AI-Based Meeting Documentation Systems: A Comprehensive Review of Speech Recognition and Summarization Technologies

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]*Head and Guide, Department of Computer Science and Information Technology, MBES College of Engineering, Ambajogai, India*

***Abstract*—The growth of remote collaboration platforms has transformed the way organizations communicate, coordinate projects, and make decisions. However, documenting online meetings remains a challenging and time-consuming activity. Artificial Intelligence (AI) has emerged as a powerful solution for automating meeting transcription, summarization, and documentation. This review paper presents a comprehensive study of AI-powered online meeting summary generation systems, focusing on speech recognition, natural language processing, transformer-based architectures, large language models, and browser-based deployment approaches. The paper reviews recent research contributions, analyzes technologies such as Web Audio API, DeepSpeech, and Gemma-based language models, and discusses their application in real-time meeting intelligence systems. A detailed examination of system architecture, data flow, implementation modules, benefits, limitations, and research gaps is provided. The study concludes that intelligent meeting assistants can significantly improve productivity, reduce manual effort, and enhance organizational knowledge management.**

I. INTRODUCTION

Online meetings have become an integral part of modern workplaces due to the rise of remote and hybrid work models. Platforms such as Zoom, Microsoft Teams, and Google Meet allow geographically distributed teams to collaborate effectively. Despite these benefits, capturing meeting outcomes, decisions, and action items remains a significant challenge. Manual note-taking often leads to incomplete records and consumes valuable time. 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 transformer architectures and large language models have significantly improved the quality of generated summaries. Organizations increasingly seek intelligent tools capable of extracting decisions, tracking action items, and providing searchable meeting archives. Such capabilities improve productivity, accountability, and knowledge retention. The purpose of this review paper is to examine the technologies, methodologies, and research trends related to AI-powered meeting summary generation systems and to identify opportunities for future innovation.

II. BACKGROUND AND FUNDAMENTALS

Meeting summarization systems integrate multiple AI disciplines. Automatic Speech Recognition (ASR) converts spoken language into text. NLP techniques process transcripts to identify topics, keywords, and relationships. Summarization algorithms condense lengthy conversations into concise representations while preserving essential information.

Speech recognition systems have evolved from statistical approaches to deep learning-based models. DeepSpeech utilizes recurrent neural networks and connectionist temporal classification to perform end-to-end speech recognition. Transformer-based

language models further improve contextual understanding and language generation.

Meeting summarization can be extractive or abstractive. Extractive summarization selects important sentences directly from transcripts, while abstractive summarization generates new sentences that capture the overall meaning. Large language models have made abstractive summarization increasingly practical and effective.

Modern systems also employ noise reduction, speaker segmentation, and sentiment analysis to improve output quality. These components collectively enable intelligent documentation workflows.

III. LITERATURE REVIEW

Research on meeting summarization has expanded rapidly over the last decade. Transformer-based summarization models introduced significant improvements in contextual understanding and coherence. The work 'Transforming Meeting Transcript Summarization using Transformers' demonstrated that transformer architectures can outperform traditional recurrent neural network approaches.

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.

Research on automated meeting minutes generation explored the extraction of decisions and action items from transcripts. Speech Emotion Recognition-based systems attempted to identify emotional signals that influence meeting outcomes. These approaches demonstrated the potential of combining speech analytics with summarization.

Recent studies emphasize action-item extraction, conversational intelligence, and long-document summarization. Large language models have significantly improved summary quality but continue to face challenges related to hallucinations, privacy, and computational cost.

Overall, the literature demonstrates substantial progress while revealing research gaps in multilingual support, speaker diarization, privacy-preserving deployment, and real-time performance optimization.

IV. TECHNOLOGIES AND ALGORITHMS

The proposed meeting summary generator integrates several technologies. Web Audio API provides browser-based access to audio streams and enables real-time audio capture without requiring complex installations. DeepSpeech performs speech-to-text conversion and supports offline deployment scenarios. Natural Language Processing pipelines perform tokenization, sentence segmentation, keyword extraction, and transcript cleaning. Transformer-based architectures provide contextual understanding and semantic representation of conversations. Large language models such as Gemma generate concise summaries and extract action items.

The processing pipeline typically consists of audio acquisition, preprocessing, speech recognition, transcript generation, NLP analysis, summarization, and report generation. This modular architecture enables scalability and maintainability.

Advantages of these technologies include automation, accessibility, and improved user experience. Challenges include computational requirements, dependence on audio quality, and model optimization needs.

V. SYSTEM ARCHITECTURE

The proposed architecture follows a layered design.

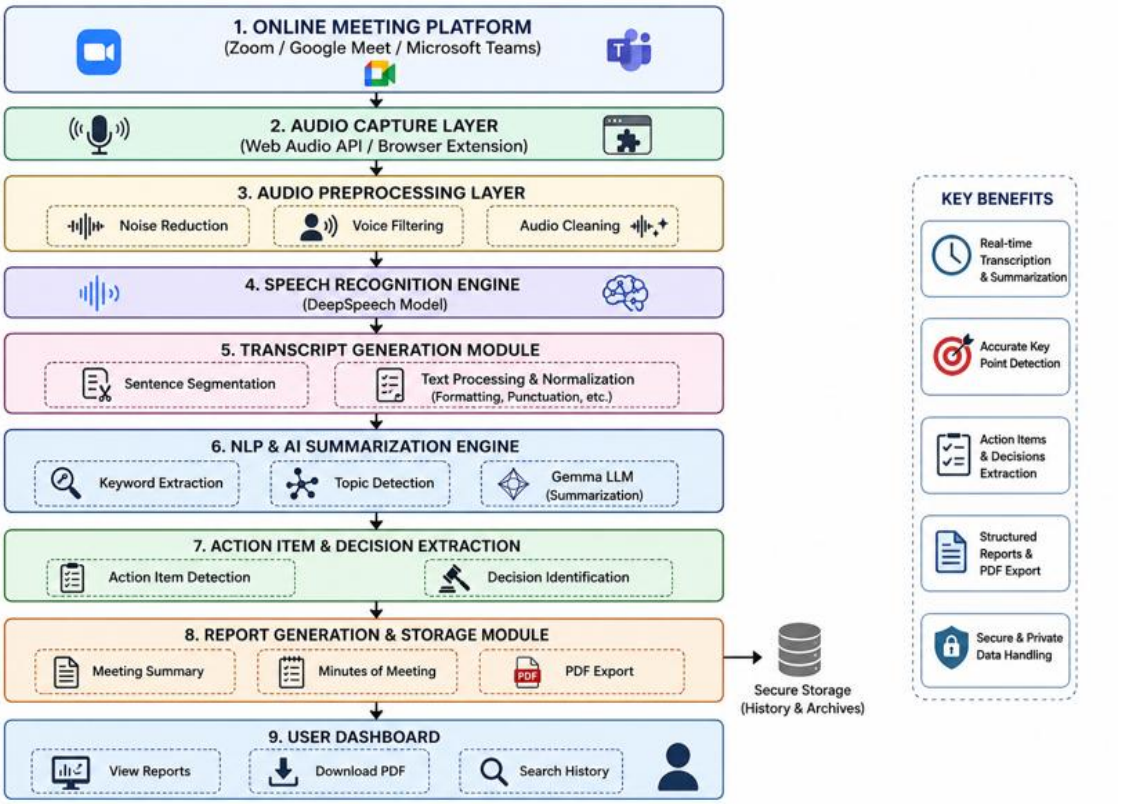

Figure 1. Overall Architecture of AI-Powered Meeting Summary Generator.

Layer 1 consists of browser-based audio capture using Web Audio API.

Layer 2 performs preprocessing including noise reduction and filtering.

Layer 3 executes speech recognition using DeepSpeech.

Layer 4 applies NLP processing and large language model-based summarization.

Layer 5 generates structured reports and downloadable documents.

This architecture separates responsibilities and allows independent optimization of individual modules. The modular approach also simplifies integration with existing collaboration platforms.

VI. METHODOLOGY AND DATA FLOW

The methodology begins with audio acquisition from virtual meeting platforms. Captured audio is filtered to remove background noise and irrelevant signals. The cleaned audio stream is processed by speech recognition models to generate transcripts.

The transcript is then analyzed using NLP techniques. Keywords, entities, discussion topics, and contextual relationships are identified. Summarization models generate concise descriptions of meeting discussions. Action-item extraction modules identify tasks, responsibilities, and deadlines.

Figure 2 Placeholder: Data Flow Diagram.

The final output consists of structured reports containing meeting summaries, decisions, and action items. These reports can be exported as PDF or integrated into collaboration tools.

VII. COMPARATIVE ANALYSIS

Traditional note-taking depends heavily on human effort and is prone to omissions. AI-powered systems provide greater consistency and scalability. Compared

to cloud-only solutions, browser-assisted systems offer stronger privacy controls.

DeepSpeech provides open-source speech recognition capabilities, whereas cloud APIs often achieve higher accuracy at the expense of privacy. Transformer-based summarizers provide superior contextual understanding compared to extractive methods. Gemma-based models offer a balance between performance and deployment flexibility.

The selection of technology depends on organizational requirements including privacy, scalability, cost, and accuracy.

VIII. ADVANTAGES, LIMITATIONS AND RESEARCH GAP

AI-powered meeting summarization systems reduce administrative effort, improve productivity, and provide searchable knowledge repositories. They support better decision tracking and facilitate collaboration. However, limitations remain. Speaker identification remains challenging in multi-participant conversations. Noisy environments can reduce transcription accuracy. Large language models may occasionally generate inaccurate or incomplete summaries.

Current research primarily focuses on transcription accuracy. Less attention has been devoted to privacy-preserving deployment, multilingual meeting support, contextual understanding, and intelligent action-item prediction. These areas represent significant opportunities for future investigation.

IX. 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.

Advanced analytics may provide meeting quality scores, participation insights, and predictive recommendations. Integration with project management platforms will allow automatic creation of tasks and follow-up activities.

The emergence of efficient large language models will further improve accessibility and enable real-time deployment on resource-constrained devices.

X. CONCLUSION

AI-powered online meeting summary generators represent an important advancement in workplace productivity and knowledge management. By combining speech recognition, NLP, and large language models, these systems automate documentation workflows and reduce manual effort. The reviewed literature demonstrates substantial progress while highlighting challenges related to privacy, multilingual processing, and contextual understanding. Future research focusing on multimodal AI, intelligent analytics, and privacy-preserving deployment will further enhance the capabilities of meeting intelligence systems.

Table 1. Literature Survey Comparison

Paper	Year	Technology	Contribution	Gap
Transformer Summarization	2021	Transformers	Abstractive summaries	Limited datasets
QMSum	2021	Transformer Models	Benchmark dataset	Consistency issues
Meeting Summary Generator	2022	SER + NLP	Automated minutes	Audio variability
Action Item Summarization	2024	LLMs	Task extraction	Complexity

Table 2. Technology Comparison

Technology	Purpose	Advantages	Limitations
Web Audio API	Audio Capture	Browser Native	Browser Dependency
DeepSpeech	Speech Recognition	Open Source	Accent Sensitivity
Transformers	Summarization	Context Aware	Resource Intensive
Gemma	LLM Summary	High Quality	Optimization Required

REFERENCES

[1] Schneider, F., Turchi, M., and Waibel, A., "Policies and evaluation for online meeting summarization," arXiv preprint, arXiv:2502.03111, 2025.
[2] Rakshitha, S. R., Naik, S. P., Sanjana, V. S., Suprasanna, V., and Nayana, C. P., "Real-time audio transcription with automated PDF summarization and contextual insights," International Journal of Innovative Science and Research Technology, vol. 9, no. 11, 2024.
[3] Chen, H., et al., "MISP-Meeting: A real-world dataset with multimodal cues for long-form meeting transcription and summarization," 2025.
[4] Arriaga, C., et al., "Evaluation of real-time transcriptions using end-to-end ASR models," 2024.
[5] Raffel, C., 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.
[6] Google Cloud, "Speech-to-Text documentation." [Online].
[7] OpenAI, "GPT models and applications." [Online].
[8] Fireflies.ai, "AI meeting assistant." [Online].
[9] Graves, A., Mohamed, A., and Hinton, G., "Speech recognition with deep recurrent neural networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.
[10] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019.