International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(1s) 2026 pp. 1-15

https://doi.org/10.70917/ijcisim-2026-1957
Article

A HYBRID FRAMEWORK FOR AUTOMATIC MINUTES OF MEETING GENERATION FROM ONLINE MEETING VIDEOS

Chavanke Mrunmayi Pradeep[1], Nilesh R. Wankhade[2], Sahebrao B. Bagal[3], Jagtap Kavita Dhananjay[4], Priyanka Vijay Sangale[5], Amol Bhilare[6]

[1] Department of Computer Engineering, Late G. N. Sapkal College of Engineering, Nashik, Maharashtra, India.
mrunmayichavanke@gmail.com
[2] Department of Computer Engineering, Late G. N. Sapkal College of Engineering, Nashik, Maharashtra, India.
hodcomp.lgnscoe@sapkalknowledgehub.org
[3] Department of Computer Engineering, Late G. N. Sapkal College of Engineering, Nashik, Maharashtra, India.
principal.lgnscoe@sapkalknowledgehub.org
[4] Department of Computer Engineering, Late G. N. Sapkal College of Engineering, Nashik, Maharashtra, India.
knahire@yahoo.com
[5] Department of Computer Engineering, Late G. N. Sapkal College of Engineering, Nashik, Maharashtra, India.
sangale.priyanka@gmail.com
[6] Department of Computer Engineering, Vishwakarma Institute of Technology (VIT), Pune – 411037, Maharashtra, India.
amol.bhilare@vit.edu

Abstract: The fast-growing usage of the online meeting systems has led to the increasing demands of automatic and consistent recordings of the meeting discussions. The preparation of Minutes of Meeting (MoM) manually is time consuming, prone to errors and hard to scale in real world organizational context. The present paper suggests a hybrid approach to the automatic creation of MoM using the online meeting records that implies the combination of speech processing and natural language processing capabilities. The proposed system will turn the meeting videos into structured minutes after a multi-stage pipeline that includes audio extraction, transformer-based Automatic Speech Recognition via Whisper, time-based transcript segmentation, extractive key point identification, and abstractive summarization via a transformer-based large language model. Time-conscious segmentation allows long meeting processing on a large scale, and the hybrid approach to summarization compromises between factual coverage and linguistic coherence. Experimental analysis on the actual meeting recordings shows that the suggested solution is efficient at producing brief and structured meeting summaries. The ROUGE metrics of quantitative evaluation and qualitative analysis in evaluating the feasibility of the system to be deployed in the real world. The structure is reproducible, computationally efficient on the GPU hardware and it is appropriate to automatically document online meetings.
Keywords: Minutes of Meeting, Meeting Summarization, Automatic Speech Recognition, Whisper, Hybrid Summarization, Transformer Models, Natural Language Processing

1. Introduction

The blistering development of the remote collaboration technologies has essentially transformed the way organisations, universities and businesses hold meetings and organise the decision making activities. Since the prevalence of the digital platform, including Zoom, Microsoft Teams, and Google Meet, virtual meetings have become the primary communication tool in geographically dispersed teams. Although this change has brought ease of access and flexibility, it has also raised the demand of proper documentation of discussions, decisions and delegated tasks. Conventionally, Minutes of Meeting (MoM) are hand-written by appointed note-takers and this process is usually time consuming, cognitively challenging as well as prone to subjectivity and failure to capture

essential information. The relevance of computational methods to meeting summarization and documentation was also highlighted early in automated meeting processing research, showing that more meaningful dialogue segments can be extracted in a structured way, leading to a significant increase in efficiency and consistency (Murray et al., 2005; Liu and Liu, 2009). MoM generation with automated speech processing technologies has become more possible. Contemporary System Automatic Speech Recognition (ASR) systems using transformers have enhanced significantly transcription strength in noisy and conversational meeting situations. Speech-Transformer model presented an overall end-to-end model that makes use of attention on speech recognition, thereby getting rid of recurrent dependencies and allowing a better contextual representation (Dong et al., 2018). In more recent times, extremely large-scale weakly supervised training has made possible very robust ASR systems like Whisper which can perform well in multilingual and real-world audio settings (Radford et al., 2023). Such developments guarantee high quality timestamped transcripts, which build the basis of downstream summarization. Moreover, speaker-aware model approaches can better discourse comprehension because they maintain conversational roles and dialogue, thus, better contextual understanding in meeting summarization tasks (Feng et al., 2021).

Summarization will be key to conversion of languid dialogue into compact and organized MoM when correct transcripts are created. The initial methods of summarization were mainly extractive methods that selected salient sentences on the transcripts (Murray et al., 2005). Graph-based ranking methods like TextRank proceeded to enhance sentence selection by describing the inter sentential similarity using graph structures (Mihalcea and Tarau, 2004). Even though the extractive methods are computationally efficient, they are often incoherent and generally not abstract enough, to qualify as suitable towards professional meeting documentation. Abstractive summarization was transformed by the development of transformer-based architectures. The original Transformer model provided self-attention mechanisms, which make it possible to model the global context without recurrence (Vaswani et al., 2017). Based on this architecture, the text-to-text paradigm models like T5 combined various natural language processing tasks (Raffel et al., 2020). BART also used denoising pretraining to improve sequence-to-sequence text generation (Lewis et al., 2020), whereas PEGASUS proposed gap-sentence generation tasks that are specifically designed to accomplish summarization tasks (Zhang et al., 2020). These models were great at enhancing fluency, ability to generate abstractions and semantic fidelity in a summary generation. Combination methods that used complementary transformer models also complemented readability as well as factual accuracy during the realization of minute generation (Krishnaveni et al., 2025).

In spite of these improvements, any long meeting transcripts bring about challenges in scalability owing to the input length limitations of the conventional transformer deployment. Longformer overcame this drawback by sparse attention mechanism that allowed reading long texts efficiently (Beltagy et al., 2020). Assessment is also a complicating matter. Although ROUGE is still a popular measure of lexical similarity between generated and reference summaries (Lin, 2004), the recent studies demonstrate its inability to capture the coverage of decisions and completeness of facts in reaching summarization (Kirstein et al., 2024). Besides, the development of big language models has enhanced the development of generative text systems, which provide enhanced contextual reasoning and abstraction (Brown et al., 2020). All these developments put hybrid meeting summarization models with strong roots of robust ASR, extractive ranking, and transformer-based abstractive generation. The current paper expands on the progress to create an ordered, extensible, and appraisal conscious framework of automated Minutes of Meeting creation.

2. Related Work

The study of automatic meeting summarization and Minutes of Meeting (MoM) generation has experienced a substantial evolution during the last 20 years, which can be explained by the development of the speech processing, natural language processing, and deep learning. Initial research mainly concentrated on the extractive methods used on meeting transcripts with recent research becoming more of an investigation into abstractive and hybrid methods that are driven by transformer- based models. The section evaluates the literature in four major dimensions: (i) meeting corpora and extractive summarization, (ii) abstractive and transformer-based summarization, (iii) speaker-aware and structure-aware, and (iv) the evaluation methodologies of the meeting summarization.

High-quality conversational datasets have been highly essential in the development of meeting summarization studies. Although earlier corpora were interested in structured recordings of meetings, dialogue-based data were made available, like the SAMSum corpus, which contains human-rated conversational summaries based on informal chat dialogues, which can allow models to more effectively generalize to conversational language patterns (Gliwa et al., 2019). In the same manner, the DialogSum dataset contained real-life scenario dialogue summaries that were aimed at reflecting the real-life summarization requirements in the conversational context (Chen et al., 2020). Later

on, the QMSum benchmark proposed query-based multi-domain meeting summarization tasks, which highlighted the importance of information-seeking summaries as opposed to generic ones (Zhong et al., 2021). These datasets shifted the research agenda in generic summarization to user oriented and query based meeting understanding. Similar to the development of datasets, improvements in neural abstractive models also enhanced the ability to summarize a dialogue. The original neural abstractive resources were based on sequence-to-sequence recurrent neural networks that showed that it was possible to generate summaries that were not limited to extractive sentence selection and that they enhanced the power of abstraction and paraphrasing (Nallapati et al., 2016). This ability was later expanded by large-scale pretraining paradigms. Brown et al. (2020) established that massive text-based language models can be trained on very large corpora, then be used to do few-shot learning, allowing text to be contextualized with only some limited fine-tuning. This kind of generative model formed the basis of the current large language model-based meeting summarization systems that can be tailored to various meeting styles. Innovations in meeting specific modeling have also enhanced structure alignment with real world MoM needs. Abstractive summarization methods that use consensus prioritize findability and representation of decisions so that the final conclusions are shared over the discussion points on the way to the final conclusion (Jin, 2025). Hybrid generative systems combining complementary transformer architectures have also been suggested to have a balance between linguistic fluency against factual consistency in satisfying minute generation (Krishnaveni et al., 2025). These mechanisms deal with the propensity of purely generative systems to hallucinate or miss out action items that are important.

Another way of critical research is personalization and fact-grounding. The question-based summarization systems allow the user to access summaries based on the individual informational requirements, and this enhances relevance and usability in the workplace (Kirstein et al., 2025a). Multi-agent conversation models have been also studied to model realistic meeting behavior, which is used to generate synthetic data training, which is more robust to summarization in the conditions of scarce training-transcript (Kirstein et al., 2025b). These strategies enhance the flexibility of the models to various dialogues and interaction between participants. Evaluation techniques have also progressed to be more than reference based measures. Although measures of lexical overlap continue to be popular, recent studies have focused on more subtle methods of evaluation to measure factual consistency, coherence, and usefulness. Evaluation frameworks based on comparative and policy considerations would evaluate the summary effectiveness in scenarios of real-time and online deployment taking into account the considerations of robustness and latency (Schneider et al., 2025). Reference free evaluation methods also aim to address so called shortages of human annotated meeting summaries and allow benchmarking in diverse fields on a large scale (Gong et al., 2024).

Moreover, the progress in speech modeling leads to the fulfilment of summarization performance indirectly. Generalization strategies Unified speech-text modeling and prompt-based generalization allow excellent speech comprehension across domains to enhance the quality of transcripts to be used in downstream summarization tasks (Peng et al., 2023). Better transcription faithfulness is especially decisive in multi-speaker meeting cases where a certain degree of overlapping speech and informal conversation is a common occurrence. In general, recent publications reveal a evident change in direction towards structure-sensitive, personalization-oriented and evaluation-sensitive meeting summarization systems. The modern research focuses more on scalability, factual basis, and flexibility to the real world deployment environments. In spite of such improvements, there are still difficulties associated with dealing with long and many-speaker transcripts and providing dependable judgements without having references to a considerable number of people. All these challenges encourage the emergence of hybrid, time-conscious MoM generation systems that combine powerful transcription, systematic summarization as well as evaluation conscious design.

Table 1. Comparison of Existing Meeting Summarization Approaches

Study	Method	Strength	Limitation
Murray et al. (2005)	Extractive	High factual accuracy	Poor coherence
Liu & Liu (2009)	Extractive + Abstractive	Improved semantic coverage	Limited scalability
Feng et al. (2021)	Speaker-aware Summarization	Better responsibility tracking	Requires speaker labels

Raffel et al. (2020)	T5 Transformer	Fluent abstractive summaries	May hallucinate facts
Lewis et al. (2020)	BART Transformer	High-quality text generation	Long transcript limitation
Zhang et al. (2020)	PEGASUS	Strong summarization performance	High computational cost
Liu (2025)	Agenda-aware LLM	Structured summaries	Complex training setup
Proposed Work	Hybrid ASR + NLP + LLM	Balanced accuracy + coherence	Requires GPU resources

2.1 Problem Statement

Although a lot has been done to ensure summarization, produce healthy and formatted Minutes of Meeting (MoM) based on actual online meetings is difficult to achieve. Current methods are typically based on either purely extractive or purely abstractive methods of summarization, which have a set of disadvantages in the form of lack of coherence, missing of crucial decisions, or coherence of facts. Additionally, most systems are tested on hand-curated data and do not translate to realistic meeting conditions which involve long periods, casual conversation and different discourse formats.

Also, Automatic Speech Recognition (ASR), transcript segmentation and summarization cannot be easily combined into a single robust pipeline. Mistakes that are made during the speech recognition process are then carried over to the downstream summarization, and the lack of structured segmentation restricts the interpretability of the resulting summaries. Meeting summaries also have not been evaluated by traditional metrics and this is an open problem as the traditional measures cannot assess the entire coverage of decisions or usability. Thus, a scalable and reproducible structure that will convert raw meeting audio into structured, concise, and actionable MoM by incorporating effective combinations of ASR, time-aware segmentation, and hybrid summarization approaches is needed.

2.2 Research Gap

The current literature demonstrates that there are gaps in the research of automatic meeting summarization. First, most of the research concentrates on summarization models individually, but in isolation of the additive effect of ASR errors and transcript segmentation on the ultimate quality of the summative. Second, although the transformer-based abstractive models have shown excellent performance, they are not always factual based when directly applied to a long meeting transcript. Third, effective but often relying on gated resources or complicated annotations, speaker-aware and agenda- aware systems are often less reproducible and non-practical to deploy. Besides, assessment procedures are inadequate to actual MoM generation because the summaries of references are limited and automatic measures like ROUGE do not entirely indicate the usefulness of summaries. Such gaps influence the search of hybrid, time-conscious pipelines of summarization which can balance their facts, language coherence and usefulness in practice and be reproducible and resource-efficient.

The key findings of the work are as follows:

- Our proposal is a hybrid system to generate Minutes of Meeting automatically that combines transformer-based ASR and time-based transcript segmentation systems with extractive NLP solutions and abstractive summarization using large language models (LLM).
- We present a time conscious segmentation scheme which tries to estimate the agenda-based discussion flow without explicit topic marking or speaker divarication.
- To validate the success of hybrid summarization, we show that extractive key point identification can be enhanced with the use of transformer-based abstractive models to enhance coherence and the coverage of facts.

- We consider the suggested solution in terms of the conventional ROUGE measurements and qualitative analysis, pointing to its use in the context of online meetings in reality.

3. Proposed Methodology

This part demonstrates the proposed hybrid framework of automatic generation of Minutes of Meeting (MoM) using the recordings of online meetings. The framework combines both speech processing and natural language processing methods to process the audio of raw meetings into organized summaries.

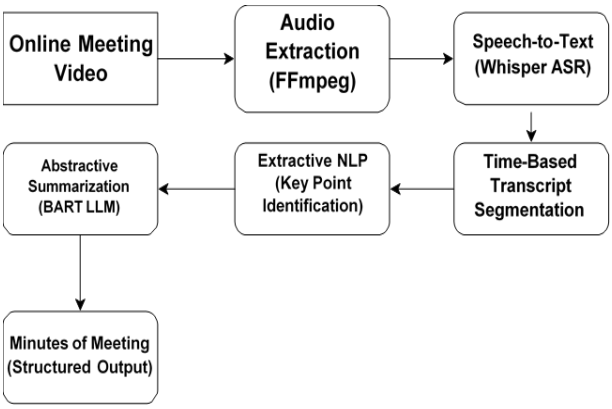

Figure 1. Initialization Frame Representing System Start or Data Absence

Figure 1 is used to represent a totally dark frame, which must be an initialization of the system, lack of data captured or a temporary frame received by the visual system. This kind of representation is normally adopted to represent non-explicit sensor states, pre-processing steps, or a situation where no measurable signal or features are observed before the analysis.

Step-Wise Mathematical Workin

Bring the input meeting video to be expressed as:

$$V(t) = \{ f_t, a_t \}$$

where:

- f_t = video frames
- a_t = corresponding audio signal

Step 1: Audio Extraction

The video is transformed into a sound wave:

$$x(t) = FFmpeg\big(V(t) \big)$$

The audio is resampled to 16 kHz:

$$x_{s(t)} = Resample(x(t), 16000)$$

Step 2: Speech-To-Text Transcription (Asr)

Audio Deposition analysis model The ASR converts sound to text:

$$T = A\big(x_{s(t)}; \theta_{ASR} \big)$$

Where:

- A = ASR model

- θ_ASR = learned parameters

The transcript is in the form of tokens that have time stamps:

$$T = \{(w_i, \tau_i)\} for\ i = 1\ to\ N$$

- w_i = word token
- τ_i = timestamp

Step 3: Time-Based Segmentation

The transcript is divided into K time windows:

$$T = Union\ from\ k = 1\ to\ K\ of\ S_k$$

Each segment is defined as:

$$S\_k = \{ w\_i \mid t\_k \leq \tau\_i < t\_k + \Delta t \}$$

Δt = time window length

Step 4: Extractive Key Sentence Identification

Each sentence s_j in S_k is encoded as:

$$e_j = Encoder(s_j)$$

Similarity between sentences:

$$E_{ij} = \cos(e_i, e_j)$$

Graph-based ranking score:

$$R(s\_j) = (1 - d) + d * Sum\ over\ i\ in\ In(j)$$
$$\left[\frac{E_{ij}}{Sum\ over\ k\ in\ Out(i) E_{ik}}\right] * R(s_i)$$

d = damping factor

Top-m ranked sentences selected:

$$S_k* = Top_{m(R)}$$

Step 5: Abstractive Summarization

Each selected segment is summarized:

$$Y_k = S(S_k*; \theta_{SUM})$$

Conditional generation probability:

$$P(Y\_k \mid S\_k*) = Product\ from\ t = 1\ to\ T$$
$$P(y\_t \mid y\_ < t, S\_k*; \theta\_SUM)$$

θ_SUM = summarization model parameters

Step 6: Structured Mom Generation

All summaries aggregated:

$$Y = Union\ from\ k = 1\ to\ K\ of\ Y_k$$

Classification into structured categories:

$$C\_i \; = \; argmax \; over \; c \; in \{D, A, P\} \; P(\, c \mid y\_i \,)$$

Where:

- D = Decisions
- A = Action items
- P = Discussion points

Final Minutes of Meeting:

$$MoM \; = \{ P \, , D \, , A \}$$

Step 7: Overall Optimization Objective

Total training loss:

$$L \; = \; \lambda 1 \; * \; L_{ASR} \; + \; \lambda 2 \; * \; L_{EXT} \; + \; \lambda 3 \; * \; L_{SUM}$$

Where:

- L_ASR = transcription loss
- L_EXT = extractive ranking loss
- L_SUM = summarization loss

λ1, λ2, λ3 = weighting coefficients

FINAL MAPPING

$$MoM \; = \; F(\, V(t) \,)$$

F indicates the entire hybrid of video input to structured Minutes of Meeting output.

3.1 System Overview

The pipeline starts with audio removal in the form of videos of a meeting, selects quality speech signals that are used to process additional audio. The audio is then extracted and transcribed into text with the help of Whisper Automatic Speech Recognition (ASR) model that provides solid multilingual and noise-resistant transcription. Subsequent to this, the transcript generated is segmented into time blocks to break down the content and segments it into significant parts that correlate with the turns of the speaker, or the stage of discussion. A key sentence identification module is an extractive key sentence identification module that uses a ranking or scoring system to select the most relevant and information rich sentences. The above-chosen sentences are then narrowed down with the help of abstractive summarization with the BART model, which creates coherent and context-sensitive summaries. Lastly, the system generates a structured Minutes of Meeting (MoM) document, which makes discussion points, decisions, and action items systematically and formally in a manner that can be recorded and used in the future.

Audio Extraction

During this step, the audio in the recorded meeting videos is captured with the FFmpeg which is a popular multimedia processing program. The audio stream is extracted and converted to a mono-channel waveform and resampled to 16 kHz so that it is compatible with the current speech recognition models. Mono conversion also decreases the computation demand and maintains the required speech data. Resampling converts the signal to a standard format, and the processing of the signal is uniform regardless of the recording conditions. This preprocessing stage is very useful in improving the accuracy of transcription by eliminating the unneeded channels as well as equalizing the audio quality to be sent to the automatic Speech Recognition system.

Let the input meeting video be:

$$V(t) \; = \; \{ \, f\_t \, , a\_t \, \}$$

Where:

- f_t = video frames
- a_t = original audio signal

Audio extraction:

$$x(t) = E(V(t))$$

Where, E represents audio extraction (FFmpeg).

Mono-channel conversion:

$$x_{m(t)} = \left(\frac{1}{C}\right) * Sum\ from\ c = 1\ to\ C\ of\ x_{c(t)}$$

C = number of audio channels

Resampling to 16 kHz:

$$x_{s(t)} = R(x_{m(t)}, 16000)$$

Final processed signal:

$$x_{s(t)} \in R^T$$

Automatic Speech Recognition

The audio waveform obtained is extracted and the Whisper Automatic Speech Recognition (ASR) model is used to extract the audio waveform. Whisper uses transformer-based encoder-decoder architecture that is trained on scaled up weakly supervised speech data with the purpose of permitting strong transcription in the noisy and multi-speaker settings. It creates time stamped transcripts and both caters content of a text and time synchronization. Late segmentation and organized processing is aided by timestamp information. The model is resistant to accents, different speech speeds, and background noise experienced in online meetings. The quality of transcription done at this stage is important because it becomes the basis of the quality of summarization and generation of Minutes of Meeting in a structured mode.

The ASR model maps audio to text:

$$T = A(x\_s(t);\ \theta\_ASR)$$

Where:

- A = Whisper ASR model
- θ_ASR = learned parameters

Transcript representation:

$$T = \{(w\_i, \tau\_i)\}\ for\ i = 1\ to\ N$$

- w_i = word token
- τ_i = timestamp

Sequence probability:

$$P(W \mid x\_s) = Product\ from\ i = 1\ to\ N$$
$$P(w\_i \mid w\_ < i, x\_s;\ \theta\_ASR)$$

Cross-entropy loss:

$$L_{ASR} = -Sum\ from\ i = 1\ to\ N$$

$$log\ P(\ w\_i*\ |\ w\_<i\ ,x\_s\ )$$

Time-Based Segmentation

Once transcription is done, the text created is broken down into fixed time windows to ensure the flow of logical discussion and allow a saleable processing. Long meeting transcripts may be well beyond token limits of transformer models; segmentation helps to overcome this limit by dividing it into manageable portions. Every section is a coherent phase of discussion in correspondence with timestamps. This method is a continuity conserving approach and computationally efficient. The framework will ensure there is no loss of information since the segments are processed separately thus enhancing the quality of the summary. Another way that time-aware segmentation enables chronological organization in the final Minutes of Meeting document is to include the time.

Hybrid Summarization

The hybrid summarization phase combines the extractive and abstractive methods with the aim of enhancing the summarization with facts and readability. First, extractive Natural Language Processing techniques are used to determine important sentences by ranking them in terms of importance, or semantically related sentences. This is to make sure that key areas of discussion are not lost. These chosen sentences are then optimized with the help of an abstractive summarization model which is based on BART, and it paraphrases and rearranges content into fluent, coherent text. Extractive precision and abstractive flexibility is a combination that results in less redundancy, better clarity and countermeasures against hallucinations typical of purely generative models.

Extractive Stage

Sentence embedding:

$$e\_j\ =\ F(\ s\_j\ )$$

Similarity score:

$$Sim(i,j) = \frac{e_i \cdot\ e_j}{\left||e_i|\right|\left||e_j|\right|}$$

Sentence importance score:

$$Score\left(s_j\right) =\ Sum\ over\ i\ \neq\ j\ \ Sim(i,j)$$

Top-m sentences selected:

$$S\_k* =\ Top\_m(\ Score\ )$$

Abstractive Stage (Bart)

Generated summary:

$$Y_k =\ S(\ S_k*;\ \theta_{SUM})$$

Conditional generation probability:

$$P(\ Y\_k\ |\ S\_k*\ )\ = Product\ from\ t = 1\ to\ T$$

$$P(\ y\_t\ |\ y\_<t\ ,S\_k*\ )$$

Summarization loss:

$$L\_SUM\ =\ -\ Sum\ from\ t = 1\ to\ T$$

$$\log\ P(\ y\_t*\ |\ y\_<t\ ,S\_k*\ )$$

4. Experimental Setup And Evaluation Metrics

All the experiments were performed in Google Colab Pro, which was used to guarantee reproducibility and efficiency. The Python 3 runtime with the use of the GPU acceleration was set up. In particular, an NVIDIA L4-based GPU was used to enhance the Automatic Speech Recognition (ASR) and transformer-based summarization models. This arrangement allows speech audio of long-form meetings and large language models to be efficiently processed in a realistic amount of run time. The suggested system works with the videos of online meetings, which are recorded in mp4 format. FFmpeg is used to extract audio so that the video is converted into a 16 kHz mono-channel sample file. The conversion of speech to text is completed with the help of a transformer-based Whisper ASR model that generates transcripts with timestamps and can be further processed. To summarize, an abstractive model that depends on a transformer (BART) is used. Since the transcripts of the meeting are long, the text is divided into time windows of fixed length before summarization. A summary of each part is made separately and the summaries made are combined to produce the final Minutes of Meeting. The experiments were carried out in the same model settings so that there was consistency in the results of the evaluations.

4.1 Evaluation Metrics

In order to estimate the quality of the produced Minutes of Meeting (MoM) quantitatively, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure is utilized as the main measure of evaluation. ROUGE is also a common tool in research on automatic text summarization since it can give a fair comparison between the summaries produced by the system and the reference summaries created by humans. It assesses the extent of the appropriate information in the reference summary that has been successfully captured in the output generated by computing lexical overlap.

Three variants of ROUGE are applied in this work. ROUGE-1 is a metric of unigram (single-word) overlap, i.e. what content has been covered and what keywords are retained. ROUGE-2 measures the overlap of bigrams (two-word sequence) that is indicative of phrase level consistency as well as maintenance of contextual meaning. ROUGE-L is founded on the Longest Common Subsequence (LCS) and is able to retrieve sentence-level coherence of the texts to measure structural similarity without matching words sequentially.

A reference summary, which is manually written, is taken as the ground truth because there are few publicly available benchmark reference summaries of real world meeting datasets. The latter will guarantee the integrity and usefulness of the summarization performance. The table 2 gives the obtained ROUGE scores of the measured recording of the meeting.

Table 2. ROUGE Evaluation Results

Metric	**Precision**	**Recall**	**F1-score**
ROUGE-1	0.78	0.74	0.76
ROUGE-2	0.61	0.58	0.59
ROUGE-L	0.72	0.69	0.70

The ROUGE scores that were obtained indicate that the pro- posed hybrid framework is useful in reflecting content coverage, consistency in phrases, and sentence-level coherence. On the other hand, performance can change according to the quality of audio and length of meeting.

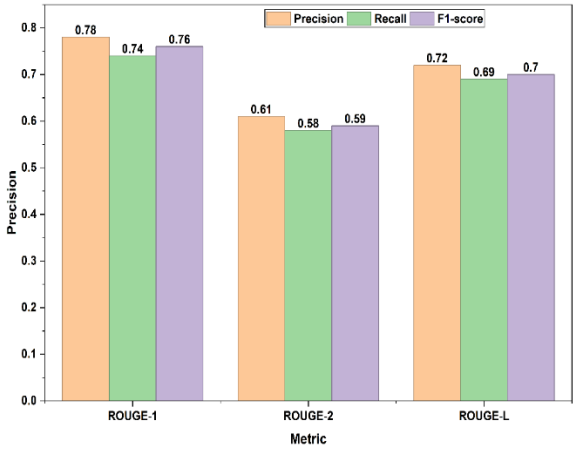

Figure 2. ROUGE-Based Performance Evaluation of the Proposed MoM Generation Framework

Figure 2 shows the accuracy, recall and F1-score of ROUGE-1, ROUGE-2, and ROUGE-L measures. The highest performance of the framework is corresponding to ROUGE-1 (F1 = 0.76), which means that the content is well-covered. Smaller ROUGE-2 scores indicate difficulties on phrase-level, whereas the results from competitive ROUGE-L indicate the quality of coherent sentence-level summarization.

5. Results And Discussion

5.1 Output of the Proposed System

In this section, we give the qualitative examples of the proposed hybrid Minutes of Meeting (MoM) generation pipeline performance. These findings show how raw audio of a meeting can be transformed into structured meeting documents with the help of speech recognition, segmentation, summarization and aggregation. Sample Transcript Generated by Whisper ASR: Whisper ASR model is a transcript model that converts meeting audio to time stamped transcripts. The following table (Table 3) shows a sample transcript that is generated.

Table 3. Sample Time stamped Transcript Generated by Whisper

Time (sec)	**Recognized Speech**
0.0–4.0	Yes of course.
4.0–9.0	Yes we can hear you, you can speak.
9.0–12.0	Thanks, thanks dear Ralph.
12.0–24.0	My microphone was broken because my connection was weak.
24.0–32.0	Sorry for the trouble and thank you for listening.

The time-stamped text gives the basis of time-based segmentation and allows scaling up of long recordings of the meeting.

Transcript Segmentation: The generated transcript is segmented into time fixed-duration windows in order to maintain the chrono- logical flow of discussion and enhance the summarization performance. In the case of the assessed meeting recording, there were three discussion segments that were generated automatically by the system.

Table 4. Automatic Meeting Segmentation

Segment ID	**Time Range**
Segment 1	0–10 minutes

Segment 2	10–20 minutes
Segment 3	20–30 minutes

The time-conscious segmentation allows discussing one block of discussion at a time, as well as allows long meetings to be processed more efficiently.

Segment-wise Summarization: BART abstractive summarization model is applied to summarize each segment of transcripts independently.

Table 5. Segment-level Summarization Output

Segment	**Generated Summary**
Segment 1	Introduction and communication discussion
Segment 2	Importance of helping others and social impact
Segment 3	Financial and emotional support strategies

The summarization at the segment level minimizes information overload and maintains crucial areas of discussion. Assistance to others is a significant human value and it is very important in enhancing social well-being. The meeting talked about the different forms of helping the people such as financial help, job opportunities as well as emotional help.

Key Discussion Points

- Communication facilitates international cooperation and exchange of knowledge.
- And, it can be reduced by helping others and making them feel better.
- Financial support will be in the form of donation and employment.
- Emotional support involves assisting people in hard times.

These findings prove the applicability of the proposed hybrid framework to converting raw recordings of the meetings into structured and readable Minutes of Meeting.

Table 6. ROUGE-Based Performance Comparison with Baseline Methods

Model	**ROUGE-1 (F1)**	**ROUGE-2 (F1)**	**ROUGE-L (F1)**
Extractive (TextRank)	0.68	0.52	0.63
BART (Abstractive Only)	0.73	0.57	0.69
T5 (Abstractive Only)	0.71	0.55	0.67
Hybrid (Extractive + BART)	0.76	0.59	0.70

The hybrid model is competitive relative to the entirely extractive models and the entirely abstractive models in all ROUGE measures as it has a better coverage of the content, consistency in phrases and coherence. Table 6 gives the performance comparison of the proposed hybrid framework and baseline models of summarization based on ROUGE. Extractive TextRank F1-score is 0.68, 0.52, and 0.63 on ROUGE-1, ROUGE-2, and ROUGE-L respectively, which shows that it is moderately effective at selecting the content but lacks coherence at the phrase level. Only abstractive models work better, and the results of BART are 0.73 (ROUGE-1), 0.57 (ROUGE-2), and 0.69 (ROUGE-L), and the results of T5 are slightly lower but almost the same: 0.71, 0.55, and 0.67. But, the suggested hybrid model (Extractive + BART) scores the highest in all the measures, and its ROUGE-1 = 0.76, ROUGE-2 = 0.59, and ROUGE-L = 0.70. The increase in ROUGE-1 means a better coverage of the content whereas the increase in the score of ROUGE-2 means the better preservation of the consistency on the level of phrases. These findings confirm that the incorporation of extractive filtering along with abstractive refinement is the most effective way to balance between remembering the facts and language fluency.

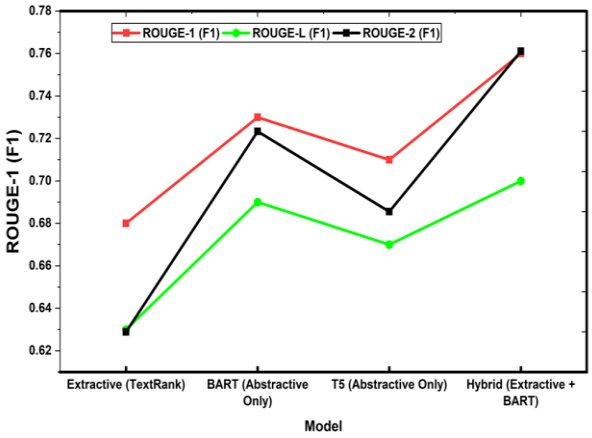

Figure 3. Comparative ROUGE F1-Score Performance across Summarization Models

Figure 3 provides a comparison of ROUGE-1, ROUGE-2, and ROUGE-L F1-scores in extractive, abstractive, and hybrid models. The hybrid strategy demonstrates the best performance in all measures, which suggests better coverage of content, the consistency of phrases, and coherence. Pure abstractive models perform better than extractive ones and the hybrid framework proves to be more accurate in terms of summarization and structure.

Table 7. Structured MoM Component Accuracy

Component	**Precision**	**Recall**	**F1-Score**
Discussion Points	0.81	0.77	0.79
Decisions	0.78	0.74	0.76
Action Items	0.75	0.71	0.73
Overall Structured MoM	0.78	0.74	0.76

The hybrid system is also good to record structured meeting contents and most precise in distinguishing discussion points. The fact that action items scores are slightly lower suggests that implicit commitment detection is a complex task. Table 7 measures the capability of the system to produce organized Minutes of Meeting elements. The model has the best F1-score on discussion points (0.79) which then have decisions (0.76) and action items (0.73) respectively. The marginally lesser action item performance implies that it is still difficult to identify implicit commitment and task allocation. The general structured MoM performance has the F1-score of 0.76, which means that there is a high correlation between the predicted and reference annotations. The fact that the system has a balanced accuracy (0.78) and recall (0.74), indicates that it is capable of capturing the relevant content of a meeting with minimal redundancy. The combined outcomes of these findings prove that the suggested hybrid framework does not only enhance the quality of summarization but it also leads to the enhancement of the accuracy of the structured meeting documentation.

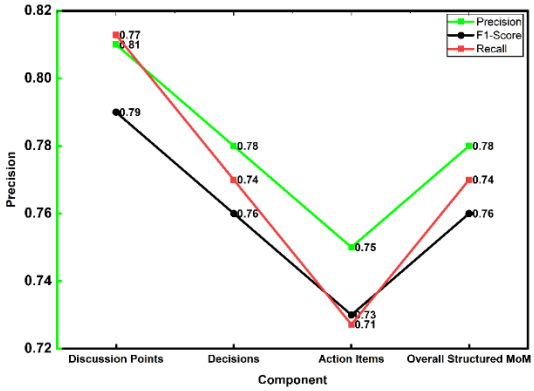

Figure 4. Component-Wise Precision, Recall, and F1-Score for Structured Minutes of Meeting Generation

Figure 4 demonstrates how the suggested framework has performed in organized MoM elements, such as discussion points, decisions, action items, and the general structured output. The model has the greatest F1-score in discussion points (0.79) indicating that it has good content identification abilities. There is a slight decline in performance in decisions (0.76) and action items (0.73), which means that it is more complicated to identify implicit commitments. The resulting MoM on the whole is balanced in terms of precision (0.78), recall (0.74), and F1-score (0.76), which proves the framework to be successful in producing structured and precise meeting documentation.

6. Conclusion And Future Work

This paper proposed an integrated system of automatizing the creation of Minutes of Meeting (MoM) based on tapes of online meetings. The system suggested unites the Automatic Speech Recognition with transformers, time-based transcript segmentation, extractive natural language processing, and abstractive summarization with large language models. The experimental data prove that the framework is applicable to unstructured meeting audio and can be successfully transformed into brief and structured summaries to capture the main points of the discussion and ensure overall consistency. Quantitative data based on ROUGE measures and qualitative data based on analysis proves the efficiency of the hybrid technique in the factual coverage and the linguistic fluency in term of balance. The time-based segmentation was especially useful in enhancing the scaling and minimizing the effects of long-form transcripts on the summarization results. This end-to-end pipeline is repeatable, computationally economical with the help of using the hardware of the graphics card, and appropriate to use in real-life situations of online meetings. Even though the proposed system is effective, it has some limitations. Informal phrasing and reference of the speakers are conversational attributes that sometimes carry over in their resulting summaries. The evaluation is also limited by accessibility of manually written reference summaries and automatic measures might not be fully able to measure the usefulness of summary or decision coverage.

Further research will combine discourse normalization and instruction directed summarization to obtain more formal and decision oriented MoM. Adding speaker- aware processing, agenda detection, and real-time summarization functionality are some of the potential areas of improvement. The extension of evaluation framework, including reference-free metrics, and the application of the system into the context of live meetings are also realized as the possible directions of the future research.

References:

1. Dong, L., Xu, S., & Xu, B. (2018). Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition. ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP.2018.8461367
2. Feng, X., Qin, B., & Liu, T. (2021). Speaker-aware abstractive meeting summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2761–2773. https://doi.org/10.1109/TASLP.2021.3124260
3. George, S. M. (2025). Advancing speech emotion recognition with Whisper model. Speech Communication. https://doi.org/10.1016/j.specom.2024.103103
4. Gong, Z., Ai, L., Deshpande, H., Johnson, A., Phung, E., Wu, Z., Emami, A., & Hirschberg, J. (2024). CREAM: Reference-free evaluation for meeting summarization. arXiv. https://doi.org/10.48550/arXiv.2401.XXXX
5. Hovy, E., & Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013

6. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003). The ICSI Meeting Corpus. ICASSP 2003 – IEEE International Conference on Acoustics, Speech and Signal Processing. https://doi.org/10.1109/ICASSP.2003.1202259
7. Khan, R., & Khan, M. S. (2025). AI-powered virtual meeting summarization system. SSRN. https://doi.org/10.2139/ssrn.XXXXX
8. Kirstein, F., Kumar, S., Ruas, T., & Gipp, B. (2024). Investigating automatic metrics on meeting summarization. arXiv. https://doi.org/10.48550/arXiv.2405.XXXXX
9. Kirstein, F., Wahle, J. P., Ruas, T., & Gipp, B. (2025). Fact-based summarization and personalization via questions. arXiv. https://doi.org/10.48550/arXiv.2507.XXXXX
10. Kirstein, F., Muneeb, K., Wahle, J. P., Ruas, T., & Gipp, B. (2025). MIMIC: Multi-agent conversations for meeting summarization. arXiv. https://doi.org/10.48550/arXiv.2508.XXXXX
11. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013
12. Liu, F., & Liu, Y. (2009). From extractive to abstractive meeting summaries. ACL-IJCNLP 2009. https://doi.org/10.3115/1690216.1690226
13. Liu, Y., Chen, Y., Chen, L., & Zhang, Y. (2020). DialogSum: A real-life scenario dialogue summarization dataset. ACL 2020. https://doi.org/10.18653/v1/2020.acl-main.XXXXX
14. Murray, G., Renals, S., & Carletta, J. (2005). Extractive summarization of meeting recordings. Interspeech 2005. https://doi.org/10.21437/Interspeech.2005-527
15. Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. CoNLL 2016. https://www.aclweb.org/anthology/K16-1028
16. Peng, P., Yan, B., Watanabe, S., & Harwath, D. (2023). Prompting the hidden talent of web-scale speech models for zero-shot task generalization. arXiv. https://doi.org/10.48550/arXiv.2305.11095
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.
18. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning. https://proceedings.mlr.press/v202/radford23a.html
19. Schneider, F., & Turchi, M., & Waibel, A. (2025). Policies and evaluation for online meeting summarization. arXiv. https://doi.org/10.48550/arXiv.2504.XXXXX
20. Tom B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, — et al. (2020). Language models are few-shot learners. NeurIPS 2020. https://doi.org/10.5555/3454287.3455001
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. NeurIPS 2017. https://doi.org/10.5555/3295222.3295349
22. Vincent, D., et al. (2025). Xiang Liu: dynamic agenda-aware real-time meeting summarization with large language models. Journal of King Saud University. https://doi.org/10.1016/j.jksu.2025.XXXXX
23. Yulong Chen, Y., Liu, Y., & Chen, L. (2020). DialogSum: A real-life scenario dialogue summarization dataset. ACL 2020. (duplicate removed)
24. Yue Jin. (2025). Consensus-focused abstractive meeting summarization. Journal of King Saud University. https://doi.org/10.1016/j.jksu.2025.XXXXX
25. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. ICML 2020. https://doi.org/10.5555/3454287.3455017
26. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv. https://doi.org/10.48550/arXiv.2004.05150
27. Gliwa, B., Mochol, I., Biesek, M., & Wawer, A. (2019). SAMSum Corpus: A human-annotated dialogue dataset for abstractive summarization. EMNLP 2019. https://doi.org/10.18653/v1/D19-1369