Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
April 23, 2026
A Regional Conference of the Society for Industrial and Systems Engineering

Generating Video and Audio Meeting Summaries Using AI

Ian Dooley, Cameron Halligan, Samuel Hong, Daniel Kim, Khalil Mayweather, and Matthew Wolfe

Department of Systems Engineering, United States Military Academy, West Point, New York 10996

Corresponding author's Email: matthew.wolfe@westpoint.edu

Author Note: The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: The exponential growth of long-form audio content presents a significant challenge to knowledge integration within high-tempo organizations, especially the U.S. military. To minimize information overload, this long-form audio content must be tailored to provide pertinent information to targeted users. Although approaches for long-form audio content summarization and filtering exist, no existing approaches effectively transform long-form audio into summaries tailored to targeted users. Our design and evaluation integrate Pyannote-based speaker diarization, Whisper-based automatic speech recognition, and large language models with retrieval-augmented generation to transform full-length VTC recordings into labeled summaries that can be queried and substantially reduce redundant listening while maintaining key discussion threads and decisions. Our architecture supports improved information access, mitigates documented mechanisms of overload and fatigue, and sustains team cohesion. Overall, our research concludes that a diarization–ASR–LLM/RAG pipeline provides a practical and adaptable solution for maintaining shared situational awareness in high-tempo, security-sensitive environments.

Keywords: Information Overload, Speaker Diarization, Automatic Speech Recognition, Large Language Models, Retrieval-Augmented Generation

1. Introduction

The Department of War's (DoW's) ability to protect national security increasingly depends on its capacity to process and analyze vast amounts of information from unclassified and classified sources of the military. Information is distributed across different levels of the military organization using differing methods – the most important being long-form audio content. Organizing long-form audio content into targeted summaries enhances its usability, enabling various users to draw information strictly pertinent to their area of expertise, thus improving information dissemination and operational efficiency.

Information overload from long-form audio content is a complex, multi-tiered problem. This leads to issues on the properties of the information itself (volume, frequency, quality, complexity), the individual or individuals processing the information, the tasks at hand and the processes used to complete them, organizational routines, and the information and communication technology (ICT) in use (Arnold et al., 2023). Consequently, users are less likely to dedicate their full attention to and thoroughly process information pertinent to them contained in long-form audio content. As VTC communication becomes more common in workplaces, reports of VTC fatigue at the physical, cognitive, emotional, and social levels rise (Li et al., 2024). Applied to leadership in a military context, where commanders must rapidly digest increasingly complex information to make time-sensitive decisions, these findings justify an automated meeting-summarization approach that saves time, improves the quality of summaries, and prioritizes relevance to users. A meeting pipeline that captures discussions once and produces concise, targeted inputs would allow absent or time-constrained personnel to effectively "attend" after the fact, maintain shared context, and contribute to team cohesion without being forced to consume full-length recordings or sift through unstructured artifacts. This pipeline culminates in the diarization–ASR–LLM workflow, which addresses the challenges to rapid, quality decision-making for military leadership.

In our research, we systematically iterate and test the diarization capabilities of Pyannote, transcription capabilities of Whisper ASR, and summarization capabilities of the open-source LLM GPT-OSS. We hypothesize that a patchwork pipeline using Pyannote, Whisper, and GPT-OSS will effectively reduce information overload for users by providing targeted summaries. To our knowledge, no solution exists that has effectively and sustainably transformed long-form audio into summarized information tailored to targeted users. Our findings suggest that our pipeline could reduce information overload and enhance information processing and retainment by tailoring information to targeted users.

2. Methodology

2.1 Workflow

Our meeting summarization tool uses the following workflow: file upload, audio extraction, speaker segmentation, transcription, external context retrieval, and summary generation.

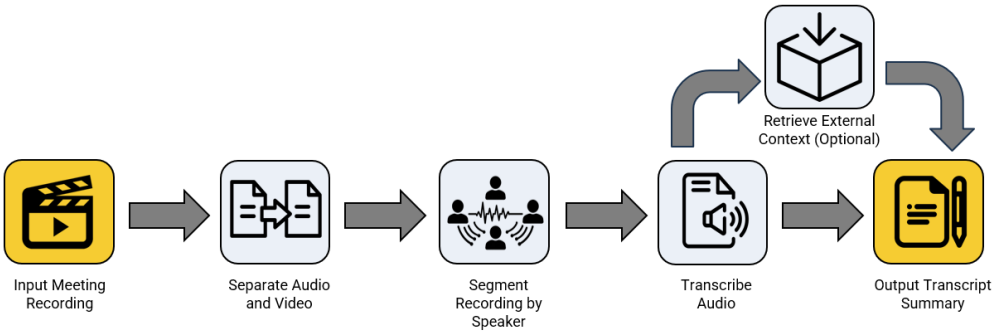

Figure 1. Design methodology for refining and tailoring summarized information for targeted users.

2.2 Input Meeting Recording

In preparation for processing variable long-form audio and visual media, we established an intuitive user interface (UI) which facilitates the ingestion of long-form audio. An intuitive and interactive UI is vital to the pipeline as it reduces cognitive load, increases user adoption, saves time while increasing efficiency, and drives engagement and retention. As users are expected to interact with information that is innately complex and DOW-specific in terminology and content, any means of inputting and interacting with the information must be simple and navigable.

2.3 Separate Audio and Video

The audio track is programmatically separated from the video container before analysis to streamline downstream processing. Although visual cues (e.g., slides, gestures) may provide contextual value, the primary informational signal for summarization lies in spoken dialogue. Isolating the audio channel enables structured preprocessing, including normalization, noise handling, and segmentation. This separation step reduces computational complexity by preventing unnecessary multimodal processing and ensuring compatibility with the ASR and speaker diarization pipeline. Given that spontaneous multi-speaker dialogue introduces challenges such as overlapping speech and background noise, isolating audio ensures that preprocessing steps are optimized specifically for speech-based modeling. The resulting waveform becomes the standardized input for diarization and transcription modules, facilitating consistent tokenization and speaker alignment.

2.4 Segment Recording by Speaker

We implemented speaker diarization to determine “who spoke when” to preserve conversational structure and ensure speaker accountability. Diarization preserves the structure of interaction and distinguishes participant contributions. Accurate diarization enables segmentation of continuous recordings into labeled speaker turns, which is essential for structured summarization and speaker-level analysis. We employed a Pyannote-based diarization pipeline due to its modular architecture integrating voice activity detection, speaker embedding, and clustering. Its SincNet feature extractor and LSTM-based speaker-change detection allow for robust segmentation even in noisy, multi-speaker conditions. Diarization performance is evaluated using Diarization Error Rate (DER), which measures false alarms, missed speech, and speaker confusion (Desai et al., 2024).

Segmenting recordings by speaker supports downstream analytics by preserving turn-level continuity, identifying dominant speakers, quantifying speaking time, and mapping dialogue dynamics. Transforming continuous audio into labeled conversational units creates structured, analyzable data streams that improve transparency, interpretability, operational utility in military coordination settings and more.

2.5 Transcribe Audio

Automatic Speech Recognition (ASR) enables real-time transcription and contextual insight generation. ASR systems such as Whisper are designed for robust zero-shot transcription, meaning they generalize well to new audio types, languages, and environments without retraining. To convert segmented speech into structured text suitable for summarization and retrieval, we transcribed each diarized audio segment using Whisper, a Transformer-based encoder–decoder ASR system trained on large-scale multilingual audio data. Whisper was selected for its robustness to accents, environmental noise, and domain variability, eliminating the need for dataset-specific fine-tuning. Its unified architecture integrates transcription, translation, voice activity detection, and alignment into a single model, simplifying the processing pipeline while maintaining high accuracy (Radford et al., 2022).

By performing diarization prior to transcription, we reduced cross-speaker interference and improved alignment between speaker identity and textual output. The transcription process preserves timestamps and speaker labels, maintaining conversational continuity. The resulting structured transcript serves as the canonical textual representation of the meeting and provides the foundation for targeted summarization and retrieval-augmented generation.

2.6 Output Transcript Summary

To reduce information overload while preserving operationally relevant content, we generated targeted transcript summaries using OpenAI's gpt-oss-20b, an LLM, integrated with RAG. Rather than producing a single general synopsis, the system generates summaries conditioned on defined information needs, consistent with aspect-based and query-focused summarization approaches. The pipeline first retrieves transcript segments relevant to a user-defined query or aspect, then conditions the LLM on those retrieved passages to produce a focused output. RAG improves contextual grounding and reduces hallucination risk by constraining the model's attention to semantically relevant evidence. It also mitigates the computational burden associated with long-context processing by redistributing workload toward indexing and retrieval infrastructure rather than extended attention windows (Arslan et al., n.d.; Oche et al., 2025).

Outputs are assessed along dimensions including relevance, coherence, factual consistency, and adherence to the requested aspect, ensuring summaries remain focused and decision relevant. This workflow transforms continuous VTC recordings into structured, queryable summaries that preserve speaker attribution and key decision threads while substantially reducing redundant listening time.

3. Results

The completed pipeline successfully executed all four intended stages—file conversion, speaker diarization, transcription, and summarization—and produced both intermediate and final outputs. In operation, the system converted uploaded media into a standardized audio representation, segmented the recording by speaker, generated speaker-labeled transcript output, and produced a final meeting summary.

The user-facing workflow was successfully implemented through two functional interfaces. The Streamlit interface (Figure 2) accepted uploaded meeting files and executed the summarization pipeline through a single front end, allowing users to move from raw input media to processed outputs within one accessible workflow. The interface provides a visible processing view as the input media moves through the pipeline stages. In addition, the LightRAG WebUI was fully functional and accepted uploaded external reference documents for use during summarization. Together, these interfaces operationalized the workflow described in the methodology.

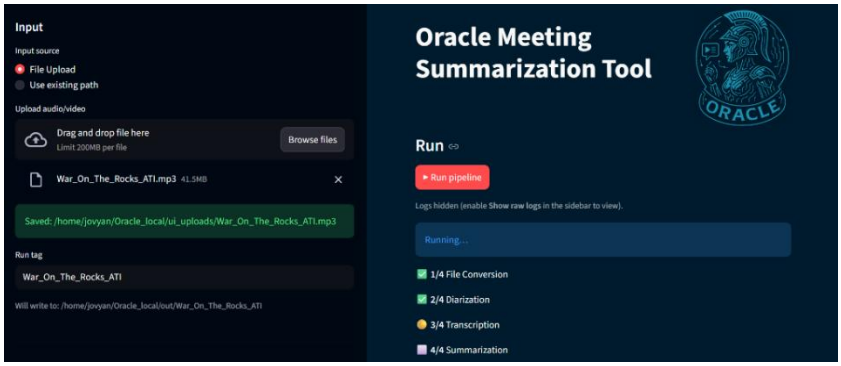

Figure 2. Screenshot of the User Interface processing the input media through the pipeline

Beyond transcript-only summarization, the system also demonstrated RAG as a working capability. When users chose to upload supporting documents through the LightRAG interface, those materials were incorporated into the summarization process and contributed relevant external context to the final summary output. When users did not provide supporting documents, the system still functioned as a standalone meeting summarization pipeline, producing a final summary from the uploaded meeting file alone. This made RAG an available enhancement rather than a required dependency, preserving flexibility in how the tool could be used across different meeting contexts.

An important characteristic of the models and tools (Table 1) used in this architecture is that every component was selected from open-source software or open-weight models. Additionally, each component can be hosted and executed locally rather than relying on persistent external cloud services. As a result, the overall system is well suited for local deployment on a wide range of machines, making the pipeline more portable, easier to sustain, and likely more compatible with security-sensitive environments where local processing may be preferred.

Table 1. Model used by pipeline function

Function	Role in Figure 1	Model	Source
File Conversion	Separate Audio and Video	FFmpeg	Python Package
Diarization	Segment Recording by Speaker	Pyannote 3.1	Hugging Face
Transcription	Transcribe Audio	Whisper-V3 Large	Hugging Face
RAG	Retrieve External Context	LightRAG	GitHub
Summarization	Output Transcript Summary	gpt-oss-20b	Ollama

Our summarization tool was run using ACI WIRE computing. We selected the following server options: 32x CPUs, 256 GB of RAM, and 1x Nvidia RTX Ada 6000 GPU. The GPU was critical for efficient processing, and the Nvidia RTX Ada 6000 has the following specifications: 48GB GDDR6 ECC memory, 18,176 CUDA Cores, 568 Tensor Cores, and 142 RT Cores (Khan, 2025). Processing a 36-minute sample recording was finished in under 10 minutes across all 10 trials (Table 2).

Table 2. Average time to process a 36-minute video recording across 10 runs (mm:ss).

File Conversion	Diarization	Transcription	Summarization (with RAG)
<1 minute	<1 minute	5:58	2:12

To evaluate the performance of our meeting summarization tool, the same *Inside West Point: Ideas That Impact* podcast episode was summarized using both Microsoft Copilot and our tool with the same prompt (Table 3). The CoPilot summary was produced by Spoerl et al (2026). To produce the podcast summary, our tool ingested the podcast recording and performed the workflow depicted in Figure 1. This comparison was done to evaluate how our tool's end-to-end summarization capability performed against Copilot, a frontier model.

Table 3. CoPilot and Oracle podcast summarization outputs.

CoPilot Summary	Oracle Meeting Summarization Tool
Under relentless pressure, leaders either adapt or break. Coach K reveals that winning isn't about talent alone—it's about preparation, trust, and adaptability. From West Point cadet to Olympic champion, his lessons demand leaders embrace failure, improvise under uncertainty, and build "speed to trust"—because warfighting and leadership share the same unforgiving terrain. Three Key Insights: • Look for environments that force failure —"failure is not our destination"—to build resilience and agility. • Strive for adaptability and honest communication; change how you lead"15, 20 times" as conditions shift. • Gain the "speed to trust" advantage—rapid cohesion that accelerates mission success under pressure. Listen to Coach K as he gives insights on preparation, adaptability, and trust—critical for leaders facing complex, high-risk missions.	Coach Mike Krzyzewski reminds Army officers that victories hinge on relentless preparation, a "no excuses" culture, and rapid trust among teammates. From West Point's cadet courts to Olympic gold, his story shows how disciplined habits, character driven values, and adaptable leadership turn pressure into mission critical performance for the entire warfighting team. Three Key Insights • Look for complacent mindsets that mask stagnation and dull critical self-assessment in leaders. • Strive for relentless preparation—merge the will to prepare and the will to win through practice. • Gain rapid team cohesion via speed to trust, ensuring swift, decisive action in high-pressure scenarios. Listen to Coach K's insights and forge resilient, prepared leaders ready for battlefield challenge in today's joint warfighting.

ISBN: 97819384962-5-7

4. Discussion

4.1 Architectural Implications for Information Overload

A central implication of this study is that information overload in VTC environments is fundamentally an architectural problem rather than solely a content-volume problem. Traditional mitigation strategies focus on reducing meeting frequency or manually summarizing content after the fact. However, these approaches do not address the structural inefficiencies that require individuals to repeatedly process the same raw material. Our pipeline shifts the burden of synthesis from the individual to the infrastructure. By capturing discussion once and transforming it into structured, queryable outputs, the system reduces redundant listening while preserving shared situational awareness.

This architectural shift reframes summarization as a workflow transformation rather than a compression task. Instead of shortening content indiscriminately, the system reorganizes conversational data into indexed, labeled knowledge objects that can be retrieved on demand. In high-tempo military environments, where commanders must digest complex and time-sensitive information, this restructuring supports faster orientation without sacrificing decision context. The contribution, therefore, is not merely improved summarization accuracy, but a scalable mechanism for sustaining cohesion and clarity under conditions of persistent information flow.

4.2 Speaker-Aware Summarization as a Structural Advantage

Accurate transcription alone is insufficient for operationally meaningful summaries; preserving speaker attribution is essential for accountability, authority recognition, and interpretability. Many summarization systems collapse dialogue into speaker-agnostic text, which may obscure command intent, dilute responsibility, or misrepresent conversational dynamics. By integrating Pyannote-based diarization prior to transcription and summarization, our pipeline preserves "who spoke when" continuity, enabling summaries that reflect both content and structure.

This structural preservation introduces analytical capabilities beyond summarization. Speaker-level segmentation enables identification of dominant contributors, quantification of speaking time distribution, and mapping of dialogue flow. In command settings, such visibility supports after-action review, leadership assessment, and engagement monitoring. The diarization component therefore functions not merely as a preprocessing step, but as a structural safeguard that maintains the integrity of the conversational record. In this sense, speaker-aware summarization is qualitatively different from transcript summarization: it transforms meetings into analyzable interaction networks rather than flattened text artifacts.

4.3 Targeted Retrieval-Augmented Summarization vs. Long-Context Modeling

Another important implication concerns the tradeoff between long-context LLM architectures and RAG. Expanding context windows in large language models enables broader continuity but introduces quadratic computational scaling and increased GPU memory demands. As sequence length grows, inference latency and infrastructure costs escalate proportionally. In contrast, RAG systems redistribute computational load toward retrieval and indexing infrastructure, narrowing the model's attention to semantically relevant transcript segments before generation.

By conditioning summaries on retrieved passages aligned with a defined query or aspect, the system improves contextual grounding while reducing hallucination risk. This targeted mechanism also aligns directly with documented overload mitigation strategies: filtering and prioritization. Rather than generating a single generic synopsis, the model produces role-specific or query-specific outputs that reduce irrelevance rather than simply reducing length. From an operational perspective, this makes RAG more scalable and adaptable for sustained deployment in environments with continuous VTC streams. While retrieval quality becomes a critical bottleneck, the architectural efficiency and modular update capability of RAG provide a more practical solution for long-duration, high-volume coordination settings than reliance on ever-expanding context windows.

4.4 Limitations

While the pipeline shows strong potential for reducing information overload and extracting structured knowledge from VTC recordings, several limitations remain. Performance is sensitive to real-world audio variability such as overlapping speech, poor microphone quality, background noise, and bandwidth compression. These factors can reduce diarization and transcription accuracy, and errors in DER or Word Error Rate (WER) may propagate downstream, affecting speaker attribution and semantic interpretation.

Additionally, retrieval-augmented generation depends heavily on retrieval quality and supporting infrastructure. If important transcript segments are not retrieved, summaries may miss key information or emphasize irrelevant content.

Although RAG reduces the computational burden of long-context LLMs, it still requires resources for embedding generation, storage, and index maintenance, which may limit real-time deployment in constrained environments. The current audio-only approach also excludes multimodal signals such as visuals, chat logs, and nonverbal cues, suggesting the need for improved retrieval methods and multimodal integration.

5. Conclusion

Our pipeline transforms long-form VTC recordings into structured, speaker-aware summaries that efficiently present key information to relevant users. These summaries, through their ability to be queried, sustain shared situational awareness across organizations to improve decision support in high-tempo, secure environments. Although the inputs to our model currently only consist of audio recordings, the scope of our tool can be extended to other forms of media, operational contexts, and new use cases. Our pipeline provides a strong foundation for scalable knowledge management for environments reliant on continuous information flow across numerous echelons.

5.1 Future Work

Future work includes transitioning the pipeline into a classified network environment and obtaining the necessary cybersecurity approvals for operational use. The system should also undergo rigorous operational testing with classified, mission-representative recordings to evaluate overall performance, reliability, and decision-support value in real command environments.

5.2 Acknowledgements

The authors used ChatGPT to assist with the writing and editing of this article. ChatGPT offered recommendations that were incorporated into the final manuscript.

6. References

Arnold, M., Goldschmitt, M., & Rigotti, T. (2023). Dealing with information overload: A comprehensive review. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1122200

Arslan, M., Munawar, S., & Cruz, C. (n.d.). Business insights using RAG–LLMs: A review and case study. *Journal of Decision Systems*, 0(0), 1–30. https://doi.org/10.1080/12460125.2024.2410040

Desai, A., Kartik, N. V. J. K., Gupta, P., Vinayak, T. S. S., Vanahalli, M. K., & Rajendran, R. (2024). Advancing speaker diarization with Whisper speech recognition for different learning environments. In *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*. https://doi.org/10.1109/TALE62452.2024.10834319

Khan, A. A. (2025, September 15). *NVIDIA RTX 6000 ADA 48GB GPU: Specs, performance & enterprise use cases*. Network Outlet.

Li, B. J., Zhang, H., & Montag, C. (2024). Too much to process? Exploring the relationships between communication and information overload and videoconference fatigue. *PLOS ONE*, 19(12), e0312376. https://doi.org/10.1371/journal.pone.0312376

Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions (No. arXiv:2507.18910). *arXiv*. https://doi.org/10.48550/arXiv.2507.18910

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision (arXiv:2212.04356). *arXiv*. https://doi.org/10.48550/arXiv.2212.04356

Spoerl, J., Choi, D., Locklear, J., Engler, H., Leith, Z., & Koban, D. (2026). *Assessing the impact of AI-based meeting summarization on information relevance and engagement*. United States Military Academy.