeting agenda, key points discussed, decisions made, action items, participant information and next steps. To protect user privacy, the model is trained to avoid including sensitive data such as passwords, credit card numbers and social security information.

6. Model Parameters and Optimization

To achieve optimal performance, the generative AI model is fine-tuned using various parameters:

Max_tokens_to_sample: This parameter controls the maximum number of tokens generated in the summary, allowing for customization based on meeting length and desired level of detail.

Temperature: A lower temperature setting produces more deterministic and focused summaries, ensuring that the generated content is faithful to the factual information.

Top_p: This parameter allows the model to explore a broader range of token options while still favoring the most likely candidates, enabling more natural and abstractive summarization.

Top_k: This parameter limits the selection of the next token to the top k highest-probability vocabulary tokens, ensuring coherence and relevance.

6.1. Ethical Considerations: The development of this system raises important ethical concerns[15]. The model must be trained to avoid generating harmful, unwarranted, or illegal content. Additionally, measures must be implemented to prevent unauthorized access to sensitive information and to protect against prompt overriding or hacking attempts.

6.2. Conclusion: This research proposes a system that can automate the transcription and summarization of meeting audio recordings. By leveraging advanced AI techniques, this system can provide valuable insights to organizations and improve the efficiency of their meeting processes. Future work will focus on refining the model's performance, addressing ethical concerns and exploring additional applications for this technology.

7. Applications Beyond security cameras

The generative AI and LLM approach suggested in this study can be used for more than just summarizing audio files, because it can process and create natural language text[16]. The following are some possible uses:

Translation and analysis of the meeting: Real-time transcription: The system can be added to videoconferencing platforms so that meeting conversations are transcribed in real

time[17]. This way, people who have trouble hearing or who need to review the material later can still access them.

Keyword extraction: The model can pull out important words and sentences from the transcribed text, which lets users find important conversations and topics quickly.

Sentiment analysis: The system can look at how people feel during meetings and give information about the general mood and tone of the talk.

7.1. Document Summarization

Summarizing research papers: The model can be used to summarize research papers, which helps researchers quickly understand the main points and results.

Legal document summarization: The system can summarize legal papers like court transcripts or contracts to help lawyers find the information they need quickly.

Summarizing customer feedback: The model can outline customer feedback, which can show how satisfied customers are and help find ways to make things better.

7.2. Making and creating content

Generating articles: The model can write articles or blog posts based on a theme or prompt, which helps people who make content make better content.

Creative writing: The model can be used to come up with creative text, like plays, poems, and stories. It can also be used to inspire writers and give them new ideas.

Translation: The model can translate text between languages[18], but the accuracy will depend on how complicated the text is and how many languages are used.

7.3. Question Answering

Answering questions from customers: The model can be used to help customers and answer their questions, which will make the jobs of human customer service reps easier.

Help with research: The model can answer questions about study.

8. Conclusion

Using generative AI and LLMs, this study has come up with a new way to automatically summarize meetings[1]. The suggested system does a good job of fixing the problems with current manual methods by correctly recording, comprehending, and summarizing how meetings work.

Organizations can get a lot out of the system's ability to transcribe audio recordings, pull out important points, and make useful summaries. Automating this process can help businesses save time, get more done, and learn important things from the talks they have in meetings.

The suggested solution could also be used for other jobs besides just summarizing audio files. These include summarizing documents, making content, and answering questions. As technology keeps getting better, we can expect to see even more creative and useful uses for it.

Based on the research provided in this paper, generative AI and LLMs can automatically summarize meetings. This shows that this technology has the potential to completely change how organizations run and run meetings. More work will be done in the future to improve the model's performance, deal with social issues and look for other uses for this technology.

9. References

1. Ouyang L, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems, 2022;35:27730-27744.

2. Rogerson-Revell P. "The changing face of meetings." Work, employment and society 31(1):163-179.

3. Li P, et al. "A survey of conversational speech recognition systems." IEEE Access 8:134531-134553.

4. Weinstein S. "The impact of videoconferencing on education, health care and business." Videoconferencing and telecommunications, 1983;1(1):3-18.

5. Sarangi S, et al. "VCast: a scalable videoconferencing system." Proceedings of the 2003 ACM SIGCOMM workshop on Internet network management, 2003.

6. Robertello J. "Finding the right signaling protocol for your real-time communications app." Twilio blog, 2022.

7. Johnston A, et al. "An analysis of the real-time transport protocol (RTP)." Internet Engineering Task Force (IETF), 1996.

8. Kemper A, Eickler A. Data management: DBMS and database technology. Walter de Gruyter GmbH and Co KG, 2017.

9. Tanenbaum AS, Wetherall DJ. Computer networks. Pearson Education India, 2011.

10. Raca D and Lehnert J. "Network-aware peer-to-peer video conferencing." Proceedings of the 17th International Workshop on Network and Operating Systems Support for Digital Audio and Video. 2007.

11. Elsweiler D, et al. "Taking meeting minutes: An exploratory study of strategies and their effectiveness." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2007.

12. Janin A, et al. "The ICSI meeting corpus." 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003.

13. Mehdad Y, et al. "A survey on automatic text summarization." arXiv preprint arXiv:1908.08345, 2019.

14. Prasad R et al. "The AMI meeting corpus: A pre-announcement." Machine learning for multimodal interaction. Springer, Berlin, Heidelberg, 2006.

15. Brundage M, et al. "The malicious use of artificial intelligence: Forecasting, prevention and mitigation." arXiv preprint arXiv:1802.07228, 2018.

16. Radford A, et al. "Language models are unsupervised multitask learners." OpenAI blog, 2019:9.

17. Johnson M, et al. "Google's multilingual neural machine translation system: enabling zero-shot translation." Transactions of the Association for Computational Linguistics 5 (2017): 339-351.

18. Bahdanau D, et al. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014). Advances in Neural Information Processing Systems, 2014;35(2022):27730-27744.