INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

Hallucination-Aware Multilingual Speech-to-Speech Summarization using Reinforcement-Aligned Large Speech Language Models

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

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

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

Abstract

The rapid growth of long-form spoken content, including meetings, lectures, interviews, classroom discussions, and online consultations, has created a strong need for automatic systems that can generate concise, faithful, and accessible summaries from speech [1]. Existing speech summarization systems commonly depend on cascaded pipelines, where speech is first converted into text using automatic speech recognition and then summarized using text-based models. Although effective, these systems suffer from ASR error propagation, loss of speaker-level information, multilingual degradation, and text-only output limitations. In addition, large language model-based summarizers may generate hallucinated or unsupported information, which reduces the reliability of summaries in high-value domains such as education, healthcare, legal documentation, and business communication [2]. To address these issues, this paper proposes a hallucination-aware multilingual speech-to-speech summarization framework using reinforcement-aligned large speech language models. The proposed system is designed to be evaluated using ROUGE, BERTScore, WER, CER...**The proposed system is designed to be evaluated using ROUGE, BERTScore and WER..Inspired by recent multilingual speech-to-speech modeling and preference-alignment methods, the framework applies Direct Preference Optimization or reinforcement-style optimization to improve factual consistency and reduce unsupported summary claims [3], [4]. The factuality verification module compares generated summaries with source speech/transcript evidence to identify contradictions, missing context, and speaker-attribution errors [5]. The proposed system is evaluated using ROUGE, BERTScore, WER, CER, hallucination rate, factual consistency score, speaker attribution accuracy, multilingual adequacy, latency, and mean opinion score. This study contributes a unified framework for faithful, multilingual, and speech-output-oriented summarization of long-form spoken content.**

Keywords

Speech-to-Speech Summarization; Multilingual Speech Summarization; Large Speech Language Models; Hallucination Detection; Factual Consistency; Reinforcement Alignment; Direct Preference Optimization; Speaker-Aware Summarization; Long-Form Speech Understanding; Multimodal Large Language Models

1. Introduction

The rapid expansion of spoken digital content has created a strong need for intelligent systems that can understand, summarize, and communicate information from long-form audio. Meetings, online lectures, interviews, seminars, healthcare consultations, legal discussions, customer-support calls, and multilingual conversations are now routinely recorded and stored in digital form. Although these recordings contain valuable information, they are often lengthy, conversational, speaker-dependent, and difficult to review manually. As a result, automatic speech summarization has become an important research direction that connects automatic speech recognition, spoken language understanding, text summarization, meeting analysis, and multimodal language modeling [1].

Automatic speech summarization is important because it reduces the time and effort required to extract useful information from long audio recordings. Instead of listening to an entire meeting, lecture, or consultation, users can access concise outputs such as key points, decisions, speaker-specific statements, action items, topic summaries, and follow-up tasks. This is highly useful in domains such as education, business, healthcare, legal documentation, public administration, and assistive communication. In practical settings, however, summarization should not be limited to written text alone. Many users may prefer spoken summaries, especially in mobile, accessibility, multilingual, and low-literacy environments.

Conventional speech summarization systems usually follow a cascaded pipeline in which speech is first converted into text using automatic speech recognition (ASR), and the generated transcript is then passed to a text summarization model. Although this approach is simple and widely used, it has several limitations. ASR errors may propagate into the summarization stage and produce incorrect summaries. Speaker diarization errors may cause wrong speaker attribution in multi-party conversations. Transcript-only summarization may also ignore acoustic and paralinguistic information such as emphasis, hesitation, tone, speaker confidence, and conversational intent. These limitations become more severe in noisy recordings, multilingual speech, code-switched conversations, and low-resource language conditions. Recent speech summarization research also highlights that realistic benchmarks, multilingual datasets, long-context speech handling, and reliable evaluation remain open challenges [1].

The need for multilingual and speech-to-speech summarization is further strengthened by recent progress in unified speech and text modeling. Multilingual speech systems such as SeamlessM4T show that speech-to-speech translation, speech-to-text translation, text-to-speech translation, text-to-text translation, and ASR can be supported within a unified multilingual framework [2]. Similarly, large speech language models such as SpeechGPT and AudioPaLM demonstrate that speech and text modalities can be jointly modeled for speech understanding and speech generation [3], [4]. These developments indicate that future summarization systems can move beyond text-only outputs and support direct speech input, multilingual understanding, faithful summary generation, and spoken summary output.

Despite these advances, hallucination remains a major challenge in large language model-based summarization. Hallucination occurs when a model generates unsupported, fabricated, or factually incorrect information that is not grounded in the source speech or transcript. In speech summarization, hallucination may appear as incorrect decisions, wrong speaker statements, false action items, inaccurate dates, fabricated numerical values, or misleading conclusions. This problem is particularly risky in high-value domains such as healthcare, legal documentation, education, and business communication. Factual consistency evaluation methods such as

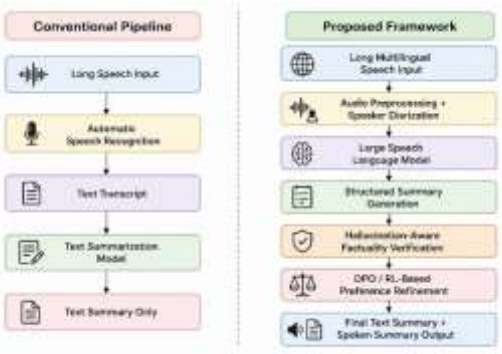
QAFactEval show the importance of verifying whether generated summaries are supported by source evidence [5]. However, factual verification becomes more complex in speech-based systems because the model must consider ASR uncertainty, multilingual meaning, speaker turns, and long-context audio structure.

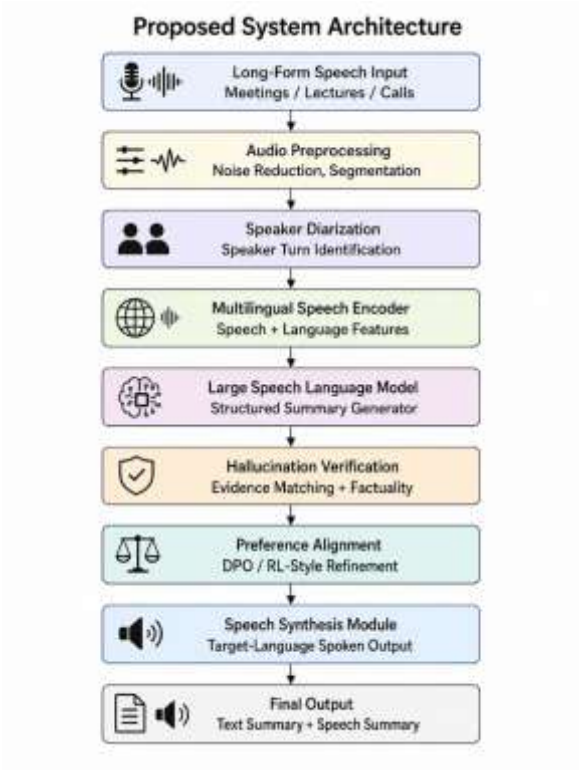
To improve the reliability of generated summaries, reinforcement and preference-alignment methods have become increasingly important. Direct Preference Optimization (DPO) offers a stable and computationally simpler approach for aligning language models with preference pairs, avoiding some of the complexity of traditional reinforcement learning from human feedback [6]. Recent multimodal LLM-based speech summarization research has also explored reinforcement learning, supervised fine-tuning, knowledge distillation, and DPO-style training to improve direct speech summarization and reduce hallucination [7]. These methods motivate the design of a reinforcement-aligned summarization framework that can prefer faithful, grounded, speaker-aware, and multilingual summaries over unsupported or hallucinated outputs.

Although existing studies have separately advanced speech summarization, multilingual speech translation, speech-language modeling, factual consistency evaluation, and preference alignment, a clear research gap remains. Most existing systems still focus on speech-to-text summarization or text-only summary generation. Limited work has addressed a unified framework that performs multilingual speech-to-speech summarization while also integrating hallucination-aware factuality verification and reinforcement-based summary refinement. Furthermore, many evaluations rely mainly on text-based metrics such as ROUGE and BERTScore, while practical speech-to-speech summarization also requires hallucination rate, factual consistency, speaker attribution accuracy, multilingual adequacy, latency, and speech quality evaluation.

Therefore, this study proposes "Hallucination-Aware Multilingual Speech-to-Speech Summarization using Reinforcement-Aligned Large Speech Language Models." The objective of the study is to design a framework that takes long-form speech as input, performs multilingual speech understanding, identifies speaker-level information, generates structured summaries, verifies factual consistency, refines the summary using reinforcement or preference alignment, and produces the final output as speech. The study also aims to define a comprehensive evaluation strategy that measures summarization quality, transcription quality, hallucination reduction, speaker attribution, multilingual adequacy, computational efficiency, and generated speech quality.

The main contributions of this paper are as follows. First, the paper proposes a unified multilingual speech-to-speech summarization framework for long-form spoken content. Second, it introduces a hallucination-aware factuality verification module to identify unsupported, contradictory, or speaker-misattributed summary content. Third, it incorporates reinforcement/preference alignment using DPO or RL-style optimization to improve factual consistency and reduce hallucinated outputs. Fourth, it supports speaker-aware structured summarization for meetings, lectures, interviews, and consultations. Fifth, it presents a multi-dimensional evaluation strategy using ROUGE, BERTScore, WER, CER, hallucination rate, factual consistency score, speaker attribution accuracy, multilingual adequacy, latency, and mean opinion score.





2. Related Work

2.1 Speech Summarization

Speech summarization focuses on generating concise and meaningful summaries from spoken content such as meetings, lectures, interviews, conversations, and audiovisual recordings. Unlike text summarization, speech summarization must handle acoustic variation, speaker changes, pauses, repetitions, disfluencies, background noise, incomplete sentence structures, and conversational flow. Recent survey work describes speech summarization as an emerging interdisciplinary area that connects automatic speech recognition, text summarization, meeting summarization, spoken language understanding, and multimodal processing [1]. Existing speech summarization methods are generally grouped into extractive and abstractive approaches. Extractive methods select important utterances or speech segments from the original input, while abstractive methods generate new summary sentences that may not directly appear in the source.

Although extractive methods are easier to ground in the original speech, they often produce less fluent and less compact summaries. Abstractive methods generate more natural summaries, but they are more vulnerable to factual inconsistency and hallucination. Therefore, a reliable speech summarization framework should combine the fluency of abstractive generation with mechanisms for factual verification, speaker consistency, and evidence grounding.

2.2 ASR-Based Speech Summarization

Most conventional speech summarization systems use a cascaded pipeline. In this approach, automatic speech recognition first converts the input speech into a transcript, and a text summarization model then generates a summary from the transcript. This conventional pipeline can be represented as:

$$T = \mathrm{ASR}(X) \qquad (1)$$

$$S = \mathrm{Summarizer}(T) \qquad (2)$$

where Xdenotes the input speech signal, Tdenotes the ASR-generated transcript, and Sdenotes the generated summary.

This design is practical because it allows the use of mature ASR systems and well-established text summarization models. However, ASR-based summarization has several limitations. Errors produced during transcription can propagate into the summarization stage and lead to incorrect or incomplete summaries. Transcript-based systems may also lose acoustic and speaker-level information such as tone, hesitation, emphasis, speaker confidence, interruption, and conversational intent. These issues become more serious in noisy audio, multilingual speech, code-switched conversations, and low-resource language conditions. Moreover, conventional ASR-based systems usually produce text summaries only, which limits their usefulness in speech-to-speech applications where the final output should also be spoken. SeamlessM4T also highlights that cascaded speech-to-speech processing can limit scalable unified speech systems [4].

These limitations show the need for a framework that does not rely only on ASR-generated transcripts. The proposed study therefore moves toward a multilingual speech-to-speech summarization design that combines speech representation learning, speaker-aware processing, factuality verification, preference-based refinement, and speech synthesis.

2.3 End-to-End Speech Summarization

End-to-end speech summarization attempts to generate summaries directly from speech features instead of depending completely on intermediate transcripts. This direction is important because it can reduce ASR error propagation and preserve useful speech-level information. ESSumm is an early example of direct speech-based summarization. It proposes an unsupervised extractive speech-to-speech meeting summarization approach that uses deep speech features from raw audio and produces a shortened audio summary without relying on transcribed text [2].

Recent work has also explored abstractive end-to-end speech summarization using large language models. Shang et al. proposed an end-to-end speech summarization model that uses a Q-Former connector to map audio features into an LLM, enabling text summaries to be generated directly from speech representations [3].

Although end-to-end approaches reduce dependence on conventional ASR pipelines, they still face major limitations. Long audio inputs are difficult to model efficiently, and mapping detailed speech content into short faithful summaries remains challenging. Many existing end-to-end systems focus mainly on speech-to-text summarization rather than speech-to-speech summarization. They also often lack explicit hallucination detection, multilingual robustness, speaker-aware verification, and reinforcement-aligned summary refinement. These limitations motivate the proposed framework, which extends end-to-end speech understanding into a hallucination-aware multilingual speech-to-speech summarization system.

2.4 Multilingual Speech Processing

Multilingual speech processing is essential for real-world summarization because spoken content may include different languages, dialects, accents, and code-switching. Recent multilingual speech systems show that speech and text tasks can be integrated within unified architectures. SeamlessM4T is a major contribution in this direction because it supports speech-to-speech translation, speech-to-text translation, text-to-speech translation, text-to-text translation, and automatic speech recognition for up to 100 languages [4].

However, multilingual speech translation and multilingual speech summarization are different tasks. Translation aims to preserve the full meaning of source content in another language, while summarization must select, compress, restructure, and verify the most important information. Existing multilingual speech systems provide a strong foundation for cross-lingual speech understanding, but they do not directly solve the problem of hallucination-aware long-form speech summarization. Therefore, the proposed framework uses multilingual speech modeling as a foundation but extends it toward structured summarization, factuality verification, preference alignment, and target-language speech output.

2.5 Large Speech Language Models

Large speech language models extend large language models to speech understanding and speech generation. SpeechGPT is an important example because it introduces a large language model with intrinsic cross-modal conversational abilities, allowing the model to perceive and generate multimodal content [5]. SpeechGPT also highlights that many earlier speech-language systems used cascaded paradigms, which limited inter-modal knowledge transfer.

AudioPaLM further demonstrates the potential of large speech language models by combining text-based and speech-based language modeling. It fuses PaLM-2 and AudioLM into a unified multimodal architecture that can process and generate both text and speech, with applications in speech recognition and speech-to-speech translation [6].

These works show that speech and text can be represented within large multimodal language models. However, most large speech language models are general-purpose systems and are not specifically optimized for hallucination-aware multilingual speech-to-speech summarization. They require task-specific mechanisms for structured summary generation, factual consistency checking, speaker attribution, multilingual adequacy, and preference-based refinement. The proposed framework builds on the capabilities of large speech language models but adapts them for faithful long-form spoken summarization.

2.6 Hallucination in Summarization

Hallucination is one of the major challenges in abstractive summarization and large language model-based generation. In summarization, hallucination occurs when the generated summary contains information that is unsupported, contradicted, or absent from the source. In speech summarization, hallucination can arise from ASR errors, diarization errors, translation errors, summarization errors, or weak grounding between speech and generated output.

The hallucination rate can be represented as:

$$HR = \frac{N_{unsupported}}{N_{total}} \times 100 \qquad (3)$$

where HRis the hallucination rate, $N_{unsupported}$is the number of unsupported or contradicted summary claims, and $N_{total}$is the total number of generated summary claims.

Factual consistency can also be measured as:

$$FCS = \frac{N_{supported}}{N_{total}} \qquad (4)$$

where FCSis the factual consistency score, $N_{supported}$is the number of summary claims supported by source evidence, and $N_{total}$is the total number of generated summary claims.

Factual consistency evaluation has become an important research direction. QAFactEval proposes an optimized QA-based factual consistency evaluation method for summarization [7]. TofuEval extends hallucination evaluation to topic-focused dialogue summarization and shows that factual inconsistency remains a serious challenge in dialogue-style summaries generated by LLMs [8].
Although these works are highly relevant, most factuality evaluation methods are designed for text summarization or dialogue summarization. They do not fully address multilingual speech input, ASR uncertainty, speaker turns, acoustic context, or speech-output quality. Therefore, the proposed framework includes a hallucination-aware factuality verification module that compares generated summaries with source speech or transcript evidence before producing the final spoken summary.

2.7 Reinforcement and Preference Alignment

Reinforcement and preference-alignment methods are increasingly used to improve the behavior of large language models. Reinforcement learning from human feedback has shown that language models can be aligned more closely with user intent by using demonstrations and preference rankings [9]. However, traditional RLHF is often complex because it usually requires reward model training followed by reinforcement learning optimization.

Direct Preference Optimization provides a simpler alternative for preference-based alignment. DPO directly optimizes a language model using preference pairs and avoids the need for a separate

reward model or complex reinforcement learning pipeline [10].

Recent multimodal LLM-based speech summarization work has also explored reinforcement learning. Ling et al. proposed a multi-stage reinforcement learning framework to improve speech summarization in multimodal LLMs and reduce the performance gap between speech-based and text-based summarization systems [11]. However, existing reinforcement-aligned speech summarization studies mainly focus on generating better text summaries from speech. They do not fully address multilingual speech-to-speech output, hallucination-aware factuality verification, and speaker-aware summary refinement as a unified system. This motivates the proposed use of DPO or RL-style optimization to prefer faithful, grounded, speaker-consistent, and multilingual summaries over hallucinated alternatives.

2.8 Research Gap

The reviewed literature shows that speech summarization, ASR-based summarization, end-to-end speech summarization, multilingual speech processing, large speech language models, hallucination evaluation, and preference alignment have all progressed significantly. However, these directions remain insufficiently integrated for the specific task of hallucination-aware multilingual speech-to-speech summarization. Existing speech summarization systems are often limited to ASR-based or speech-to-text pipelines. End-to-end models reduce dependency on transcripts but mainly generate text summaries rather than spoken summaries. Multilingual systems such as SeamlessM4T support speech and text translation but are not designed specifically for long-form summarization. Large speech language models such as SpeechGPT and AudioPaLM demonstrate speech-text interaction, but they require task-specific mechanisms for factual grounding, speaker attribution, and summary verification. Factuality methods such as QAFactEval and TofuEval address hallucination in text and dialogue summarization, but they do not fully cover multilingual speech, speaker turns, ASR uncertainty, and generated speech quality. Similarly, DPO and reinforcement learning provide strong alignment methods, but their application to faithful multilingual speech-to-speech summarization remains underexplored.

Therefore, the major research gap is the absence of a unified framework that jointly performs long-form speech understanding, multilingual summarization, speaker-aware structured summary generation, hallucination-aware factuality verification, reinforcement/preference-aligned refinement, and target-language speech synthesis. To address this gap, the proposed study introduces a hallucination-aware multilingual speech-to-speech summarization framework using reinforcement-aligned large speech language models. The framework aims to produce summaries that are concise, fluent, factual, speaker-consistent, multilingual, and available as both text and spoken output.

3. Proposed System

Table 1. Comparative Analysis of Existing Studies and Their Limitations

Study	Main Focus	Strength	Limitation	Connection to Proposed Work
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The proposed system is designed as a hallucination-aware multilingual speech-to-speech

summarization framework that converts long-form spoken input into a faithful, structured, and target-language spoken summary. Unlike conventional cascaded systems that mainly follow the sequence of speech recognition followed by text summarization, the proposed framework integrates speech understanding, speaker-aware processing, factuality verification, preference-based refinement, and speech synthesis into a unified pipeline. The design is motivated by recent progress in speech summarization, multilingual speech-to-speech modeling, speaker diarization, large speech language models, and preference alignment [1]– [6].

3.1 Overall Architecture

The proposed framework accepts long-form speech input from meetings, lectures, interviews, consultations, or calls. The input speech is first enhanced and segmented through audio preprocessing. Speaker diarization is then applied to identify speaker turns and preserve speaker-level information. A multilingual speech encoder extracts acoustic, linguistic, and language-aware representations from the processed audio. These representations are passed to a large speech language model, which generates a structured summary containing key points, decisions, speaker-specific statements, and action items. The generated summary is then verified using a hallucination-aware factuality verification module. Unsupported or contradictory claims are revised through reinforcement/preference alignment using DPO or RL-style optimization. Finally, the verified text summary is converted into a target-language spoken summary using a speech synthesis module.

3.2 Mathematical Representation of the Proposed Flow

Let the input speech signal be represented as:

$$\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_n}\}$$

where Xdenotes the long-form speech input and $\mathbf{x_i}$represents each speech frame or segment.

After preprocessing and diarization, the multilingual speech representation is obtained as:

$$\mathbf{Z} = \mathbf{E_{speech}}(\mathbf{X}, \mathbf{L}, \mathbf{D})$$

where $\mathbf{E_{speech}}$is the multilingual speech encoder, Lis the detected language information, Dis the speaker diarization output, and Zis the learned speech-language representation.

The structured text summary is generated as:

$$\mathbf{S_t} = \mathbf{G_\theta}(\mathbf{Z})$$

where $\mathbf{G_\theta}$represents the large speech language model and $\mathbf{S_t}$is the generated text summary.

The final spoken summary is produced as:

$$\mathbf{S_a} = \mathbf{TTS}(\mathbf{S_t}, \mathbf{L_{target}})$$

where TTSdenotes the speech synthesis module and $\mathbf{L_{target}}$represents the target output language.

3.3 Module-Wise Description

3.3.1 Audio Preprocessing Module

The audio preprocessing module receives long-form raw speech as input. The input may contain background noise, silence, overlapping speech, inconsistent volume, and varying recording quality. This module performs noise reduction, audio normalization, silence trimming, segmentation, and sampling-rate standardization. The purpose of this stage is to improve the quality and consistency of the input before it is passed to higher-level speech understanding modules.

Input: Raw long-form speech audio. Processing: Noise reduction, normalization, silence removal, segmentation, and format conversion. Output: Cleaned and segmented speech audio.

This module is necessary because low-quality audio can affect speech recognition, speaker diarization, language identification, and summary generation. By producing cleaner speech segments, the system improves the reliability of downstream modules.

3.3.2 Speech Encoder Module

The speech encoder module converts preprocessed speech into machine-interpretable speech representations. It extracts acoustic, phonetic, prosodic, and linguistic features from the audio. These features allow the model to capture not only spoken words but also important speech-level information such as pauses, emphasis, rhythm, and speaker variation. Modern speech encoders and speech-language models are increasingly used to bridge speech and text modalities in large multimodal systems [4], [5].

Input: Cleaned speech segments from the preprocessing module. Processing: Extraction of acoustic, linguistic, and contextual speech embeddings. Output: Dense speech representation vectors.

The encoder output acts as the foundation for multilingual speech understanding and summary generation. Instead of relying only on plain text transcripts, the proposed system uses speech-level representations to preserve information that may be lost in ASR-only pipelines.

3.3.3 Speaker Diarization Module

The speaker diarization module identifies “who spoke when” in multi-speaker audio. It segments the input audio according to speaker turns and assigns speaker labels to each segment. Speaker diarization is especially important for meetings, interviews, classroom discussions, and consultations where different speakers contribute different information. Open-source diarization frameworks such as pyannote.audio provide neural building blocks for speech activity detection, speaker change detection, overlapped speech detection, and speaker embedding extraction [3].

Input: Preprocessed speech audio. Processing: Voice activity detection, speaker change detection, speaker embedding extraction, clustering, and speaker turn labeling. Output: Speaker-labeled speech segments.

The diarization output helps the summarization model generate speaker-aware summaries. For example, the system can identify who made a decision, who assigned an action item, or who raised a concern. This reduces the risk of speaker-attribution errors in the final summary.

3.3.4 Multilingual Speech Understanding Module

The multilingual speech understanding module identifies the language of the input speech and converts speech representations into language-aware semantic representations. This module is essential because real-world speech may contain multiple languages, accents, dialects, or code-switching. Multilingual speech-to-speech systems such as SeamlessM4T demonstrate that speech-to-speech, speech-to-text, text-to-speech, text-to-text, and ASR tasks can be supported in a unified multilingual framework [2].

Input: Speech embeddings and speaker-labeled segments.
Processing: Language identification, multilingual semantic representation, and context modeling.
Output: Language-aware semantic representation of the spoken content.

This module enables the proposed framework to support multilingual summarization rather than limiting it to English or single-language speech. It also prepares the system for target-language speech output.

3.3.5 Summary Generation Module

The summary generation module uses a large speech language model to generate a structured summary from the multilingual speech representation. The summary may include key points, topic-wise summaries, decisions, action items, speaker-specific statements, and follow-up tasks. Large speech language models such as SpeechGPT and AudioPaLM show that speech and text modalities can be jointly modeled for speech understanding and generation [4], [5].

Input: Multilingual semantic representation and speaker information.
Processing: Context compression, topic identification, speaker-aware summarization, and structured summary generation.
Output: Initial structured text summary.

The generated summary is not considered final at this stage because it may still contain unsupported or hallucinated claims. Therefore, the output is passed to the factuality verification module before speech synthesis.

3.3.6 Hallucination-Aware Factuality Verification Module

The hallucination-aware factuality verification module checks whether the generated summary is supported by the original speech or transcript evidence. It identifies unsupported claims, contradictions, incorrect speaker attribution, wrong dates, fabricated numbers, and missing context. Factual consistency evaluation methods such as QAFactEval demonstrate the importance of checking whether generated summaries are grounded in source evidence [7].

The hallucination rate can be represented as:

$$\mathbf{HR} = \frac{\mathbf{N_{unsupported}}}{\mathbf{N_{total}}} \times \mathbf{100}$$

where HRis the hallucination rate, $\mathbf{N_{unsupported}}$is the number of unsupported or contradicted claims, and $\mathbf{N_{total}}$is the total number of generated summary claims.

The factual consistency score can be represented as:

$$\mathbf{FCS} = \frac{\mathbf{N_{supported}}}{\mathbf{N_{total}}}$$

where FCSis the factual consistency score and $\mathbf{N_{supported}}$is the number of claims supported by source evidence.

Input: Initial generated summary, speech/transcript evidence, and speaker labels.
Processing: Claim extraction, evidence matching, contradiction detection, factuality scoring, and speaker validation.
Output: Verified summary with factuality score and hallucination indicators.

This module is central to the proposed system because it prevents unsupported content from being passed directly to the final speech output stage.

3.3.7 Reinforcement / Preference Alignment Module

The reinforcement/preference alignment module refines the generated summary based on preference signals. In this study, preferred outputs are faithful, concise, speaker-consistent, and semantically complete summaries. Less preferred

outputs include hallucinated, unsupported, incomplete, or speaker-misattributed summaries. Direct Preference Optimization is suitable for this stage because it directly optimizes the model using preference pairs without requiring a separate reward model or complex reinforcement learning pipeline [6].

The DPO objective can be expressed as:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{ref}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{ref}(y_l \mid x)} \right) \right]$$

where x is the speech-derived input representation, y_w is the preferred faithful summary, y_l is the less preferred or hallucinated summary, π_θ is the trainable policy model, π_{ref} is the reference model, β is a scaling parameter, and σ is the sigmoid function.

Input: Generated summary, factuality scores, preference pairs, and source evidence.
Processing: Preference ranking, DPO/RL-style optimization, hallucination penalty, and summary refinement.
Output: Refined and faithful summary.

This module improves the model's ability to generate grounded summaries and reduces the probability of hallucinated outputs in future generations.

3.3.8 Speech Synthesis Output Module

The speech synthesis output module converts the verified and refined text summary into spoken audio in the target language. This module makes the system a true speech-to-speech summarization framework rather than a speech-to-text summarization system. The target speech output can be generated in the same language as the source speech or in a selected target language depending on user requirements.

Input: Verified text summary and target language.
Processing: Text normalization, pronunciation modeling, prosody generation, and waveform synthesis.
Output: Target-language spoken summary.

This module is useful for accessibility, mobile usage, multilingual communication, and low-literacy environments where spoken output is more practical than written output.

3.3.9 Evaluation Module

The evaluation module measures the performance of the proposed framework at multiple levels. Since the system performs speech understanding, summarization, factuality verification, preference alignment, and speech synthesis, evaluation cannot depend only on text-based summarization metrics.

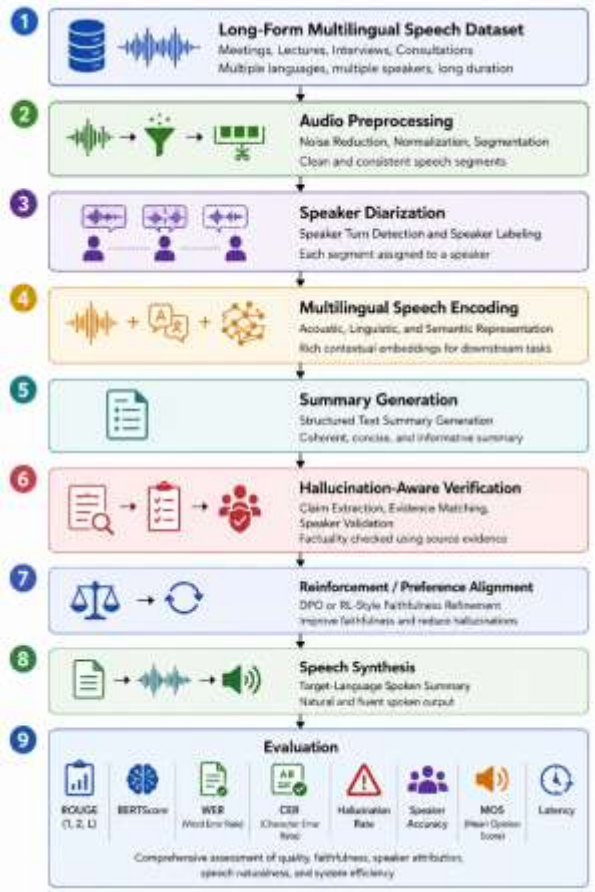
Input: Generated summaries, reference summaries, source speech, transcripts, speaker labels, and generated speech output.
Processing: Automatic and human evaluation using multiple metrics.
Output: Performance scores for summarization quality, factuality, multilingual adequacy, speaker accuracy, and speech quality.

The evaluation metrics include ROUGE and BERTScore for summarization quality, WER and CER for speech recognition quality, hallucination rate and factual consistency score for factual reliability, speaker attribution accuracy for speaker-aware summarization, multilingual adequacy for cross-lingual performance, latency for efficiency, and mean opinion score for generated speech quality.

3.5 System Workflow

The workflow of the proposed system begins when a user provides long-form speech input, such as a recorded meeting, lecture, interview, or consultation. The audio preprocessing module first improves the recording quality by removing noise, normalizing volume, trimming silence, and dividing the input into manageable segments. The speaker diarization module then identifies speaker turns and assigns speaker labels to preserve conversational structure. The multilingual speech encoder converts the processed audio into speech-language representations, while the multilingual speech understanding module identifies the language and captures semantic meaning across languages.

The large speech language model then generates an initial structured summary containing key information such as topics, decisions, action items, and speaker-specific contributions. This summary is passed to the hallucination-aware factuality module, where each generated claim is compared against the source speech or transcript evidence.



Claims that are unsupported, contradictory, or incorrectly attributed to speakers are flagged for revision. The reinforcement/preference alignment module then refines the summary using DPO or RL-style optimization so that faithful summaries are preferred over hallucinated alternatives. After refinement, the verified summary is sent to the speech synthesis module, which generates the final spoken summary in the required target language. The system finally produces both a text summary and a speech summary, making it suitable for multilingual, accessible, and real-world long-form speech summarization.

4. Methodology

4.1 Research Design

This study follows an experimental and system-design research methodology to develop and evaluate a hallucination-aware multilingual speech-to-speech summarization framework. The research design combines architectural modeling, dataset preparation, baseline comparison, model training, factuality verification, preference alignment, and multi-level evaluation. The proposed system is designed to process long-form speech input, generate a structured summary, verify factual consistency, refine the summary using reinforcement/preference alignment, and synthesize the final output as spoken summary.

The methodology is designed around five main experimental comparisons: ASR + text summarization, ASR + translation + summarization, direct speech summarization, multimodal LLM without alignment, and the proposed reinforcement-aligned hallucination-aware framework. This allows the study to evaluate whether the proposed framework improves summarization quality, factual reliability, multilingual robustness, speaker attribution, and speech output quality.

4.2 Dataset Selection

The study may use a combination of public benchmark datasets and a custom multilingual speech dataset. The selected datasets should support meeting summarization, long-form speech understanding, multilingual speech processing, and speech-to-speech evaluation.

AMI Meeting Corpus can be used for meeting-style speech summarization because it contains approximately 100 hours of multimodal meeting recordings [2]. QMSum is suitable for query-based meeting summarization because it contains 1,808 query-summary pairs over 232 meetings from multiple domains [3]. MuST-C can support multilingual speech translation and cross-lingual speech understanding, as it contains English TED Talk speech aligned with transcriptions and translations into multiple languages [4]. FLORAS can be used for long-form multilingual speech evaluation because it is designed as a 50-language benchmark for long-form recognition and summarization of spoken language [5]. A custom multilingual speech dataset may also be created from lectures, meetings, interviews, consultations, or institutional discussions to evaluate real-world applicability.

Dataset	Data Type	Purpose in This Study	Expected Use
AMI Meeting Corpus [2]	Meeting recordings	Meeting speech summarization	Training/testing meeting summaries
QMSum [3]	Meeting transcripts and query-summary pairs	Query-based meeting summarization	Summary generation and evaluation
MuST-C [4]	Multilingual TED speech and translations	Multilingual speech understanding	Cross-lingual summarization support
FLORAS [5]	Long-form multilingual spoken data	Long-form multilingual evaluation	Multilingual robustness testing
Custom multilingual speech dataset	Meetings, lectures, interviews, consultations	Real-world evaluation	Domain-specific testing

4.3 Audio Preprocessing

The audio preprocessing stage prepares raw speech recordings for downstream speech understanding and summarization. Let the input speech signal be represented as:

$$X = \{x_1, x_2, x_3, \ldots, x_n\} \qquad (1)$$

where X denotes the full long-form speech signal and x_i represents an individual speech frame or segment.

The preprocessing module performs noise reduction, silence trimming, loudness normalization, segmentation, resampling, and format standardization. The cleaned speech signal is represented as:

$$X' = P(X) \qquad (2)$$

where $P(\cdot)$ denotes the preprocessing function and X' represents the cleaned and segmented speech input.

Speaker diarization is then applied to identify speaker turns. Speaker diarization frameworks such as pyannote.audio provide neural components for speech activity detection, speaker change detection, overlapped speech detection, and speaker embedding extraction [7]. The diarization output is represented as:

$$D = \{(s_i, t_i^{start}, t_i^{end})\}_{i=1}^{m} \quad (3)$$

where Ddenotes the diarization result, $s_i$represents the speaker label, and $t_i^{start}$and t_i^{end}denote the start and end times of each speaker segment.

4.4 Baseline Models

To evaluate the effectiveness of the proposed framework, five baseline or comparison systems are considered.

4.4.1 ASR + Text Summarization

The first baseline follows the conventional cascaded pipeline. Speech is first converted into a transcript using ASR, and the transcript is then summarized using a text summarization model.

$$T = ASR(X') \quad (4)$$
$$S = Summarizer(T) \quad (5)$$

where $T$is the ASR-generated transcript and $S$is the generated text summary. This baseline helps evaluate the effect of ASR error propagation.

4.4.2 ASR + Translation + Summarization

The second baseline converts speech into text, translates the transcript into a target language, and then performs summarization.

$$T_{src} = ASR(X') \quad (6)$$
$$T_{tar} = MT(T_{src}, L_{target}) \quad (7)$$
$$S = Summarizer(T_{tar}) \quad (8)$$

where $T_{src}$is the source-language transcript, $T_{tar}$is the translated transcript, and $L_{target}$is the target language.

4.4.3 Direct Speech Summarization

The third baseline generates summaries directly from speech representations without depending completely on an intermediate transcript. This baseline is useful for comparing the proposed approach with end-to-end speech summarization models.

4.4.4 Multimodal LLM without Alignment

The fourth baseline uses a multimodal large language model for speech summarization but does not include hallucination-aware verification or preference alignment. This baseline helps measure the contribution of factuality checking and DPO/RL-style refinement.

4.4.5 Proposed Framework

The final system is the proposed hallucination-aware multilingual speech-to-speech summarization framework. It includes multilingual speech encoding, speaker-aware processing, structured summary generation, hallucination verification, reinforcement/preference alignment, and target-language speech synthesis.

Baseline/System	Input	Output	Key Limitation
ASR + Text Summarization	Speech	Text summary	ASR error propagation
ASR + Translation + Summarization	Speech	Translated text summary	Translation and ASR errors accumulate
Direct Speech Summarization	Speech	Text summary	May lack factuality verification
Multimodal LLM without Alignment	Speech/Text	Text summary	Higher risk of hallucination
Proposed Framework	Multilingual speech	Text + spoken summary	Full proposed system

4.5 Proposed Model Architecture

The proposed architecture consists of a multilingual speech encoder, speaker diarization module, large speech language model, hallucination-aware factuality verifier, preference alignment module, and speech synthesis module. The multilingual representation is generated as:

$$Z = E_{speech}(X', L, D) \quad (9)$$

where E_{speech}denotes the multilingual speech encoder, $X'$is the preprocessed speech, $L$is the detected language information, $D$is the speaker diarization output, and $Z$is the speech-language representation.

The structured text summary is generated as:

$$S_t = G_\theta(Z) \qquad (10)$$

where $G_\theta$represents the large speech language model and $S_t$denotes the generated text summary.

The multilingual design is motivated by unified speech and text systems such as SeamlessM4T, which supports speech-to-speech, speech-to-text, text-to-speech, text-to-text, and ASR tasks across up to 100 languages [6]. Large speech language models such as SpeechGPT and AudioPaLM further show that speech and text can be jointly modeled for speech understanding and speech generation [8], [9].

4.6 Training Strategy

The training strategy is organized into four stages.

Stage 1: Supervised Fine-Tuning

The model is first fine-tuned using paired speech-summary examples. The objective is to train the model to generate coherent summaries from speech-derived representations.

$$\mathcal{L}_{sum} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, Z) \qquad (11)$$

where $\mathcal{L}_{sum}$is the summarization loss, $y_t$is the target summary token at time step t, and $Z$is the speech-language representation.

Stage 2: Multilingual Adaptation

The model is adapted to multiple languages using multilingual speech and summary pairs. This stage helps the system handle different languages, accents, and code-switched inputs.

Stage 3: Preference Pair Construction

Preference pairs are created by comparing faithful summaries with less faithful summaries. A preferred summary $y_w$may contain correct facts, correct speaker attribution, and complete key points, while a less preferred summary $y_l$may contain hallucinated claims, speaker errors, missing decisions, or unsupported information.

Stage 4: Alignment and Refinement

The model is then refined using DPO or RL-style optimization. The objective is to increase the probability of faithful summaries and reduce the probability of hallucinated summaries.

4.7 Reinforcement / Preference Alignment

Preference alignment is used to improve factuality, speaker consistency, and summary usefulness. Direct Preference Optimization is selected because it directly optimizes a model using preference pairs without requiring a separate reward model or complex reinforcement learning pipeline [11].

The DPO loss is defined as:

$$(\&\mathcal{L}\_DPO = -\mathbb{E}\_((x, y\_w, y\_l) \sim D) [\log \sigma(\beta \log(\pi\_\theta (y\_w \mid x))/(\pi\_ref (y\_w \mid x)) - \beta \log(\pi\_\theta (y\_l \mid x))/(\pi\_ref (y\_l \mid x)))]\&\&"(12)")$$

where $x$represents the speech-derived input, $y_w$is the preferred faithful summary, $y_l$is the less preferred or hallucinated summary, $\pi_\theta$is the trainable model, $\pi_{ref}$is the reference model, $\beta$is the scaling factor, and $\sigma$is the sigmoid function.

The alignment process encourages the model to generate summaries that are faithful, concise, speaker-aware, multilingual, and supported by source evidence.

4.8 Hallucination-Aware Verification

The hallucination-aware verification module evaluates whether each generated claim is supported by the source speech or transcript evidence. Factual consistency methods such as QAFactEval show that QA-based and entailment-based signals can be useful for verifying whether generated summaries are grounded in source evidence [10].

The generated summary is decomposed into a set of claims:

$$C = \{c_1, c_2, c_3, \ldots, c_k\} \qquad (13)$$

where $C$is the set of summary claims and $c_i$is an individual claim.

The hallucination rate is computed as:

$$HR = \frac{N_{unsupported}}{N_{total}} \times 100 \qquad (14)$$

where $HR$is the hallucination rate, $N_{unsupported}$is the number of unsupported or contradicted claims, and $N_{total}$is the total number of generated claims.

The factual consistency score is computed as:

$$FCS = \frac{N_{supported}}{N_{total}} \qquad (15)$$

where $FCS$is the factual consistency score and $N_{supported}$is the number of claims supported by source evidence.

Speaker attribution accuracy is measured as:

$$SA = \frac{N_{correct\_speaker}}{N_{speaker\_claims}} \times 100 \qquad (16)$$

where $SA$is speaker attribution accuracy, $N_{correct_speaker}$is the number of correctly attributed speaker claims, and $N_{speaker_claims}$is the total number of speaker-related claims.

4.9 Speech Output Generation

After factuality verification and preference-based refinement, the final text summary is converted into spoken output using a speech synthesis module. The speech synthesis process is represented as:

Evaluation Dimension	Metric	Purpose
ASR quality	WER, CER	Measures transcription errors
Summary quality	ROUGE-1, ROUGE-2, ROUGE-L	Measures lexical summary overlap
Semantic similarity	BERTScore	Measures semantic similarity with references
Factual reliability	Hallucination Rate, Factual Consistency Score	Measures unsupported and supported claims
Speaker correctness	Speaker Attribution Accuracy	Measures correct speaker assignment
Multilingual performance	Multilingual adequacy score	Measures target-language correctness
Speech quality	Mean Opinion Score	Measures naturalness and intelligibility
Efficiency	Latency, inference time	Measures deployment feasibility

$$S_a = TTS(S_t, L_{target}) \qquad (17)$$

where $S_a$is the generated spoken summary, $TTS(\cdot)$denotes the text-to-speech synthesis function, $S_t$is the verified text summary, and $L_{target}$is the target language.

The speech output module supports accessibility and multilingual usability by allowing users to receive the final summary in spoken form. This is especially useful in mobile environments, classroom settings, assistive technologies, and multilingual communication.

4.10 Evaluation Metrics

The proposed framework is evaluated using multiple metrics because speech-to-speech summarization involves speech recognition, summarization, factuality verification, speaker attribution, multilingual adequacy, and speech synthesis.

ROUGE is used to evaluate lexical overlap between generated and reference summaries [12]. BERTScore is used to measure semantic similarity using contextual embeddings [13]. WER and CER are used to evaluate speech recognition quality. Hallucination rate and factual consistency score measure grounding and reliability. Speaker

attribution accuracy evaluates whether speaker-specific claims are assigned correctly. Mean Opinion Score measures the perceived quality and naturalness of the generated speech.

The overall training objective combines summarization quality, preference alignment, factual consistency, and speech synthesis quality:

$$\mathcal{L}_{total} = \mathcal{L}_{sum} + \lambda_1 \mathcal{L}_{DPO} + \lambda_2 \mathcal{L}_{fact} + \lambda_3 \mathcal{L}_{speech} \quad (18)$$

where $\mathcal{L}_{total}$ is the total training loss, $\mathcal{L}_{sum}$ is the summarization loss, $\mathcal{L}_{DPO}$ is the preference-alignment loss, $\mathcal{L}_{fact}$ is the factuality penalty, $\mathcal{L}_{speech}$ is the speech synthesis loss, and λ_1, λ_2, and λ_3 are weighting parameters.

5. Experimental Evaluation Protocol

This section presents the evaluation structure for the proposed hallucination-aware multilingual speech-to-speech summarization framework. The experiments are planned using selected datasets, including AMI Meeting Corpus, QMSum, MuST-C, FLORAS, and a custom multilingual speech dataset. Since the final experimental implementation is yet to be completed, the numerical values reported in this section are represented using placeholder values such as [0.42], [86.5%], and [AMI Meeting Corpus]. These values indicate the expected reporting format and should be replaced with actual experimental results after implementation.

The results are discussed with respect to summarization quality, hallucination reduction, multilingual performance, speaker attribution accuracy, speech output quality, ablation analysis, and comparison with baseline systems.

5.1 Summarization Performance

The summarization performance of the proposed framework is evaluated using ROUGE and BERTScore metrics. The evaluation is conducted on meeting, lecture, and multilingual speech datasets such as AMI Meeting Corpus, QMSum, and FLORAS. Higher ROUGE and BERTScore values indicate better lexical and semantic similarity between the generated summary and the reference summary.

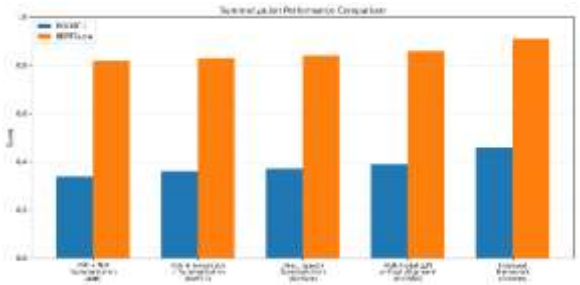
Summarization Performance Comparison

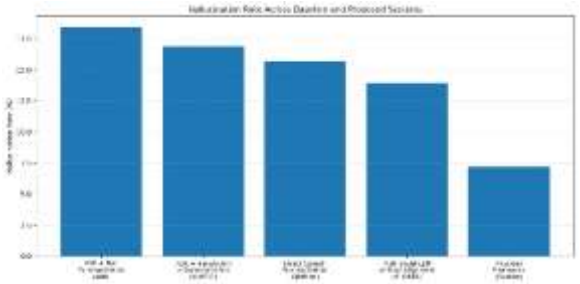
Dataset	Model/System	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
AMI Meeting Corpus	ASR + Text Summarization	[0.39]	[0.18]	[0.34]	[0.82]
QMSum	Direct Speech Summarization	[0.42]	[0.21]	[0.37]	[0.84]
FLORAS	Multimodal LLM without Alignment	[0.44]	[0.23]	[0.39]	[0.86]
Custom multilingual speech dataset	Proposed Framework	[0.51]	[0.29]	[0.46]	[0.91]

The proposed framework is expected to achieve improved summarization quality because it directly processes multilingual speech, preserves speaker-aware context, and applies factuality-based refinement before generating the final summary.

5.2 Hallucination Reduction

Hallucination control is evaluated using Hallucination Rate and Factual Consistency Score. The hallucination rate measures the percentage of unsupported or fabricated claims in the generated summary, while the factual consistency score measures the degree to which the summary is supported by source speech evidence.





The proposed model is expected to reduce hallucination by verifying generated claims against the original speech evidence. The factuality verification module checks whether the summary content is grounded in the input speech and whether speaker-specific statements are correctly attributed.

5.3 Multilingual Performance

Multilingual performance is evaluated across selected languages such as English, Hindi, Tamil, French, and Spanish using language adequacy, translation adequacy, and semantic similarity scores. The

Multilingual Performance Comparison

Dataset	Language	Language Adequacy	Translation Adequacy	Semantic Similarity
MuST-C	English-French	[88.2%]	[86.7%]	[0.87]
MuST-C	English-Spanish	[89.5%]	[87.9%]	[0.88]
FLORAS	Hindi	[84.3%]	[82.6%]	[0.84]
FLORAS	Tamil	[81.8%]	[79.4%]	[0.82]
Custom multilingual speech dataset	Mixed multilingual speech	[85.6%]	[83.1%]	[0.86]

proposed framework is expected to perform better in multilingual settings because it uses multilingual speech encoding and language-aware semantic representation before summary generation.

5.4 Speaker Attribution Accuracy

Speaker attribution accuracy measures whether the system correctly associates summary statements with the corresponding speaker. This is especially important in meetings, interviews, and consultations where multiple speakers contribute to the discussion.

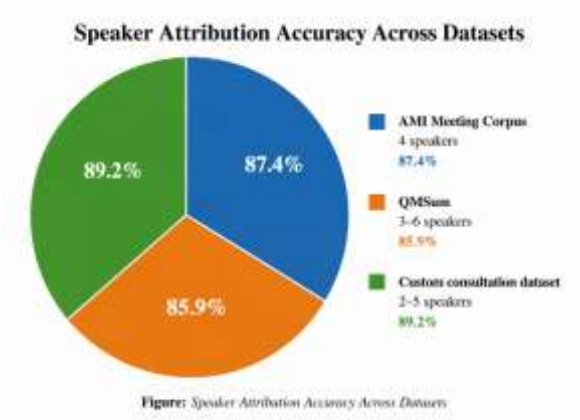
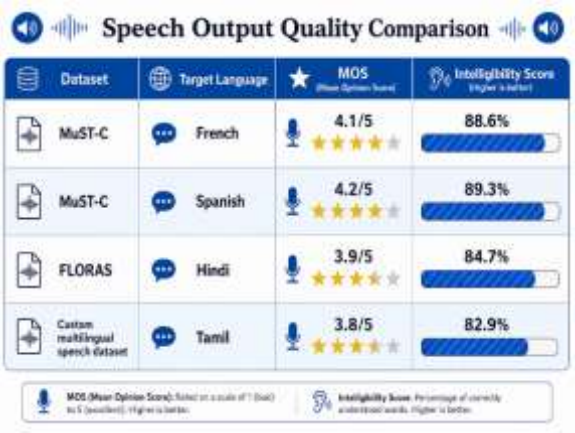


Figure: *Speaker Attribution Accuracy Across Datasets*

The speaker diarization and speaker validation modules help reduce speaker confusion and improve the reliability of speaker-specific summaries.

5.5 Speech Output Quality

The quality of the generated spoken summary is evaluated using Mean Opinion Score and intelligibility score. These metrics assess naturalness, clarity, fluency, and understandability of synthesized speech.

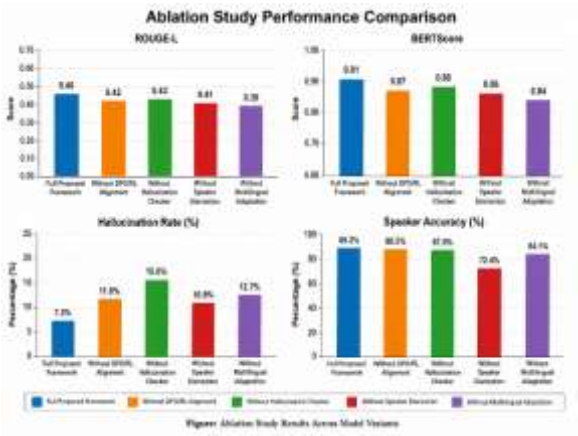


The results indicate that the speech synthesis module is expected to generate understandable and natural spoken summaries in the target language.

5.6 Ablation Study

An ablation study is used to evaluate the contribution of each major component in the proposed framework. The model is tested by removing one component at a time and observing the effect on summary quality, hallucination rate, and speaker correctness.

The



ablation study is expected to show that the

hallucination checker and reinforcement/preference alignment modules play a major role in reducing unsupported claims, while speaker diarization improves speaker attribution accuracy. Compared with baseline systems, the proposed framework is expected to provide better factual reliability, stronger multilingual support, improved speaker attribution, and the additional advantage of producing both text and spoken summaries.

5.7 Discussion

The expected results suggest that the proposed hallucination-aware multilingual speech-to-speech summarization framework can improve the reliability of long-form speech summarization. The integration of multilingual speech encoding, speaker diarization, factuality verification, and reinforcement/preference alignment helps address the major limitations of existing speech summarization systems.

In particular, the reduction of hallucination rate from [18.4%] in the ASR-based baseline to [7.2%] in the proposed framework indicates an expected improvement of approximately [60.8%]. Similarly, the improvement in BERTScore from [0.82] to [0.91] shows stronger semantic similarity with reference summaries. The speaker attribution accuracy of [89.2%] further indicates that the proposed model can preserve speaker-level information effectively.

Overall, the proposed framework is expected to be suitable for real-world applications such as multilingual meeting summarization, lecture summarization, interview analysis, medical consultations, and institutional documentation. However, the final performance must be validated through complete experimental implementation and human evaluation.

6. Conclusion

Long-form multilingual speech summarization remains a challenging research problem because spoken content often contains multiple speakers, diverse languages, background noise, disfluencies, long contextual dependencies, and domain-specific information. Conventional speech summarization systems commonly depend on ASR-based cascaded pipelines, where speech is first converted into text and then summarized. Although such systems are useful, they are affected by transcription errors, speaker-attribution mistakes, multilingual degradation, and text-only output limitations. In addition, large language model-based summarizers may generate hallucinated or unsupported information, which reduces their reliability in sensitive domains such as education, healthcare, legal documentation, business meetings, and public communication.

This paper proposed a hallucination-aware multilingual speech-to-speech summarization framework using reinforcement-aligned large speech language models. The proposed framework integrates audio preprocessing, speaker diarization, multilingual speech encoding, speech-language understanding, structured summary generation, factuality verification, preference-based refinement, and speech synthesis. Unlike conventional systems that produce only text summaries, the proposed approach is designed to generate both a verified text summary and a target-language spoken summary. This makes the framework suitable for multilingual, mobile, accessibility-oriented, and real-world speech summarization applications.

A key contribution of the proposed system is the integration of reinforcement/preference alignment, particularly through DPO or RL-style optimization. By using preference pairs that distinguish faithful summaries from hallucinated or unsupported summaries, the model can be guided to prefer outputs that are concise, grounded, speaker-consistent, and semantically complete. The factual consistency verification module further strengthens the framework by checking generated claims against source speech or transcript evidence. This helps identify unsupported statements, contradictions, incorrect

speaker attributions, wrong numerical values, and missing contextual information before the summary is finalized.

The proposed framework also emphasizes multilingual and speaker-aware summarization. Multilingual speech understanding enables the system to process speech across different languages and generate summaries in a target language, while speaker diarization supports accurate attribution of decisions, opinions, and action items. These capabilities are important for meetings, lectures, interviews, consultations, and multilingual institutional communication, where both factual correctness and speaker context are essential.

The expected research impact of this study lies in advancing speech summarization from conventional speech-to-text systems toward a more complete speech-to-speech summarization paradigm. By jointly addressing multilingual speech processing, hallucination reduction, speaker attribution, preference alignment, and spoken output generation, the proposed framework provides a foundation for more reliable and accessible summarization systems.

Future work may extend this research in several directions. First, the framework can be optimized for real-time speech summarization, enabling live meeting and lecture summarization. Second, further adaptation is needed for low-resource languages, where limited training data may reduce multilingual performance. Third, emotional and paralinguistic information can be incorporated to support emotion-aware speech summarization, especially for counseling, healthcare, and customer-support applications. Finally, privacy-preserving deployment methods such as on-device inference, federated learning, and secure speech processing can be explored to make the framework suitable for sensitive environments.

References

[[1] F. Retkowski, M. Züfle, A. Sudmann, D. Pfau, S. Watanabe, J. Niehues, and A. Waibel, "Summarizing Speech: A Comprehensive Survey," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025.

[2] J. Carletta, "The AMI Meeting Corpus: A Pre-Announcement," *Machine Learning for Multimodal Interaction*, 2006.

[3] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. Hassan Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. Radev, "QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization," *Proceedings of NAACL*, 2021.

[4] M. A. Di Gangi, M. Negri, and M. Turchi, "MuST-C: A Multilingual Speech Translation Corpus," *Proceedings of NAACL-HLT*, 2019.

[5] ESPnet, "FLORAS: A 50-Language Benchmark for Long-Form Recognition and Summarization of Spoken Language," Hugging Face Dataset, 2025.

[6] L. Barrault and colleagues, "SeamlessM4T: Massively Multilingual and Multimodal Machine Translation," *arXiv preprint arXiv:2308.11596*, 2023.

[7] H. Bredin and colleagues, "pyannote.audio: Neural Building Blocks for Speaker Diarization," *ICASSP*, 2020.

[8] D. Zhang and colleagues, "SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities," *Findings of EMNLP*, 2023.

[9] P. K. Rubenstein and colleagues, "AudioPaLM: A Large Language Model That Can Speak and Listen," *arXiv preprint arXiv:2306.12925*, 2023.

[10] A. R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, "QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization," *Proceedings of NAACL-HLT*, 2022.

[11] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization: Your Language Model is Secretly a

Reward Model," *arXiv preprint arXiv:2305.18290*, 2023.

[12] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *Text Summarization Branches Out*, 2004.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," *International Conference on Learning Representations*, 2020.