AI-Powered Online Meeting Summary Generator Using Speech Recognition and Large Language Models

Nitin Nagnath Sugare[1], Dr. Sushil Kulkarni[2]
[1]*M. Tech 3rd Semester Student, Department of Computer Science and Information Technology, MBES College of Engineering, Ambajogai, India*
[2]*Guide, Head and Guide, Department of Computer Science and Information Technology, MBES College of Engineering, Ambajogai, India*

***Abstract*—This paper presents the design and implementation of an AI-powered online meeting summary generator that automates transcription, summarization, and report generation from virtual meetings. Artificial Intelligence (AI) has emerged as a powerful solution for automating meeting transcription, summarization, and documentation. The system combines browser-based audio capture, speech recognition, natural language processing, and large language model-based summarization to generate concise meeting reports with minimal manual effort. The proposed architecture utilizes the Web Audio API for audio acquisition, Mozilla DeepSpeech for speech-to-text conversion, and the Gemma language model for abstractive summarization and action-item extraction. The implementation focuses on privacy-aware processing, real-time usability, and support for modern collaboration platforms. Experimental observations demonstrate that the proposed approach significantly reduces manual note-taking effort while producing structured summaries suitable for professional documentation.**

***Index Terms*—Online Meeting Summarization, Speech Recognition, DeepSpeech, Large Language Models, Gemma, Natural Language Processing, Web Audio API, Artificial Intelligence.**

I. INTRODUCTION

The widespread adoption of remote work and virtual collaboration platforms has created a growing demand for intelligent meeting documentation systems. Traditional note-taking is time-consuming, inconsistent, and often fails to capture all important discussions, decisions, and action items. Automated meeting summarization addresses these limitations by converting spoken conversations into structured textual reports.

AI-powered meeting assistants address this challenge by automatically capturing audio, converting speech into text, identifying important discussion points, and generating concise summaries. These systems combine speech recognition, natural language processing (NLP), machine learning, and large language models to transform unstructured conversations into actionable information.

Recent advances in speech recognition and large language models have enabled the development of systems capable of understanding conversational context and generating concise summaries. Unlike conventional transcription tools that only convert speech into text, modern AI-powered systems can identify key discussion points, extract decisions, and produce professional meeting minutes.

The proposed implementation aims to provide a browser-based solution that captures meeting audio, filters irrelevant noise, transcribes speech, generates summaries, and creates downloadable reports. The system is designed to support major online meeting platforms while minimizing privacy concerns through local audio processing wherever possible.

II. LITERATURE REVIEW

Several researchers have investigated automatic meeting summarization using deep learning and natural language processing techniques. Transformer-based models have demonstrated strong performance in abstractive summarization tasks, while benchmark datasets such as QMSum have enabled standardized evaluation of meeting summarization systems.

The QMSum benchmark introduced a structured evaluation framework for query-based meeting summarization. Researchers used encoder-decoder models and transformer architectures to analyze meeting conversations and generate targeted summaries. The study highlighted challenges related to factual consistency and evaluation methodologies.

Recent studies have explored action-item extraction, topic segmentation, and query-focused summarization. However, many existing solutions depend heavily on cloud-based processing, lack multilingual support, or struggle with noisy audio conditions. Research on browser-based privacy-preserving meeting intelligence systems remain relatively limited.

The literature indicates that combining speech recognition with large language models offers significant potential for improving meeting documentation quality. Nevertheless, challenges related to speaker identification, factual consistency, and real-time processing continue to require further investigation.

III. RESEARCH GAP AND PROBLEM STATEMENT

Existing meeting summarization systems often focus on either transcription accuracy or summary generation but rarely provide an integrated browser-based implementation suitable for everyday professional use. Many available solutions require extensive cloud processing, raising privacy concerns for organizations handling sensitive discussions.

The primary problem addressed in this work is the development of a real-time system capable of accurately transcribing online meetings and generating concise, actionable summaries while reducing manual effort and maintaining practical deployment simplicity.

IV. PROPOSED SYSTEM

The proposed system consists of five major stages: audio capture, preprocessing, speech recognition, transcript analysis, and summary generation.

Audio from online meetings is captured through the browser using the Web Audio API. The captured signal undergoes preprocessing to reduce background noise and improve speech clarity. The processed audio is then passed to the DeepSpeech engine for transcription. The resulting transcript is analyzed using natural language processing techniques before being summarized by the Gemma large language model. Finally, the system generates a structured report containing key points, decisions, and action items.

The architecture is modular, allowing individual components to be upgraded independently as newer speech recognition or language models become available.

V. SYSTEM ARCHITECTURE

The system architecture follows a sequential processing pipeline:

Meeting Platform → Audio Capture → Speech Recognition → Transcript Processing → AI Summarization → Report Generation.

The browser extension captures audio streams from meeting applications such as Google Meet, Zoom, and Microsoft Teams. Audio preprocessing improves recognition quality, while the transcription module converts speech into text. The summarization engine then generates concise meeting notes and extracts actionable tasks before the documentation module creates a downloadable report.

VI. METHODOLOGY

The implementation uses the Web Audio API for real-time audio acquisition within the browser environment. Mozilla DeepSpeech performs speech-to-text conversion using a deep neural network trained on large speech datasets. Transcript preprocessing includes noise filtering, sentence segmentation, and text normalization.

For summarization, the Gemma model is employed to generate abstractive summaries that preserve the meaning of the original discussion while reducing unnecessary details. Prompt engineering techniques guide the model to produce structured outputs containing meeting objectives, discussion points, decisions, and action items.

The complete workflow includes:

1. Capture meeting audio.
2. Filter background noise.
3. Convert speech to text.

4. Clean and segment the transcript.
5. Generate an AI summary.
6. Extract action items.
7. Produce a downloadable report.

VII. IMPLEMENTATION DETAILS

The frontend is implemented as a browser-based interface that provides controls for starting and stopping recording sessions. The backend processing pipeline is developed in Python and integrates DeepSpeech and the Gemma API.

The system includes four primary modules:

- Audio Extraction Module – captures meeting audio streams.
- Transcription Module – converts speech into text.
- Summarization Module – generates concise summaries and action items.
- Documentation Module – creates structured reports in downloadable formats.

The modular design simplifies maintenance and future feature expansion.

VIII. EXPERIMENTAL SETUP AND RESULTS

Testing was performed using recorded online meeting sessions containing multiple speakers and varying background noise levels. The evaluation focused on transcription quality, summary readability, processing time, and usability.

The system successfully generated structured meeting summaries that captured major discussion topics and action items. Local audio processing reduced privacy concerns compared with fully cloud-based solutions. The implementation demonstrated practical usability for academic and professional meetings.

Although quantitative benchmarking was limited in this implementation study, qualitative evaluation indicated that the generated summaries were substantially more concise and organized than raw transcripts.

IX. ADVANTAGES AND LIMITATIONS

Advantages of the proposed system include real-time transcription, automated summarization, reduced manual effort, browser-based operation, and automatic report generation. The architecture also supports integration with popular online meeting platforms.

The current implementation has several limitations. Speaker identification is not fully implemented, performance may decrease in highly noisy environments, and internet connectivity is required for AI-based summarization. Browser compatibility is currently optimized for Chromium-based browsers.

X. CONCLUSION

This paper presented the implementation of an AI-powered online meeting summary generator that combines speech recognition and large language models to automate meeting documentation. The proposed system captures meeting audio through the browser, performs transcription using DeepSpeech, and generates structured summaries using the Gemma model.

The implementation demonstrates that integrating speech recognition with modern language models can significantly reduce the effort required for meeting documentation while producing concise and actionable reports. The modular architecture provides a practical foundation for future enhancements and real-world deployment.

XI. FUTURE SCOPE

Future systems are expected to integrate multimodal AI capable of analyzing speech, text, images, and shared documents simultaneously. Speaker diarization and multilingual summarization will improve usability across global organizations.

Future improvements may include speaker diarization, multilingual summarization, sentiment analysis, automatic task assignment, integration with project management platforms, and deployment of more advanced transformer-based speech recognition models.

REFERENCES

[1] Vaswani *et al*., "Attention Is All You Need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017.

[2] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proc. 2020 Conf. on Empirical Methods in Natural Language

Processing: System Demonstrations (EMNLP), pp. 38–45, 2020, doi: 10.18653/v1/2020.emnlp-demos.6.

[3] Mozilla, "DeepSpeech: Open-Source Speech Recognition Engine," GitHub repository. [Online]. Available: https://github.com/mozilla/DeepSpeech

[4] L. Zhong, J. Yin, W. Shang, X. Jiang, Z. Yang, Y. Chai, W. Wang, C. Li, and N. Duan, "QMSum: A New Benchmark for Query-Based Multi-Domain Meeting Summarization," in Proc. 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 5905–5921, 2021.

[5] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009.