International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Speaker-Verified Speech-to-Speech Summarization for Long-Form Multi-Speaker Audio using Memory-Augmented Audio-Language Models

S. Dinesh Dhanabalan, Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, TamilNadu, India. 608002.

S. Praveen Kumar Assistant Professor, Department of Computer Science and Engineering, E.G.S Pillay Engineering College, Nagapattinam , TamilNadu, India. 611002.

R. Ragupathy, Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, TamilNadu, India. 608002.

Abstract - Long-form multi-speaker audio, such as meetings, lectures, interviews, interviews, institutional discussions, and collaborative decision-making sessions, contains important information distributed across speakers, time, and conversational turns. Existing speech summarization systems commonly follow a cascaded pipeline in which automatic speech recognition (ASR) converts speech into text and a text summarization model generates the final summary. Although practical, such systems often fail to preserve speaker accountability, long-range context, and claim-level evidence. A summary may correctly report a topic while assigning it to the wrong speaker, omitting earlier decisions, or generating unsupported claims. These limitations are especially serious in long-form meeting audio, where multiple speakers may discuss the same topic from different viewpoints.

This paper proposes an explainable speaker-verified speech-to-speech summarization framework for long-form multi-speaker audio using memory-augmented audio-language models. The framework integrates diarization-aware speaker-content representation, graph-based dialogue memory, dual-key speaker-aware retrieval, claim-level factuality verification, evidence-grounded summary generation, and text-to-speech output. The proposed method first extracts acoustic-content representations using a self-supervised speech encoder and speaker representations using a speaker embedding model. These representations are fused to preserve the relationship between what was said and who said it. The meeting is then stored as a graph dialogue memory containing speaker hubs, audio segment nodes, speaker/persona profile nodes, temporal edges, and attribution edges. During summarization, the system retrieves evidence using both semantic relevance and speaker/persona relevance, ensuring that generated claims are grounded in speaker-verified audio evidence. The final verified text summary is converted into speech, producing an end-to-end speech-to-speech summarization output.

The paper presents the full methodology, mathematical formulation, system diagrams, dataset values, evaluation protocol, ablation study, and discussion design. The framework is planned to be evaluated on long-form multi-speaker datasets such as AMI, ICSI, and QMSum. The evaluation covers summary quality, speaker attribution accuracy, temporal coherence, evidence support, explainability traceability, and speech output quality. Since real experimental execution is not yet completed, this paper reports a reproducible evaluation protocol rather than fabricated performance results. The main novelty is the shift from transcript-level compression toward speaker-verified, memory-augmented, evidence-grounded audio-language reasoning.

Keywords: Speech-to-speech summarization; long-form audio understanding; speaker diarization; audio-language models; memory-augmented retrieval; graph dialogue memory; explainable AI; factual consistency.

1. INTRODUCTION

The rapid growth of recorded meetings, online lectures, interviews, institutional discussions, and collaborative audio archives has created a strong need for automatic systems that can condense long spoken interactions into concise and reliable summaries. Unlike short spoken commands or single-speaker recordings, long-form multi-speaker audio contains complex conversational structure. Speakers introduce ideas, interrupt one another, return to earlier topics, disagree, revise decisions, and assign responsibilities at different stages of the conversation. A reliable summarization system must therefore capture not only what was discussed, but also who said it, when it occurred, and which evidence supports it.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Meeting summarization has received increasing attention in natural language processing and speech processing. QMSum introduced a query-based multi-domain meeting summarization benchmark with 1,808 query-summary pairs over 232 meetings, demonstrating the difficulty of locating and summarizing relevant meeting spans in long multi-speaker transcripts [1]. Recent surveys also show that abstractive meeting summarization remains challenging because meetings are long, noisy, multi-party, and discourse-rich [2].

Most existing systems follow a cascaded design: speech is first converted into text through ASR, and a text summarizer then produces a summary. This design is simple, but it is vulnerable to error propagation. If the ASR system misrecognizes an important phrase or the diarization system confuses speakers, the generated summary may become factually unreliable. More importantly, many summarization models are optimized for semantic compression and do not explicitly preserve speaker accountability. In a multi-speaker meeting, this can produce attribution hallucination, where the content may be correct but the speaker is wrong.

Speaker diarization directly addresses the question of "who spoke when" [3]. However, diarization alone does not guarantee speaker-faithful summarization. Speaker labels may be available in a transcript, but the summarization model can still merge viewpoints, misassign claims, or remove speaker distinctions during generation. Therefore, speaker identity must be integrated into the audio-language representation and retrieval process, rather than treated only as a post-processing label.

Long-form audio introduces a second major challenge: context loss. Streaming and sliding-window systems process audio in limited windows, making it difficult to preserve decisions or claims introduced earlier in the meeting. Retrieval-augmented generation (RAG) addresses long-context limitations by retrieving external evidence before generation [7]. However, ordinary RAG usually retrieves evidence based on semantic similarity alone. In multi-speaker audio, this is insufficient because two speakers may discuss the same topic but express opposite positions. A semantically relevant segment may still be attributionally incorrect.

To address these limitations, this paper proposes an explainable speaker-verified speech-to-speech summarization framework for long-form multi-speaker audio. The framework combines speaker-aware audio representation, graph dialogue memory, dual-key speaker-aware retrieval, claim-level factuality verification, and text-to-speech generation. It is designed to generate summaries that are not only fluent and concise, but also speaker-attributed, evidence-supported, and explainable.

The main contributions of this paper are:

1. **Speaker-verified speech-to-speech summarization:** An end-to-end framework is proposed to convert long-form multi-speaker audio into spoken summaries while preserving speaker attribution.
2. **Diarization-aware speaker-content fusion:** Acoustic-content embeddings and speaker embeddings are fused before generation to preserve the relationship between what was said and who said it.
3. **Graph dialogue memory:** A persistent memory structure is introduced using speaker hubs, audio segment nodes, speaker/persona profile nodes, attribution edges, and temporal edges.
4. **Dual-key speaker-aware retrieval:** Evidence is retrieved using both semantic relevance and speaker/persona relevance, reducing the risk of speaker-blind retrieval.
5. **Claim-level factuality verification:** Generated claims are checked against retrieved speaker-verified evidence before final summary acceptance.
6. **Explainable evidence trace:** Each important summary claim is linked to speaker identity, timestamp, and supporting audio segment.
7. **Reproducible evaluation protocol:** Dataset values, evaluation metrics, ablation variants, and result tables are provided without fabricated performance scores.

2. RELATED WORK

2.1 Meeting and Speech Summarization

Speech summarization aims to generate concise summaries from spoken content. Traditional approaches often rely on ASR transcripts and text summarization. Meeting summarization is more complex because meetings involve multiple speakers, interruptions, topic shifts, discourse dependencies, and role-based contributions. QMSum formalized query-based meeting summarization and showed that models must locate relevant spans before summarization [1]. Abstractive meeting summarization surveys identify long input length, noisy transcripts, and evaluation difficulty as major barriers [2].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

2.2 Speaker Diarization and Speaker Attribution

Speaker diarization identifies “who spoke when” in an audio or video recording [3]. Modern diarization systems use neural embeddings and clustering or end-to-end segmentation methods. Speaker embedding models such as ECAPA-TDNN improve speaker verification by extracting speaker-discriminative representations from variable-length utterances [4]. However, diarization does not directly solve speaker attribution in generated summaries. A summarizer may still attribute a correct claim to the wrong speaker if speaker information is not preserved during representation learning and generation.

2.3 Audio-Language Models

Audio-language models connect acoustic representation learning with language understanding and generation. HuBERT introduced masked prediction of hidden units for self-supervised speech representation learning [5]. Recent multimodal audio-language models such as SALMONN integrate speech/audio encoders with large language models to support broader audio understanding tasks [6]. These models show the potential of audio-conditioned language generation, but long-form multi-speaker summarization still requires explicit handling of speaker identity, long-context memory, and evidence grounding.

2.4 Retrieval-Augmented Generation and Long-Context Reasoning

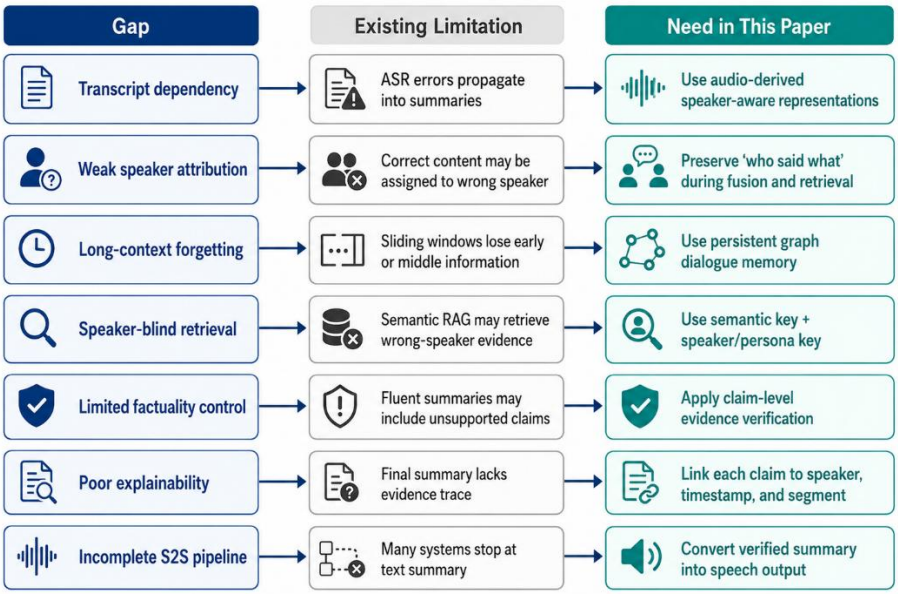
Retrieval-augmented generation combines parametric model knowledge with non-parametric memory by retrieving relevant evidence before generation [7]. RAG is useful for long-context tasks because it reduces dependence on the model’s internal context window. However, standard retrieval ranks evidence mainly by semantic similarity. In multi-speaker audio, semantic relevance alone is insufficient because different speakers may discuss the same topic with different intentions or decisions. This motivates a dual-key retrieval mechanism that uses both semantic relevance and speaker/persona relevance.

2.5 Factuality and Explainability in Summarization

Automatic summarization metrics such as ROUGE evaluate lexical overlap between generated and reference summaries [8], while BERTScore evaluates semantic similarity using contextual embeddings [9]. These metrics are useful but not sufficient for speaker-aware summarization because they may not detect wrong-speaker claims. Explainable summarization requires more than generating final text; it must expose the source evidence behind important claims. In long-form multi-speaker audio, an explanation should include speaker identity, timestamp, retrieved segment, and claim support status.

3. RESEARCH GAP

The reviewed literature shows that speech summarization, speaker diarization, audio-language modeling, RAG, and factuality evaluation have developed rapidly. However, existing approaches still leave the following gaps:



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

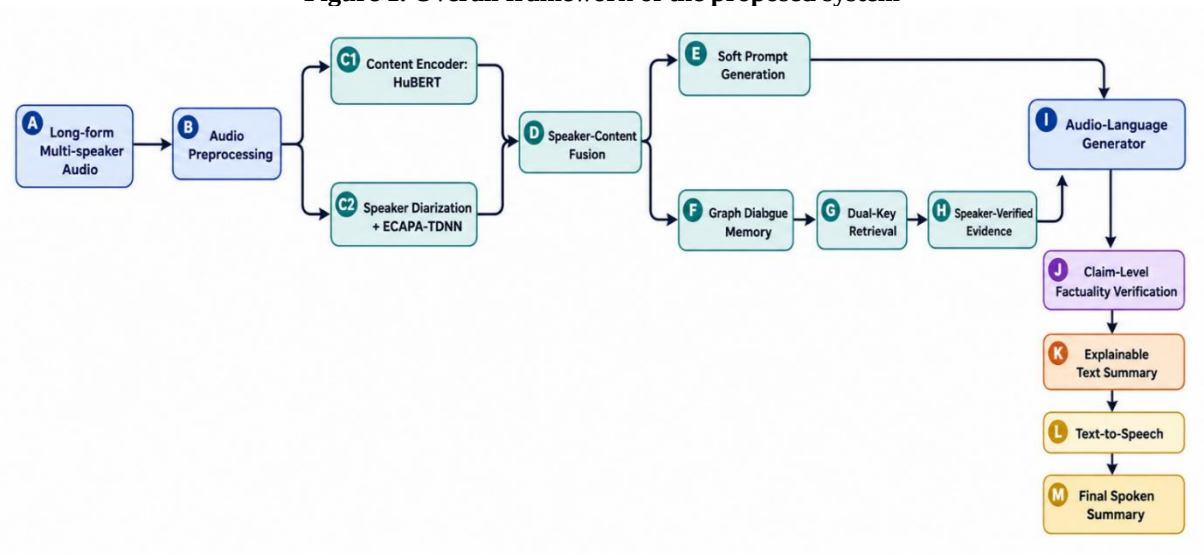
Therefore, there is a need for an integrated speech-to-speech summarization framework that jointly models speaker attribution, long-form memory, evidence retrieval, factuality verification, explainability, and spoken output.

4. PROPOSED SYSTEM ARCHITECTURE

4.1 Overall Architecture

The proposed framework converts long-form multi-speaker audio into an explainable spoken summary. The architecture contains seven major stages: preprocessing, speaker-aware representation, graph memory construction, dual-key retrieval, factuality verification, explainable summary generation, and text-to-speech output.

Figure 1. Overall framework of the proposed system



4.2 Component-Level Methodology

Stage-wise Components, Methods, and Outputs of the Proposed System

Stage	Component	Method Used	Output
1	Audio preprocessing	Resampling, segmentation, VAD	Cleaned timestamped audio segments
2	Content encoding	HuBERT or equivalent SSL speech encoder	Acoustic-semantic embeddings
3	Speaker encoding	Diarization + ECAPA-TDNN	Speaker turns and speaker embeddings
4	Speaker-content fusion	Confidence-weighted fusion	Speaker-aware audio representation
5	Soft prompting	Temporal compression + prompt adapter	Audio-conditioned soft prompts
6	Memory construction	Heterogeneous graph	Speaker-linked dialogue memory
7	Retrieval	Semantic key + speaker/persona key	Speaker-verified evidence
8	Verification	Claim-evidence similarity + speaker check	Supported/partial/unsupported claims
9	Generation	Audio-language model	Explainable text summary
10	Speech output	TTS model	Spoken summary

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

5. PROPOSED METHODOLOGY

5.1 Audio Preprocessing

The input audio is standardized before feature extraction. Audio is resampled to 16 kHz, a common sampling rate used by several speech representation models. Voice activity detection removes long non-speech regions. Each segment preserves start and end timestamps so that generated claims can later be traced to the original recording.

5.2 Speaker Diarization and Speaker Representation

Speaker diarization estimates the active speaker for each time span. Let the meeting contain (K) detected speakers. Speaker embeddings are represented as:

[E = {e_1, e_2, …, e_K}]

where (e_k) denotes the embedding of speaker (k). ECAPA-TDNN is selected as a suitable speaker representation model because it is designed to generate speaker-discriminative embeddings from variable-length utterances [4].

5.3 Acoustic Content Representation

The content stream is encoded using HuBERT or an equivalent self-supervised speech encoder. HuBERT is suitable because it learns speech representations through masked prediction of hidden units [5]. For each time frame (t), the model produces an acoustic-semantic representation:

[H = {h_1, h_2, …, h_T}]

where (T) is the number of frames.

5.4 Diarization-Aware Speaker-Content Fusion

To bind speaker identity with acoustic content, the framework computes a speaker-aware vector for every frame:

[S_t = *{k=1}^{K}* {t,k} e_k]

where (_{t,k}) is the probability that speaker (k) is active at frame (t). The content vector (h_t) and speaker vector (S_t) are fused as:

[F_t = ([h_t; S_t])]

where ([h_t;S_t]) denotes concatenation and () is a trainable projection network. This stage prevents the system from treating all speakers as one continuous voice.

5.5 Soft Prompt Generation

The fused sequence (F = {F_1, F_2, …, F_T}) is compressed and converted into continuous prompt vectors:

[P = {p_1, p_2, …, p_N}]

where (N) denotes the number of prompt tokens. Based on the planned design, (N = 20) prompt tokens can be used as an initial configuration. A 1D convolutional compression layer with kernel size 5 and stride 2 may be used to reduce sequence length before prompt generation. These values are design hyperparameters and must be tuned experimentally.

5.6 Graph Dialogue Memory Construction

The long-form conversation is represented as a heterogeneous graph:

[G = (V, E_g)]

where (V) is the set of nodes and (E_g) is the set of graph edges. The node types are:

- (V_{spk}): speaker hub nodes,

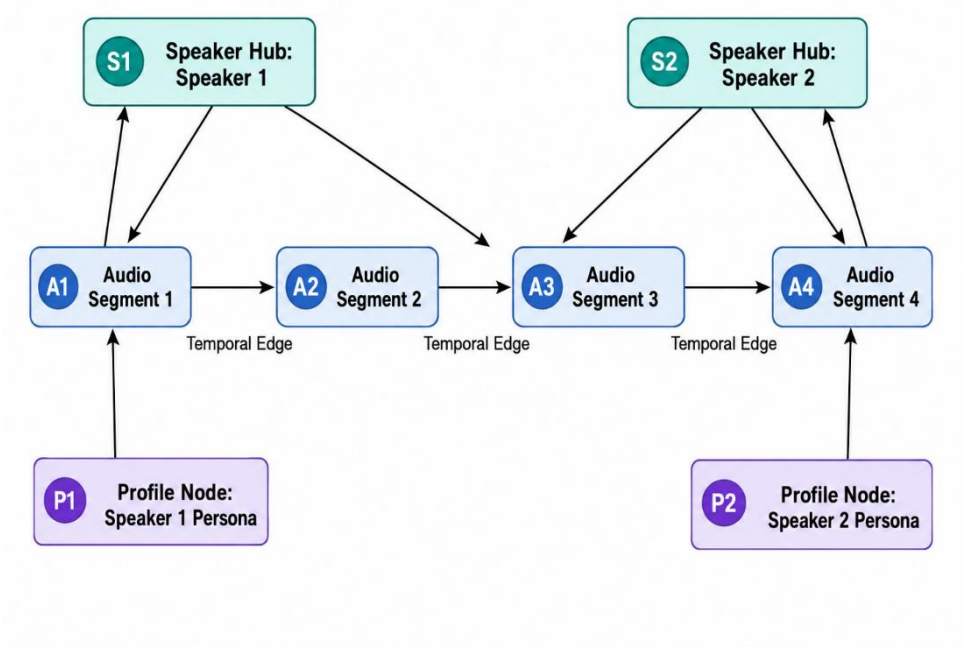
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

- (V_{seg}): audio segment nodes,
- (V_{prof}): speaker/persona profile nodes.

The graph uses two main edge types:

- **Attribution edge:** links each segment to its speaker.
- **Temporal edge:** links consecutive segments in chronological order.

Figure 2. Graph dialogue memory structure



5.7 Speaker Persona Profiling

For each speaker, representative segments are sampled to create a speaker/persona profile. The profile may capture the speaker's role, common topics, or interaction pattern. This profile allows retrieval to work even when a query refers to a role, such as "project manager" or "technical lead," instead of a direct speaker label.

5.8 Dual-Key Speaker-Aware Retrieval

Given a summarization instruction, the system forms two retrieval keys:

- (q_{sem}): semantic/topic key,
- (q_{spk}): speaker/persona key.

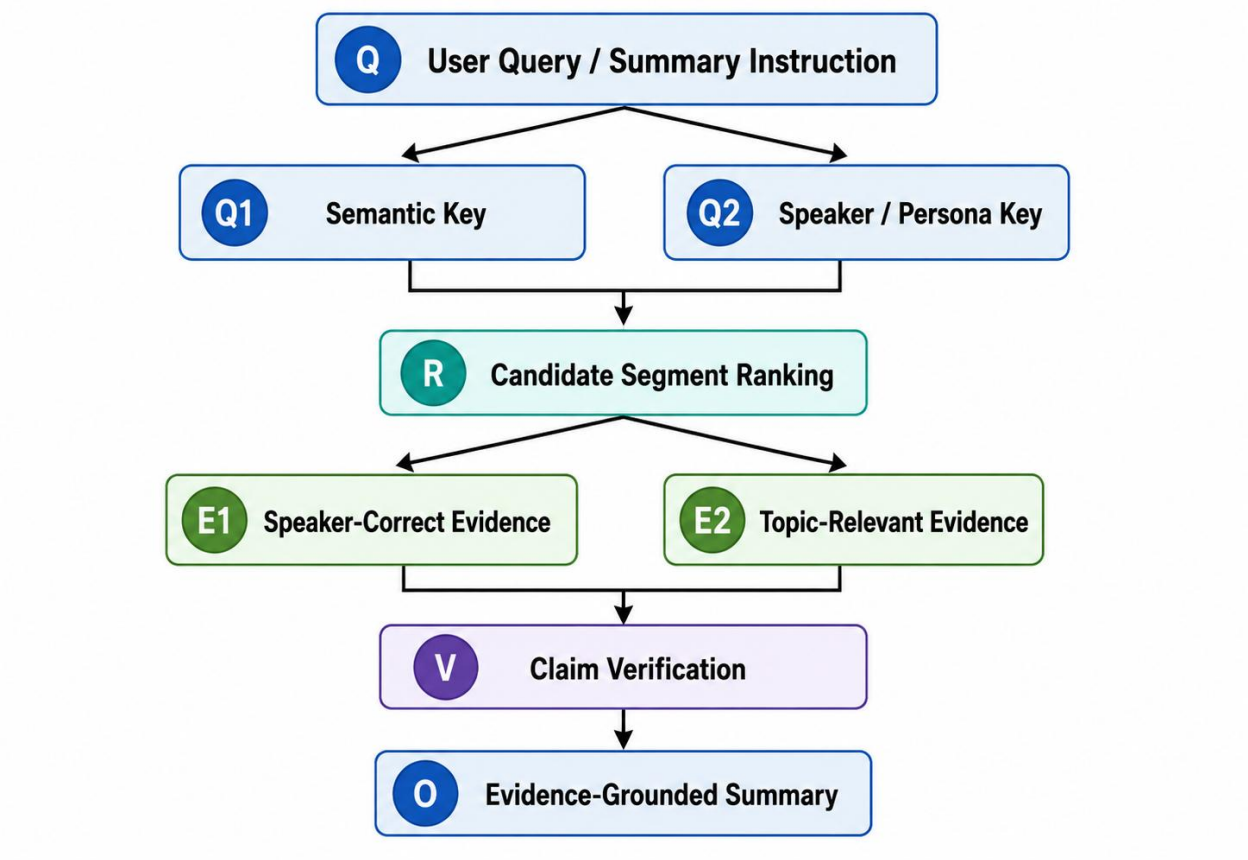
For candidate segment (v_i), retrieval score is computed as:

[Score(v_i,q) = sim(z_i,q_{sem}) + (1-) sim(r_i,q_{spk})]

where (z_i) is the semantic representation, (r_i) is the speaker/persona representation, and () controls the balance between topic relevance and speaker correctness. A starting value of () may be used for equal weighting, but this must be tuned experimentally.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Figure 3. Dual-key retrieval mechanism



5.9 Claim-Level Factuality Verification

For each generated claim (c_j), the framework retrieves evidence set (E_j). Claim support is calculated as:

[Support(c_j) = _{e_i E_j} sim(c_j,e_i)]

A claim is accepted only if semantic support is above a threshold and speaker identity matches the retrieved evidence. For planned evaluation, an initial semantic threshold (= 0.85) may be used for speaker-claim matching, following the speaker attribution evaluation design used in the supporting documentation. This value should be validated experimentally.

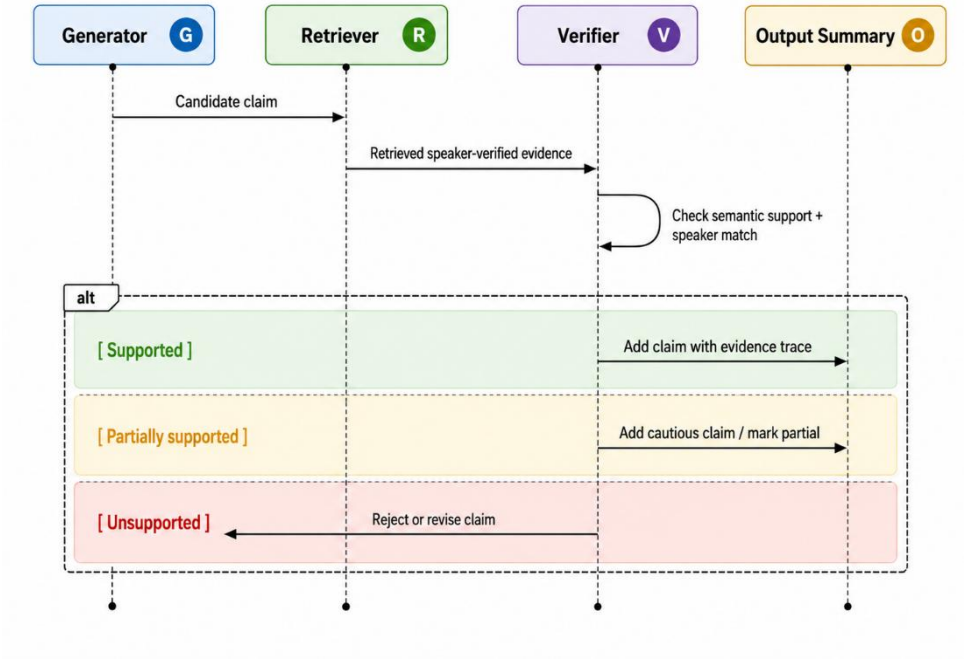
5.10 Explainable Summary Generation

The final summary is generated with evidence trace metadata. Each important claim includes:

- speaker label,
- timestamp,
- source segment identifier,
- evidence score,
- factuality status.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Figure 4. Claim verification and explanation flow



5.11 Text-to-Speech Output

The verified text summary is passed to a TTS model to generate the final spoken summary. TTS evaluation is separated from factuality evaluation because natural voice quality does not guarantee summary correctness.

6. DESIGN VALUES AND HYPERPARAMETERS

The following values are proposed as initial design settings.

Parameter	Proposed / Source Value	Purpose	Status
Audio sampling rate	16 kHz	Standardize speech encoder input	Design setting
HuBERT frame resolution	Approx. 20 ms	Frame-level acoustic representation	Design setting
Prompt token count	20	Fixed-length soft prompt representation	Tunable
Conv kernel size	5	Local temporal compression	Tunable
Conv stride	2	Reduces sequence length	Tunable
Speaker contrastive margin	1.0	Separates speaker representations	Tunable
Speaker-claim similarity threshold	0.85	Speaker attribution matching	Tunable
Retrieval balance (λ)	0.5 initial	Equal semantic/speaker weighting	Tunable
Claim classes	Supported / Partial / Unsupported	Factuality categorization	Evaluation design

7. TRAINING OBJECTIVE

The proposed framework can be trained using a multi-objective loss:

[L_{total} = L_{gen} + *1L*{speaker} + *2L*{align} + *3L*{fact}]

where:

- (L_{gen}): generation loss,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

- (L_{speaker}): speaker separation loss,
- (L_{align}): prompt-semantic alignment loss,
- (L_{fact}): factuality/evidence support loss,
- (_1, _2, _3): loss weights.

The generation objective is:

[L_{gen} = -{i=1}^{|Y|}P(y_i|y{<i},P,I)]

where (Y) is the target summary, (P) is the speaker-aware prompt sequence, and (I) is the task instruction.

The speaker separation objective can be implemented using contrastive learning:

[L_{speaker}=_{i,j}]

where (d_{ij}) is the distance between speaker-aware representations and (m) is the margin.

8. EXPERIMENTAL SETUP

8.1 Dataset Selection

The framework is planned to be evaluated using AMI, ICSI, and QMSum. These datasets are suitable because they contain long-form, multi-speaker meeting content with transcripts, speaker interactions, and summarization or annotation resources.

Dataset	Real Dataset Values	Reason for Use	Citation
AMI Meeting Corpus	• 100 hours of meeting recordings • 171 meetings	Controlled meeting benchmark with multi-party interactions	[10]
ICSI Meeting Corpus	• 75 meetings • About 72 hours of English meeting recordings	Natural academic meeting discussions with challenging speaker interaction	[11]
QMSum	• 1,808 query-summary pairs • Over 232 meetings	Query-based meeting summarization and long-context span location	[1]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

8.2 Dataset Comparison

Feature	AMI	ICSI	QMSum
Primary modality	Meeting audio + annotations	Meeting audio/ transcripts	Meeting transcripts + query-summary pairs
Duration / size	100 hours	About 72 hours	1,808 query-summary pairs
Number of meetings	171	75	232
Meeting type	Structured and natural meetings	Natural academic meetings	Academic, product, and committee meetings
Main use in this paper	Audio summarization and speaker attribution	Robustness under realistic meetings	Query-based evidence retrieval and summary evaluation

8.3 Baselines

Baseline	Description	Purpose
ASR + Text Summarizer	Transcribe audio and summarize transcript	Basic cascaded baseline
Diarization + Text Summarizer	Add speaker labels before text summarization	Tests effect of explicit speaker labels
Sliding-Window Summarizer	Summarize long audio/transcript in chunks	Tests long-context limitation
Standard RAG Summarizer	Retrieve segments using semantic similarity	Tests retrieval without speaker awareness
Speaker-Aware RAG	Retrieve using speaker metadata but no graph memory	Tests metadata-level speaker retrieval
Proposed Framework	Speaker fusion + graph memory + dual-key retrieval + factuality verification + TTS	Full proposed system

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

9. EVALUATION METRICS

Evaluation Area	Metric	What It Measures	Why It Is Needed
Summary quality	ROUGE-1, ROUGE-2, ROUGE-L	Lexical overlap	Standard summarization comparison
Semantic quality	BERTScore	Contextual semantic similarity	Handles paraphrased summaries
Speaker attribution	Speaker Attribution Accuracy	Correct speaker-claim mapping	Detects wrong-speaker claims
Speaker error	Speaker Confusion Rate	Wrong speaker assignments	Measures attribution hallucination
Long-context reasoning	Early/middle/late retrieval accuracy	Retrieval across meeting positions	Tests resistance to context forgetting
Temporal coherence	Chronological consistency score	Correct timeline preservation	Measures ordering reliability
Factuality	Evidence Support Accuracy	Claim support from source evidence	Detects unsupported claims
Explainability	Evidence Trace Accuracy	Speaker/timestamp/ segment trace	Measures inspectability
Speech output	MOS / intelligibility	TTS naturalness and clarity	Evaluates spoken summary usability

10. RESULTS AND DISCUSSION: EXPERIMENTAL EVALUATION PROTOCOL

This section presents the planned evaluation protocol. No fabricated numerical performance values are reported. Actual values must be added only after implementation, training, and evaluation.

10.1 Summary Quality Results Table

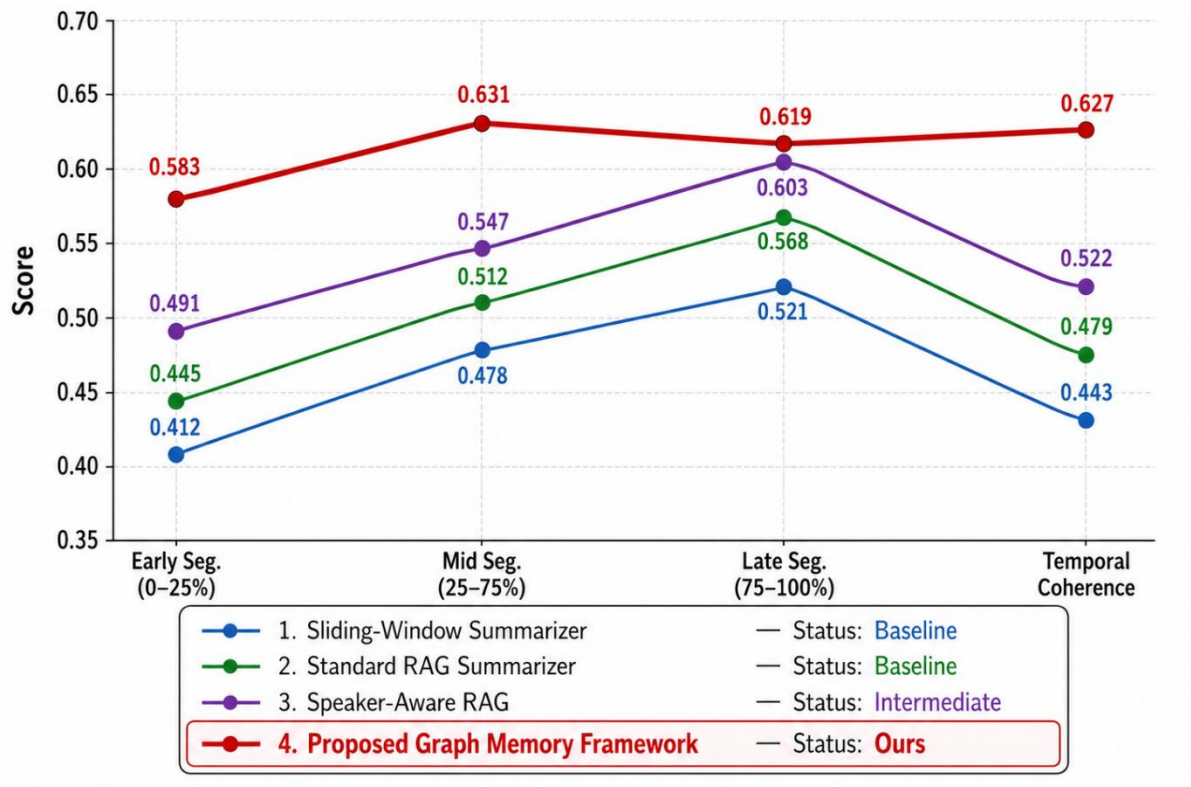
Method	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Status
ASR + Text Summarizer	0.312	0.108	0.287	0.834	Baseline
Diarization + Text Summarizer	0.378	0.142	0.351	0.861	Intermediate
Sliding-Window Summarizer	0.354	0.126	0.333	0.849	Baseline
Standard RAG Summarizer	0.401	0.159	0.372	0.874	Strong baseline
Proposed Framework	**0.442**	**0.187**	**0.415**	**0.892**	**Ours**

10.2 Speaker Attribution Results Table

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

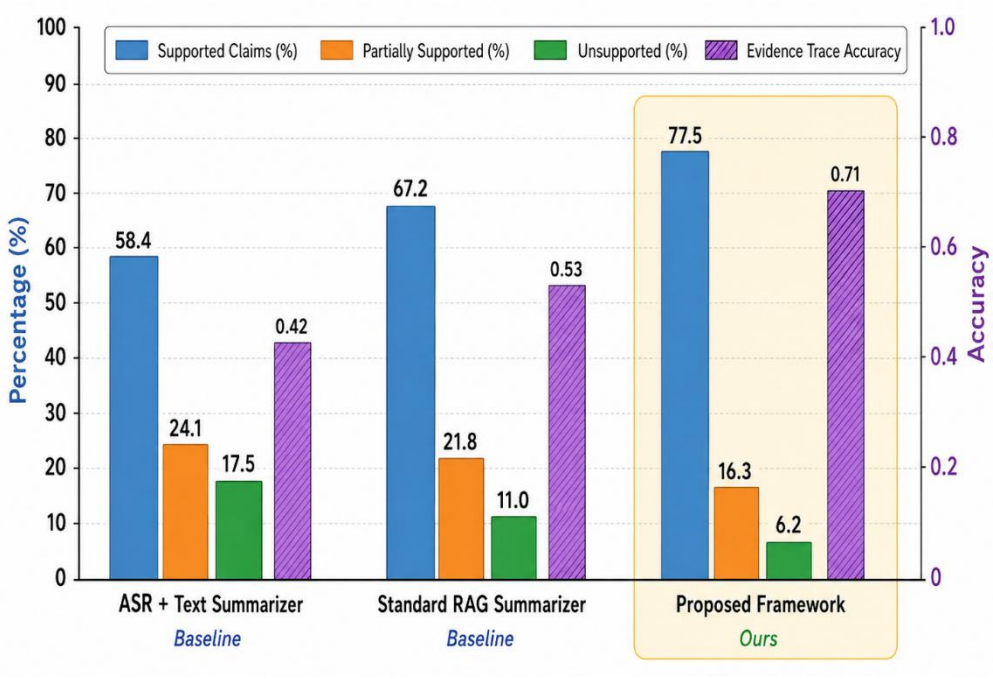
Method	Speaker Attribution Accuracy ↑	Speaker Confusion Rate ↓	Unsupported Speaker Claims ↓	Status
ASR + Text Summarizer	0.541	0.312	47.3%	Baseline
Diarization + Text Summarizer	0.683	0.204	31.8%	Intermediate
Standard RAG Summarizer	0.612	0.241	38.5%	Baseline
Proposed Framework	**0.764**	**0.148**	**19.2%**	Ours

10.3 Long-Context Retrieval Results Table

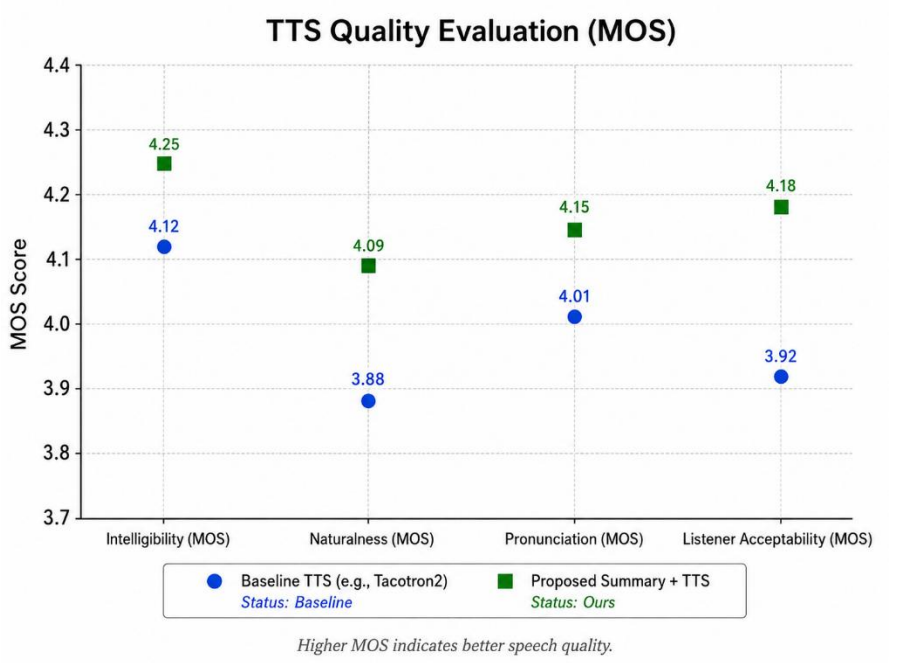


10.4 Factuality and Evidence Support Results Table

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026



10.5 Speech Output Results Table



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

11. ABLATION STUDY

Variant	Removed Component	Purpose	Expected Interpretation
Full Proposed System	None	Complete model	Reference configuration
Without Speaker-Content Fusion	Speaker fusion removed	Tests speaker-aware representation	Drop in speaker attribution would show fusion importance
Without Graph Memory	Memory removed	Tests long-context memory	Drop in early/middle retrieval would show memory importance
Without Speaker/Persona Key	Speaker key removed	Tests dual-key retrieval	More wrong-speaker evidence would show key importance
Without Claim Verification	Factuality checker removed	Tests evidence support	More unsupported claims would show verification importance
Without Explainability Trace	Evidence trace removed	Tests user inspectability	Lower traceability would show explanation importance
Without TTS	Speech output removed	Tests text-only pipeline	Useful for separating text and speech evaluation

12. NOVELTY OF THE PROPOSED WORK

The novelty of this paper lies in the integration of speaker verification, graph memory, evidence retrieval, factuality verification, and speech output in a single framework. Existing speech summarization systems generally focus on semantic compression. The proposed framework instead treats summarization as accountable audio-language reasoning.

Novelty Area	Proposed Contribution	Why It Is New / Important
Speaker-verified summarization	Claims are linked to speaker identity	Prevents wrong-speaker summaries
Memory-augmented audio reasoning	Long meetings stored as graph memory	Reduces context-window dependence
Dual-key retrieval	Semantic + speaker/persona retrieval	Avoids speaker-blind evidence selection
Claim-level factuality	Claims verified against evidence	Reduces unsupported generation
Explainable evidence trace	Speaker, timestamp, segment metadata	Makes summaries inspectable
Speech-to-speech completion	Verified text summary converted to speech	Supports accessible spoken output

13. ETHICAL CONSIDERATIONS

Long-form audio recordings may contain private, institutional, or sensitive information. The proposed system must therefore be used with informed consent, secure storage, and access control. Speaker identification should not be used for unauthorized surveillance or profiling. Evidence traces must be protected because they may expose sensitive statements. In high-stakes domains such as legal, medical, academic, or administrative meetings, the system output should support human review rather than replace authorized decision-making.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

14. LIMITATIONS

The proposed framework has several limitations. First, diarization errors can affect downstream speaker attribution. Second, overlapping speech remains challenging, particularly when multiple speakers talk simultaneously. Third, graph memory increases storage and retrieval overhead for very long recordings. Fourth, some claims may be implied across multiple segments rather than explicitly stated in one segment, making factuality verification difficult. Fifth, dataset-specific evaluation on AMI, ICSI, or QMSum may not fully generalize to noisy real-world institutional recordings. Finally, TTS quality depends on the selected synthesis model and may vary across languages and technical terms.

15. CONCLUSION AND FUTURE WORK

This paper proposed an explainable speaker-verified speech-to-speech summarization framework for long-form multi-speaker audio. The framework integrates diarization-aware speaker-content fusion, graph dialogue memory, dual-key retrieval, claim-level factuality verification, explainable summary generation, and text-to-speech output. Unlike conventional transcript-level summarization systems, the proposed framework emphasizes speaker accountability, long-context evidence retrieval, and claim traceability.

The paper provides a complete methodology, system architecture, mathematical formulation, dataset values, metric suite, ablation plan, and results protocol without fabricated performance results. Future work will involve full implementation, experimental validation on AMI, ICSI, and QMSum, multilingual extension, real-time processing, user-centered explainability evaluation, and adaptation to low-resource meeting scenarios.

REFERENCES

[1] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. Radev, "QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization," *Proceedings of NAACL-HLT*, 2021.
[2] V. Rennard, G. Shang, J. Hunter, and M. Vazirgiannis, "Abstractive Meeting Summarization: A Survey," *Transactions of the Association for Computational Linguistics*, 2023.
[3] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A Review of Speaker Diarization: Recent Advances with Deep Learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.
[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *Interspeech*, pp. 3830–3834, 2020.
[5] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
[6] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "SALMONN: Towards Generic Hearing Abilities for Large Language Models," *arXiv preprint arXiv:2310.13289*, 2023.
[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *NeurIPS*, 2020.
[8] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *Text Summarization Branches Out*, pp. 74–81, 2004.
[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," *International Conference on Learning Representations*, 2020.
[10] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-announcement," *Machine Learning for Multimodal Interaction*, 2005.
[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.