Auto MoM : Design and Implementation of Automated Minutes of Meeting Generator

Mrs. B. Bhagyasree, Ch. Krishna Chaitanya, D. Deepak, K. Reshwanth, L. Manikumar

Abstract— Meetings are critical communication and decision-making tools in both business and academic institutions, yet the task of manually documenting them into formal Minutes of Meeting (MoM) remains labor-intensive, error-prone, and inefficient. Despite the widespread use of digital meeting platforms such as Zoom and Google Meet, current technologies fail to bridge the productivity gap between conversation and actionable documentation. This project introduces AutoMoM: an AI-powered, end-to-end automated system for recording, transcribing, synthesizing, and formatting MoMs from virtual meetings.

The system leverages a multi-stage AI pipeline that integrates real-time speech recognition using OpenAI Whisper or equivalent transcription models, Natural Language Processing (NLP) based on Large Language Models (e.g., GPT-4) for content summarization, and action-item classification techniques to extract key decisions and stakeholder responsibilities. Through secure OAuth-based integration with platforms like Zoom and Google Meet, meeting recordings are autonomously fetched, processed, and converted into professional, structured MoM documents that can be exported, archived, or shared through email.

I. INTRODUCTION

In every organization, from academic institutions to global corporations, meetings are the critical forum where ideas are born, decisions are made, and progress is initiated. Yet, a fundamental productivity gap exists between these vital spoken conversations and the creation of tangible, actionable documentation. The manual process of recording, transcribing, and summarizing Minutes of Meeting (MoM) is a well-known bottleneck. It is not only labor-intensive and time-consuming but also notoriously prone to human error, personal bias, and critical omissions. This administrative burden forces participants into an impossible choice: either actively engage in the discussion or divert their focus to frantically taking notes, ensuring they cannot do either task perfectly. This chronic inefficiency leads to a cascade of significant organizational problems. Without a reliable record, key decisions become ambiguous, and nuanced context is lost almost immediately. Assigned responsibilities and deadlines become murky, leading to a lack of accountability and stalled projects.

This project, AutoMoM, is an intelligent, end-to-end system designed to solve these exact problems by fully automating the entire meeting documentation workflow. It provides a seamless experience where a user can simply provide a meeting's audio or video recording through a clean web interface. The system then takes over, initiating a sophisticated analysis that transforms the unstructured, complex, and multi-speaker audio into a set of structured, high-value assets. The final output delivered to the user is not just a raw transcript, but a coherent and actionable package: a complete and accurate transcription, a concise summary of the key discussion points, and a clearly extracted list of decisions and assigned action items.

What sets AutoMoM apart is its dedicated focus on overcoming the primary frustrations of current-generation AI tools: speed and accuracy.

To solve for speed, the system is built to perform its most complex and time-consuming analytical tasks—figuring out *what* was said and *who* said it—at the same time. This parallel-processing approach is designed to dramatically reduce the waiting period, delivering the final, actionable minutes to the user significantly faster than a traditional, sequential process. To solve for accuracy, AutoMoM introduces a collaborative feature where users can provide a manual "gist" or key context before processing begins. By feeding the system relevant information—such as attendee names, project-specific jargon, or key topics—the user "primes" the AI, enabling it to produce a far more accurate transcription of uncommon terms and a more relevant, focused summary that understands the meeting's true purpose.

The impact of this project is a fundamental shift in how teams and organizations operate, moving them from a culture of recollection to one of record. By eliminating the administrative burden of note-taking, AutoMoM directly saves countless hours of manual labor and, more importantly, liberates every meeting participant to remain fully present and engaged in the conversation. The system's unique adaptations for speed and context-aware accuracy ensure the final output is not just a novelty, but a reliable and genuinely useful tool. This creates a single source of truth that enhances communication clarity, drives accountability, and builds a permanent, searchable, and invaluable archive of institutional knowledge—turning transient conversations into a lasting, structured asset.

II. LITERATURE REVIEW

The paper [1] provides a comprehensive overview of abstractive meeting summarization, highlighting its suitability for multi-party conversations. It details the unique challenges like spontaneous speech and complex dynamics that differentiate it from traditional text summarization. The paper reviews limitations of current tools (training data, models, metrics) and advocates for new approaches specifically designed for the meeting domain.This paper [2] offers a thorough review of dialogue summarization across various domains, including meetings, chat, email, customer service, and medical dialogues. It provides a structured overview of publicly available datasets and covers the recent advances and emerging research directions in the field.

This paper [5] proposes a novel diarization-aware multi-speaker ASR system that integrates speaker diarization with LLM-based transcription. The framework processes structured diarization inputs (speaker embeddings and timestamps) alongside audio features, enabling the LLM to generate accurate, time-aligned, segment-level transcriptions, showing robust performance in complex, high-overlap meeting scenarios.This work [6] introduces a novel multimodal approach to speaker diarization that jointly utilizes audio, visual, and semantic cues. The method structures visual and semantic information into pairwise constraints to guide a semi-supervised clustering of acoustic speaker embeddings, consistently outperforming state-of-the-art methods on spontaneous, multi-party conversations.

This survey discussed in [3] presents a general overview of text summarization with a specific focus on the meeting domain. It covers essential datasets, evaluation metrics, and provides a leaderboard comparing the performance of various summarization models, serving as a practical guide to the field's benchmarks. Another survey conducted according to [4] expands the scope of traditional summarization reviews by introducing a detailed analysis of Large Language Model (LLM)-based techniques. It covers conventional extractive and abstractive methods before diving into the paradigm-flexible approaches enabled by LLMs, highlighting their superior performance in coherence and fluency.

This paper [10] proposes HMNet, a novel abstractive summarization network for meetings. It uses a hierarchical structure to handle long transcripts and incorporates a role vector to distinguish between speakers. Due to the scarcity of meeting data, the model is pre-trained on large-scale news summary data, a technique that significantly improves performance on the AMI and ICSI datasets. Another paper [11] studying on top of the AMI corpora introduces an extractive summarization system that leverages discourse structure to identify salient information. Using discourse graphs to represent semantic relations between utterances, a GNN-based node

classification model selects the most important utterances for the summary. The approach surpasses existing text-based and graph-based systems on the AMI and ICSI corpora.

This paper [12] presents BART, a denoising autoencoder for pre-training sequence-to-sequence models. BART is trained by corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. It achieves state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks. This paper [14] introduces DialogSum, a large-scale dialogue summarization dataset consisting of real-life scenarios. The dataset is designed to benchmark models on summarizing conversations that are more representative of everyday interactions than existing corpora. The work studies in the paper [15] presents Meetalk, a system that addresses key challenges in meeting summarization, including salience identification and context understanding. It uses a retrieval-augmented approach with a three-stage LLM pipeline and introduces a mechanism for adaptive personalization by learning from user corrections to improve summary quality over time.

This paper [18] discusses a system for automatic meeting minutes generation using NLP. The approach involves converting meeting audio to text, generating summaries using both extractive and abstractive methods (BART, T5), and identifying action items using a pre-trained BERT-based model to improve the efficiency and accuracy of meeting documentation.This paper [19] presents the architecture of the CALO Meeting Assistant, a comprehensive system for meeting analysis. Its components include real-time and offline speech transcription, dialogue act tagging, topic segmentation, and dedicated modules for question-answer pair identification, action item recognition, and decision extraction.

This project discussed in the paper [21] develops an AI-powered system for automatically summarizing online meetings. The methodology involves four stages: transcription, key point extraction, summarization, and action item extraction. For action item extraction, the system employs Named Entity Recognition (NER) using spaCy to identify action-based statements by detecting named entities and direct objects associated with action verbs.

This paper [22] defines a new task, query-based multi-domain meeting summarization, where models must summarize relevant spans of a meeting in response to a query. It introduces QMSum, a new benchmark for this task with 1,808 query-summary pairs over 232 meetings, and evaluates a locate-then-summarize method, revealing significant challenges for future research.

The automated meeting assistant space is currently dominated by two major commercial platforms, Otter.ai and Fireflies.ai, which have validated the market demand for AI-driven summarization. Otter.ai, in particular, has defined the user experience around real-time transcription, providing live captions and collaborative note-taking that allow teams to engage with the transcript as the meeting happens.1 Conversely, Fireflies.ai has established its market leadership through a deep focus on post-meeting workflows; its strengths lie in generating high-quality "super summaries," extracting actionable tasks, and offering a vast ecosystem of project management tools, and communication platforms, both platforms exhibit significant, persistent limitations. The most critical failure reported by users of both tools is poor speaker identification (diarization), where the systems frequently mislabel or confuse speakers, especially in fast-paced conversations. Furthermore, and their primary data-capture method relies on an "intrusive" bot joining the call, which raises privacy and usability concerns. These well-documented gaps create a clear opportunity for a system designed to improve upon these specific failures in accuracy and User Experience.

III. Comparative Analysis

The architecture designed for the AutoMoM project represents a pragmatic and innovative synthesis of the key findings, best practices, and identified gaps within the provided literature. It deliberately builds upon the established academic consensus while architecturally addressing the specific, well-documented failures of current commercial tools. The system's design is not an

attempt to invent a single, monolithic model from scratch; rather, it is a purposeful orchestration of specialized components, with its primary innovations—parallel processing and user-guided context—designed to directly solve the field's most persistent challenges of performance and accuracy.

The foundational design of AutoMoM aligns perfectly with the modern, multi-stage pipeline identified in the literature. Like the CALO Meeting Assistant and other systems, our architecture employs a decoupled, multi-stage approach: a frontend for user interaction, a backend for orchestration, a "Core AI Engine" for processing, and databases for storage. More significantly, it fully embraces the paradigm shift to Large Language Models (LLMs) for the final, high-level tasks of summarization and insight generation. While older, pre-LLM systems focused on extractive methods using discourse graphs or pre-training models like BART , our architecture leverages a state-of-the-art LLM (e.g., GPT-4o) for its superior fluency and coherence in abstractive summarization, a direction strongly advocated by the academic community.

Where the AutoMoM architecture begins to innovate is in its direct response to the "error propagation" problem inherent in traditional cascaded pipelines, a central challenge identified in the literature. Cutting-edge research proposes solving this by creating novel, deeply integrated "diarization-aware" models that process audio and speaker embeddings simultaneously. Our project, in contrast, proposes a *pragmatic architectural solution* rather than a new model. By containerizing the STT Model and the Speaker Diarization Model as separate components, our Backend API can orchestrate them in parallel. This is a critical design choice. While a sequential pipeline (diarization *then* transcription) creates a dependency that propagates error, and a fully integrated model is complex to build, our parallel approach allows the system to compute both *what* was said and *who* said it at the same time. The subsequent "merge step" in the backend is an algorithmic solution that logically combines the two outputs, significantly reducing total processing time and creating a modular system where either the STT or diarization component can be upgraded as technology improves, without re-engineering the entire pipeline.

Furthermore, the AutoMoM architecture introduces a novel solution to the problem of context and accuracy, a theme echoed in papers on hierarchical networks , query-based summarization , and adaptive personalization. While existing commercial tools like Fireflies.ai and Otter.ai are criticized for their poor accuracy with domain-specific jargon or accents, our system introduces a dual-path user-provided context feature. This "gist" provided by the user is not only passed to the LLM to inform a more relevant summary—a concept similar to the goals of Meetalk and QMSum —but is also fed directly into the STT model as a prompt. This first-pass "priming" of the transcription model is a key innovation designed to dramatically improve the base-level accuracy on the exact terms (names, acronyms, technical jargon) that automated systems typically fail to recognize.

Finally, the AutoMoM architecture is a direct response to the documented failures of the current commercial state-of-the-art. The literature review of existing tools identifies their two greatest weaknesses: poor speaker identification (diarization) and an intrusive, "bot-based" ingestion model that raises privacy concerns. Our system addresses both. By building around a modular Core AI Engine containing a dedicated Speaker Diarization Model (e.g., Pyannote), we create an architecture that is not locked into a single, underperforming component. This allows for continuous improvement by swapping in superior open-source models as they become available. Secondly, the system's fundamental workflow—centered on a user uploading a file via a secure web application—is inherently "bot-free," aligning perfectly with the market's clear demand for a more private, less obtrusive, and user-controlled alternative to existing solutions.

IV. CONCLUSION

This comparative study began by examining the foundational challenges of automated meeting summarization, a field that, despite its clear value, remains hindered by significant technical hurdles. The literature review confirmed that the core problem is not merely transcribing words but understanding the chaotic, spontaneous, and multi-party dynamics of human conversation. Our analysis of the academic landscape revealed a clear consensus: the field has matured from simple

extractive methods to sophisticated abstractive approaches powered by Large Language Models (LLMs). However, this same literature identified the field's primary unsolved problem: the fragility of cascaded processing pipelines, where failures in an early stage—particularly the notoriously difficult task of speaker diarization in high-overlap scenarios—irreversibly corrupt all downstream tasks, a problem known as "error propagation".

Our subsequent analysis of the commercial state-of-the-art demonstrated that this academic challenge is not theoretical; it is the single most significant failure point of today's market-leading products. Platforms such as Fireflies.ai and Otter.ai, while validating the immense market demand for this technology, are subject to consistent user complaints of poor speaker identification, which leads to inaccurate, confusing, and unreliable meeting minutes. Furthermore, this market review identified a second major gap in user experience: a reliance on intrusive "bot-based" ingestion models that raise significant privacy and usability concerns among users. These findings established the clear objectives for our project: to design an architecture that specifically solves the dual problems of diarization-based error propagation and the poor user experience of existing tools.

In response to these findings, we designed the AutoMoM system architecture, a pragmatic and innovative blueprint that synthesizes the strengths of the academic literature while directly addressing the failures of the commercial landscape. Our architecture adopts the modular, multi-stage pipeline advocated by numerous research papers—decoupling the `Core AI Engine` components of Speech-to-Text (STT), Speaker Diarization, and NLU (LLM) from the main application logic. This C4-style design moves beyond the limitations of its predecessors by incorporating two novel architectural adaptations that serve as the core of our contribution.

The first innovation is the implementation of an orchestrated parallel processing pipeline. While cutting-edge research attempts to solve error propagation by building highly complex, deeply integrated "diarization-aware" models, our architecture proposes a more robust and modular algorithmic solution. The `Backend API` initiates both transcription and speaker diarization tasks simultaneously. This parallel approach not only cuts the total processing time significantly but, more importantly, it breaks the dependent link in the cascaded pipeline. By merging the two independent outputs in a final "alignment step," the system prevents a diarization error from fundamentally corrupting the transcription process itself, thereby solving the core problem of error propagation at the architectural level.

The second innovation, the user-provided "context gist," is a direct solution to the accuracy and relevance failures endemic to current tools. By allowing a user to "prime" the system with project-specific jargon, attendee names, and key topics, we create a dual-path benefit. First, this context is fed to the STT model, dramatically improving its transcription accuracy on the exact domain-specific terms that generic, pre-trained models typically fail to recognize. Second, this same context is passed to the LLM during the summarization phase, ensuring the final output is not just fluent, but contextually aware and focused on the topics the user actually cares about—a practical implementation of the concepts explored in query-based and personalized summarization. Combined with an inherently "bot-free" file upload workflow, this architecture results in a system that is faster, more accurate, and more private than current solutions.

V. FUTURE SCOPE

The architecture presented in this paper for the AutoMoM system provides a robust and innovative foundation for automated meeting documentation. Its novel approaches to parallel processing and user-guided context solve key issues of performance and accuracy. This foundation opens several significant avenues for future research, development, and commercial application.

A primary avenue for expansion lies in evolving the system's ingestion model from an asynchronous, file-upload-based workflow to a real-time, end-to-end platform. This can be pursued via two distinct architectural paths. The first is a privacy-first, client-side application that would function as a desktop agent or browser extension, similar to tools like Tactiq and Granola. This "bot-free" model would perform transcription and diarization locally on the user's machine, appealing to individuals and organizations with

strict data sovereignty requirements. The second path is a full SaaS-based ingestion model, which would utilize a server-side bot (like those in Fireflies.ai or Otter.ai) to join meetings automatically. This would enable a fully managed, scalable solution for enterprise clients, seamlessly integrating with their calendars and centralizing meeting intelligence.

A second, crucial enhancement is the development of a human-in-the-loop (HITL) fine-tuning pipeline. The current architecture allows users to *provide* context, but the next logical step is to *learn* from their corrections. This would involve building a feedback mechanism into the frontend where users can easily correct speaker labels, fix transcription errors, or edit the AI-generated summaries. These verified corrections would be captured by the backend and used to create a growing, high-quality dataset. This dataset would then be used to iteratively fine-tune the Core AI Engine models, creating a system that becomes progressively more accurate and personalized to a specific user's or team's vocabulary, accents, and summarization preferences, similar to the adaptive personalization proposed in advanced systems.

To achieve enterprise-grade scale, the current backend orchestration logic could be re-architected into a fully event-driven architecture (EDA). In this more advanced model, the file upload would trigger an event (e.g., File_Uploaded) that is published to a message bus. Each component of the Core AI Engine would be a separate, containerized microservice that subscribes to these events, processes its specific task (e.g., transcription), and then publishes a new event (e.g., Transcription_Complete). This asynchronous, decoupled approach would provide immense scalability, fault tolerance, and higher throughput, as each microservice could be scaled and updated independently, ensuring the system can handle thousands of concurrent processing jobs efficiently.

A fourth area of expansion is to evolve AutoMoM from a "system of record" into a "system of action" through deep workflow integration. While the current system produces a summary and action items, future work should focus on making these outputs directly actionable. This involves building a robust integration layer with project management tools like Jira and Confluence. This would allow a user to, for example, click an action item within the AutoMoM interface ("Fix the login bug") and have the system automatically generate a pre-filled Jira ticket, complete with a description, relevant transcript snippets, and a link back to the meeting. This would seamlessly bridge the gap between conversation and project execution.

VI. REFERENCES

[1] Virgile Rennard, Guokan Shang, Julie Hunter, Michalis Vazirgiannis; Abstractive Meeting Summarization: A Survey. Transactions of the Association for Computational Linguistics 2023; 11 861–884. doi: https://doi.org/10.1162/tacl_a_00578

[2] Xiachong Feng , Xiaocheng Feng* , Bing Qin Harbin Institute of Technology, China {xiachongfeng, xcfeng, bqin}@ir.hit.edu.cn

[3] Meeting Summarization: A Survey of the State of the Art Lakshmi Prasanna Kumar IMRSV Data Labs., Ottawa, Canada lakshmi@imrsv.ai Arman Kabiri IMRSV Data Labs., Ottawa, Canada arman@imrsv.ai

[4] Yang Zhanga,b, Hanlei Jina,b, Dan Menga,b, Jun Wang1a,b, Jinghua Tana,b aSouthwestern University of Finance and Economics, Chengdu, China bEmail Addresses, wangjun1987@swufe.edu.cn, 595915575@qq.com,

[5] https://doi.org/10.48550/arXiv.2506.05796

[6] [Integrating Audio, Visual, and Semantic Information for Enhanced Multimodal Speaker Diarization on Multi-party Conversation](https://aclanthology.org/2025.acl-long.977/) (Cheng et al., ACL 2025)

[7] [ALLIES: A Speech Corpus for Segmentation, Speaker Diarization, Speech Recognition and Speaker Change Detection](https://aclanthology.org/2024.lrec-main.67/) (Tahon et al., LREC-COLING 2024)

[8] [SiTa - Sinhala and Tamil Speaker Diarization Dataset in the Wild](https://aclanthology.org/2025.chipsal-1.8/) (Thayasivam et al., CHiPSAL 2025)

[9] Playing with Voices: Tabletop Role-Playing Game Recordings as a Diarization Challenge Lian Remme Heinrich HeineUniversity Düsseldorf lian.remme@uni-duesseldorf.de Kevin Tang Heinrich Heine University Düsseldorf kevin.tang@uni-duesseldorf.de

[10] https://aclanthology.org/2020.findings-emnlp.19.pdf

[11] Leveraging Discourse Structure for Extractive Meeting Summarization Virgile Rennard1,2 , Guokan Shang3 , Michalis Vazirgiannis2,3 , Julie Hunter1 1LINAGORA, 2École Polytechnique, 3MBZUAI virgile@rennard.org guokan.shang@mbzuai.ac.ae mvazirg@lix.polytechnique.fr jhunter@linagora.com

[12] https://arxiv.org/pdf/2405.11055

[13] https://arxiv.org/abs/2004.05150

[14] https://aclanthology.org/2021.findings-acl.449.pdf

[15] Chen, Zheng & Futian, Jiang & deng, Yue & He, Changyang & Li, Bo. (2025). Meetalk: Retrieval-Augmented and Adaptively Personalized Meeting Summarization with Knowledge Learning from User Corrections. 94-110. 10.18653/v1/2025.knowllm-1.9.

[16] Yoo, C., & Lee, H. (2023). Improving Abstractive Dialogue Summarization Using Keyword

Extraction. Applied Sciences, 13(17), 9771. https://doi.org/10.3390/app13179771
[17] Zichao Yang1 , Diyi Yang1 , Chris Dyer1 , Xiaodong He2 , Alex Smola1 , Eduard Hovy1 1Carnegie Mellon University, 2Microsoft Research, Redmond {zichaoy, diyiy, cdyer, hovy}@cs.cmu.ed
[18] S. Muppidi, J. Kandi, B. S. Kondaka, C. Kethireddy and S. E. Kandregula, "Automatic meeting minutes generation using Natural Language processing," 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT), Bengaluru, India, 2023, pp. 1-7, doi: 10.1109/EASCT59475.2023.10393102. keywords: {Computational modeling;Pipelines;Speech recognition;Organizations;Evolutionary computation;Transformers;Natural language processing;Transcript;speech Recognition;summarization;Key points Extraction;Action items},
[19] Tur, Gokhan & Stolcke, Andreas & Voss, Lynn & Peters, Stanley & Hakkani-Tur, Dilek & Dowding, John & Favre, Benoit & Fernández, Raquel & Frampton, Matthew & Frandsen, Michael & Frederickson, Clint & Graciarena, Martin & Kintzing, Donald & Leveque, Kyle & Mason, Shane & Niekrasz, John & Purver, Matthew & Riedhammer, Korbinian & Shriberg, Elizabeth & Yang, Fan. (2010). The CALO meeting assistant system. Audio, Speech, and Language Processing, IEEE Transactions on. 18. 1601 - 1611. 10.1109/TASL.2009.2038810.
[20] https://www.cle.org.pk/Publication/papers/2025/Leveraging_LLMs_for_action_item_identification_in_Urdu_meetings_Dataset_creation_and_comparative_analysis.pdf
[21] https://arxiv.org/abs/2104.05938