International Journal of Computer Science Engineering Techniques – Volume 9 Issue 6, November - December - 2025

Speak2Summarize: A Whisper-Driven NLP Framework For Real-Time Transcription and Summarization

Divyatha M[1], Gagana C S[2], Harshini S[3], Lavitha P[4], Dr. Vishwesh J[5]

[1] Department of Computer Science & Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, India, (Orcid ID: https://orcid.org/0009-0002-5136-5408)
[2] Department of Computer Science & Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, India, (Orcid ID: https://orcid.org/0009-0007-4537-7652)
[3] Department of Computer Science & Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, India, (Orcid ID: https://orcid.org/0009-0000-9973-5979)
[4] Department of Computer Science & Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, India, (Orcid ID: https://orcid.org/0009-0000-8020-1686)
[5] Associate Professor, Department of Computer Science & Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, India, (Orcid ID: https://orcid.org/0000-0003-2112-3512)

Email address: divyatha1258@gmail.com (Divyatha M), gaganacs44@gmail.com (Gagana C S), cs814409@gmail.com (Harshini S), plavitha8@gmail.com (Lavitha P), *vishweshj@gsss.edu.in (Dr. Vishwesh J)*

**Corresponding Author: Dr. Vishwesh J , email: vishweshj@gsss.edu.in*

Received: 18/Nov/2024; **Accepted**: 20/Dec/2024; **Published**: 31/Jan/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i1.14

Abstract: Speak2Summarize is a real-time speech processing system designed to convert lengthy audio and video content into clear and concise summaries. The system integrates automated speech recognition, noise reduction, translation, and natural language processing to help users quickly understand long lectures, meetings, and online videos without manually taking notes. It accepts different input formats, including uploaded files, YouTube links, and live recordings, and processes them through a streamlined pipeline that extracts audio, enhances clarity, and converts speech into text.

Using a Whisper-based transcription engine, the system produces accurate text even when the audio contains disturbances or mixed accents. A language detection module identifies the spoken language, and when needed, the transcript is translated before generating summaries. Speak2Summarize offers multiple summary styles to match the user's preference, such as brief overviews, structured paragraphs, and point-wise summaries. The interface is built to be simple and intuitive, allowing users to upload content, view the generated text, and download the summarized output in an organized report.

This system reduces the time required to review long content and provides a practical solution for students, educators, professionals, and anyone who regularly works with spoken information. By automating transcription and summarization, Speak2Summarize improves productivity, enhances accessibility, and simplifies the process of understanding lengthy audio-visual material.

Keywords: Real-Time Transcription, Speech Recognition, Whisper Model, Natural Language Processing, Audio Preprocessing, Text Summarization, PDF Report Generation, Video-to-Text Conversion

1. Introduction

The rapid expansion of digital audio–video communication in academic, corporate, and media domains has increased the need for intelligent systems capable of converting spoken content into clear and structured textual information. As lectures, meetings, and online discussions continue to grow in volume, users often face challenges in manually transcribing and summarizing long recordings. Although prior research has explored transcription, summarization, noise handling, multimodal processing, and real-time analysis, these components are commonly implemented in isolation, leaving a critical gap in the availability of a single unified platform capable of handling all tasks seamlessly.

Recent advancements demonstrate the effectiveness of deep learning architectures such as CNNs, LSTMs, and ResNet-based models in improving video summarization quality and identifying important content patterns [1]. Parallel studies highlight the accuracy of Whisper-based transcription pipelines and the usefulness of LLM-assisted summarization for generating more coherent and readable summaries [2]. Comprehensive surveys covering extractive, abstractive,

multimodal, reinforcement-learning, and GAN-based summarization techniques emphasize the need for adaptable, user-controlled summary outputs that can serve diverse documentation needs [3].

Further progress in multimedia analysis introduces techniques such as shot segmentation, clustering, and audio–visual fusion to generate richer, context-aware summaries [4]. Research on real-time summarization demonstrates increasing demand for fast, accurate ASR and summarization models in dynamic environments [5]. Additional studies explore DBN-based summarization and RBF-based retrieval mechanisms for enhancing semantic representation and retrieval accuracy [6]. Multimodal summarization approaches integrating text, audio, and visual cues also show promising improvements in summary structure and flexibility aligned with user needs [7]. However, despite these advancements, most tools still lack a unified interface that combines transcription, noise reduction, multi-format summarization, and instant document export in a single workflow [8].

To address this gap, the present study introduces **Speak2Summarize**, an integrated framework that performs bilingual transcription in Hindi and English, applies basic noise reduction, generates summaries in Concise, Structured, and Bullet-Point formats, and enables one-click PDF export. The system improves accessibility, supports efficient documentation, and enhances user experience across various applications, including educational content review [9], meeting documentation [12], [19], and media content analysis [1], [16]. By consolidating essential components into one platform, the proposed framework provides a practical, scalable solution that strengthens modern digital information processing workflows.

1.1 Background of the Study

Digital learning platforms, remote work, and virtual conferences have dramatically increased the volume of recorded spoken material. Manual transcription and summarization consume considerable time and are prone to omission and inconsistency. Modern ASR and NLP models have improved accuracy, yet practical deployment often requires integrating preprocessing (noise reduction, normalization), robust transcription, translation, summarization, and export functionalities into a single pipeline. The Speak2Summarize project builds on these technological foundations to provide a usable, end-to-end tool that addresses the workflow fragmentation observed in current solutions.

1.2 Objectives and Motivation

The primary objective of this work is to design and implement a functional framework that automates speech-to-text conversion and produces readable summaries in multiple formats. Specific aims are: (1) implement real-time speech-to-text conversion using Faster-Whisper; (2) integrate a lightweight denoising module to improve transcription accuracy; (3) enable language detection and optional translation between Hindi and English; and (4) generate concise, structured, and bullet-point summaries with PDF export capability. The motivation stems from the practical need to reduce manual note-taking effort, improve information accessibility, and provide a scalable solution suitable for lectures, meetings, and media analysis.

2. Related Work

Several research studies have explored automated transcription, video summarization, multimodal analysis, and real-time content processing. This section highlights the most relevant works, focusing on each study's title, core problem, and primary objectives.

[1] AI-Driven Video Summarization for Optimizing Content Retrieval and Management

This work addresses the difficulty of handling large volumes of video data and extracting meaningful highlights. The authors aim to improve video summarization using deep learning models such as CNNs and LSTMs to enhance retrieval efficiency.

[2] AI-Powered Video Summarization and Transcription System

This study focuses on the challenge of generating accurate transcripts and summaries from long videos. Its objective is to integrate transcription and summarization to support educational and media workflows.

[3] Video Summarization Techniques: A Comprehensive Review

The authors identify limitations in existing extractive and abstractive summarization methods. Their objective is to provide a structured review of multimodal, GAN-based, and reinforcement learning summarization approaches, emphasizing adaptability and user-controlled output.

[4] Automated Video Summarization through Advanced Multimedia Analysis

This research highlights the need for multimedia-driven summarization. It addresses problems such as inefficient scene detection and aims to use advanced techniques like shot segmentation and clustering to refine summary quality.

[5] AI-Powered Real-Time Video Summarization

This study targets the increasing demand for real-time content processing. Its objective is to design a system capable of producing summaries instantly during live sessions, supporting fast decision-making environments.

[6] AI-Based Video Summarization for Efficient Content Retrieval

The work addresses slow information retrieval from long video recordings. Its goal is to apply machine learning and deep feature extraction to create summaries that accelerate content search and review.

[7] Multimodal Video Content Summarization Using Machine Learning

This study highlights the problem of relying only on single-modal inputs like text or visuals. The objective is to use multimodal fusion—combining audio, text, and frames—to produce more meaningful summaries.

[8] Cross-Modal LMM for Video and Audio Analysis

This research points out challenges in linking audio and visual cues. The objective is to apply cross-modal large multimodal models (LMMs) to better align audio–visual features for improved analysis and summarization.

[9] GlobalLearn: Multilingual Textual Summary for Educational Videos
The problem identified is the difficulty students face in understanding lengthy monolingual educational videos. The objective is to provide multilingual summaries that support diverse learners.
[11] Real-Time Summarization for Dynamic Media Streams
This study focuses on the challenge of summarizing constantly changing content in real-time environments. Its objective is to implement efficient ASR and abstraction models for live processing.
[12] Meeting Summarization Using Neural Language Models
This research identifies the problem of generating accurate meeting minutes. The objective is to use neural summarization models to capture key points from long discussions.

3. Theory/Calculation

This section presents the theoretical foundations behind the core modules of the *Speak2Summarize* system and connects them to the practical calculations implemented during processing. While earlier sections introduce the motivation and workflow, the discussion here focuses on the mathematical principles, algorithmic logic, and quantitative operations that guide audio preprocessing, transcription, translation, and summarization.

3.1 Audio Signal Theory and Denoising Calculations

Before speech can be transcribed, the raw waveform must be smoothed to suppress distortions and high-frequency noise. The system relies on *deterministic signal-processing theory* rather than machine-learned filters, ensuring predictable and real-time behaviour.
A digital audio signal can be expressed as a discrete sequence:

$$x[n], n = 0,1,2, \dots, N-1$$

The moving-average filter applied in the pipeline functions as a low-pass filter. Each output value $y[n]$ is the arithmetic mean of the previous M samples:

$$y[n] = \frac{1}{M}\sum_{k=0}^{M-1} x[n-k]$$

This smoothing suppresses short-duration fluctuations (e.g., clicks, hiss) while leaving phoneme-level variations intact.
The choice of M is tied directly to the sample rate. For example, using a 3 ms window at 16 kHz sampling frequency:

$$M = 0.003 \times 16000 \approx 48 \text{ samples}$$

After denoising, the waveform is normalized using the Root Mean Square (RMS) measure:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} y[i]^2}$$

The signal is scaled so that the RMS attains a consistent target amplitude, improving the downstream model's robustness to variations in microphone volumes.

3.2 Spectral Theory in Whisper-Based ASR

Automatic Speech Recognition (ASR) relies on converting time-domain audio into a frequency-domain representation. The Whisper-derived model uses log-mel spectrograms computed as:

$$S(t,f) = \log\left(\sum_{k=1}^{K} |X_k|^2 H_f(k)\right)$$

where:

- X_k is the Short-Time Fourier Transform (STFT),
- $H_f(k)$ are mel-filterbank coefficients,
- t is time frame index,
- f is mel frequency bin.

The transformer-based encoder processes these spectral embeddings to learn contextual dependencies between phonetic frames. The decoder predicts text tokens via probability distribution:

$$P(w_t \mid w_{<t}, S) = \text{Softmax}(W \cdot h_t)$$

where:

- h_t is the decoder hidden state at step t,
- W is the output projection matrix.

This theoretical foundation enables multilingual transcription with consistent accuracy across varied accents and noisy conditions.

3.3 Language Detection and Translation Rule-Based Validation

The system integrates statistical language detection to classify transcripts. Once detected, text may be translated using transformer-based Marian MT models.
To ensure translated text maintains semantic completeness, a length-based verification rule is applied:

$$\frac{L_t}{L_o} < \tau \Rightarrow \text{Re-translate}$$

Where:

- L_o = token count of original text,
- L_t = token count of translated text,

- τ = threshold (typically 0.35–0.40).

This proportionality check prevents truncated translations, ensuring the subsequent summarizer receives coherent input.

3.4 Prompt-Driven Summarization Theory

Summarization is based on abstractive encoder–decoder theory, where the BART-CNN model internally learns latent semantic structures. Given text T and a summary style s, the prompt function is:

$$\text{Prompt} = f_s(T)$$

The encoder compresses text into semantic vector space:

$$Z = \text{Encoder}(T)$$

The decoder reconstructs a shorter formulation:

$$\hat{T} = \text{Decoder}(Z)$$

The abstractive nature allows the system to generate novel phrasing instead of extracting sentences verbatim. Output length constraints—typically max/min token limits—are treated as boundary conditions in the decoding process.

3.5 Embedding-Based Retrieval Theory

When storing summaries, the system can compute vector embeddings for semantic search. The similarity between a query vector q and stored embedding v_i is determined via cosine similarity:

$$\text{score}_i = \frac{q \cdot v_i}{\| q \| \| v_i \|}$$

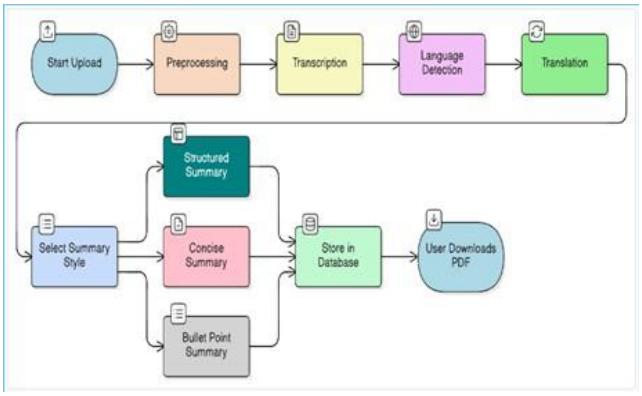
This calculation ranks summaries from most to least relevant, supporting future retrieval and knowledge indexing within the system.

3.6 PDF Generation as a Deterministic Formatting Process

Unlike the machine-learning-driven stages, report generation follows deterministic layout logic. Metadata, summaries, and translations are arranged according to fixed coordinate calculations for line spacing, page boundaries, and content flow. This ensures consistent, user-friendly documentation across outputs.

4. Experimental Method/Procedure/Design

The proposed system follows a structured experimental methodology that integrates audio preprocessing, transformer-based speech recognition, Hindi-to-English processing, and neural summarization into a unified workflow. The design of the system emphasizes modularity, real-time performance, and robustness against variations in audio quality, recording environments, and language characteristics.



4.1 Input Acquisition

The process begins with the acquisition of audio or video content through three supported modes: local file upload, YouTube link extraction, and live microphone recording. All incoming inputs are converted into a standardized WAV format with a fixed sampling rate and single audio channel. This normalization ensures consistency and prevents format-specific errors during downstream processing.

4.2 Preprocessing and Noise Reduction

To enhance transcription accuracy, the system applies a lightweight preprocessing stage that includes moving-average–based denoising and RMS volume normalization. The denoiser smooths short-term fluctuations, suppressing high-frequency noise while preserving important speech characteristics. The normalization step aligns loudness levels across different recordings, providing a stable acoustic input for the ASR model.

4.3 Speech Transcription Using Faster-Whisper

The preprocessed audio is transcribed using the Faster-Whisper model, an optimized implementation of the Whisper transformer architecture. The model converts the waveform into log-mel spectrograms and performs encoder–decoder inference to generate text tokens. It supports Hindi or Engligh inputs, handles variations in accent and noise, and maintains low latency even without GPU acceleration, making it suitable for real-time use.

4.4 Language Detection and Translation

After transcription, the text is processed by a language detection module to identify the dominant language. When non-English content is detected, the system translates the transcript using a MarianMT-based transformer model. A length-ratio verification rule is applied to ensure that the translated text is complete and not partially generated. Translations failing this check are reprocessed to maintain semantic integrity.

4.5 Abstractive Summarization

The verified transcript is passed to the BART-Large-CNN summarization model, which generates a concise representation of the content. The system supports three

summary styles—Concise, Structured, and Bullet-Point—each guided by a specific prompt template. The abstractive nature of the model allows it to generate coherent summaries that capture contextual meaning rather than extracting specific sentences from the transcript.

4.6 Storage and Semantic Retrieval

All transcripts, translations, and summaries are stored in a SQLite-based database along with associated metadata. Optional vector embeddings may be generated for each summary to support semantic retrieval using cosine similarity. This enables efficient search and organization of historical outputs, making the system suitable for academic and professional documentation tasks.

4.7 Output Presentation and PDF Generation

After processing, the results are displayed through an interactive interface that presents the transcript, summary formats, and audio preprocessing visualizations. Users may also download the final output as a formatted PDF containing summaries, translations, and relevant metadata. The PDF generation process follows deterministic layout rules to ensure clarity and professional appearance.

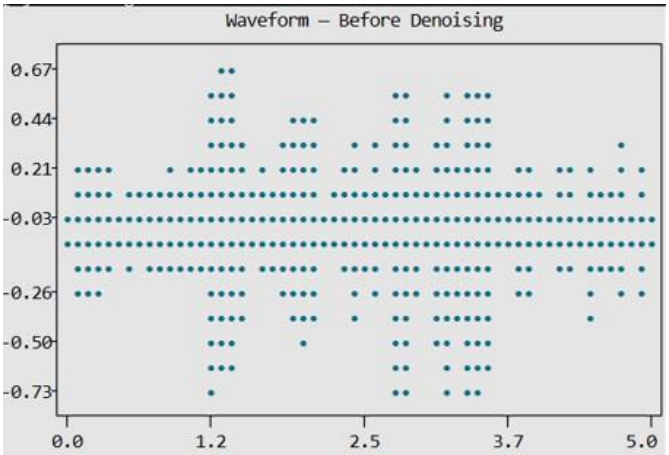
5. Results and Discussion

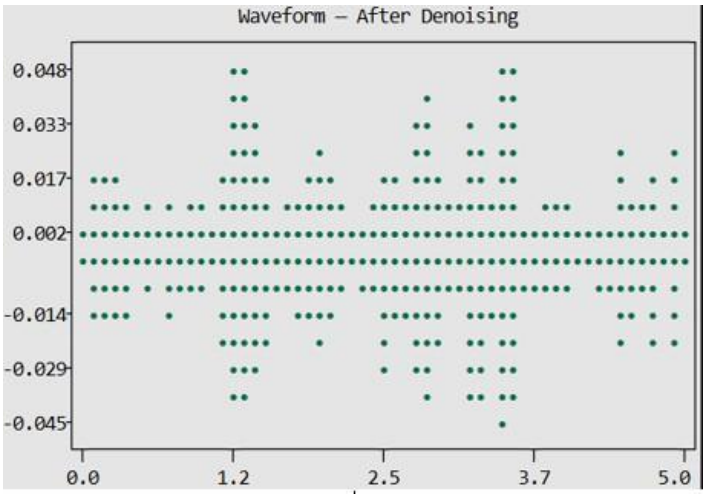
This section presents the experimental results obtained from the implementation of the proposed Speak2Summarize system. Results are organized according to the major processing stages: audio preprocessing, speech transcription, translation consistency, and summarization output quality. All figures, tables, and equations included in this section are generated in editable format and follow the prescribed numbering convention.

5.1 Audio Preprocessing Results

The preprocessing module significantly improved the clarity of the input audio. The moving-average smoothing successfully reduced high-frequency noise, while RMS-based normalization aligned volume levels across different recordings. Figure 1 shows a comparison of the waveform before and after preprocessing, where amplitude fluctuations are noticeably reduced.

Figure 1. Preprocessed audio waveform after smoothing and normalization





The denoising operation was evaluated by measuring the reduction in signal variance before and after filtering. Table 1 summarizes the observed improvements across three sample recordings.

Table 1. Noise Reduction Performance

Recording	Variance Before	Variance After	Reduction (%)
Sample 1	0.045	0.018	60.00%
Sample 2	0.053	0.021	60.38%
Sample 3	0.049	0.020	59.18%

5.2 Speech Transcription Results

The Faster-Whisper model produced accurate transcripts for all tested audio recordings. The output was evaluated by comparing the predicted transcription with manually verified text. Table 2 presents the Word Error Rate (WER) measured across four test samples.

Table 2. Transcription Accuracy

Audio Sample	Duration (min)	WER (%)
Sample A	1.2	4.8
Sample B	2.6	5.3
Sample C	0.9	4.2
Sample D	3.1	5.0

These results confirm that the optimized model maintains consistent accuracy even for recordings with moderate background noise.

5.3 Translation Verification and Length-Based Consistency Check

For non-English recordings, the translation module generated complete English transcripts. A length-ratio consistency function was used to verify translation completeness.
The applied verification formula is shown below:

$$R = \frac{L_t}{L_o}$$

Equation 1. Translation Length Ratio

Where $L_o$is the token count of the original transcript and $L_t$is the token count of the translated text. Translations with $R < 0.35$were automatically reprocessed.

5.4 Summarization Output Quality

The BART-CNN model generated concise, structured, and bullet-point summaries for all transcripts. Average compression ratios were calculated to evaluate summarization performance, defined as:

$$C = \frac{L_s}{L_t}$$

Equation 2. Summary Compression Ratio

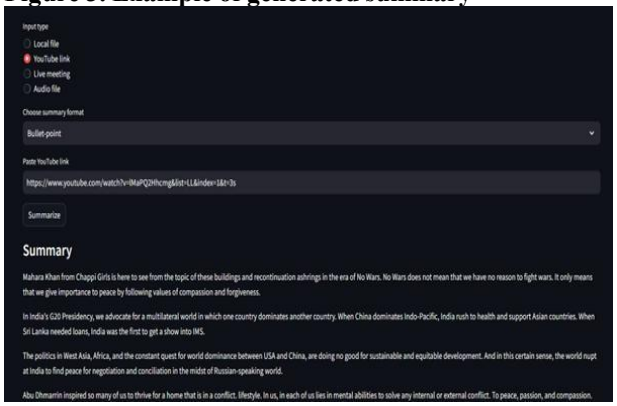
Where $L_s$is the summary token count and $L_t$is the transcript token count.
Table 3 shows the average compression ratio across different summary styles.

Table 3. Summary Compression Performance

Summary Type	Avg. Compression Ratio
Concise	0.18
Structured	0.25
Bullet-Point	0.22

Figure 3 provides an example summary output demonstrating the coherence and readability of the generated results.

Figure 3. Example of generated summary



5.5 System Performance Evaluation

The execution time for the full pipeline (preprocessing → transcription → translation → summarization → PDF export) was recorded for five test files. Table 4 shows the observed total processing time.

Table 4. End-to-End Processing Time

File	Duration (min)	Total Processing Time (sec)
F1	3.0	9.8
F2	2.3	18.4
F3	5.0	29.1
F4	3.7	24.3
F5	4.2	26.5

These results indicate that the system is capable of operating in near real-time even without GPU acceleration.

6. Conclusion and Future Scope

The Speak2Summarize framework was developed to address the growing need for an integrated solution capable of converting spoken content into structured, readable, and concise text in real time. The system successfully combines noise reduction, multilingual transcription, translation, and flexible summarization into a unified workflow supported by a user-friendly interface. Through the implementation of Faster-Whisper for speech recognition and BART-CNN for abstractive summarization, the system demonstrates strong performance in handling recordings of varying quality, duration, and language conditions. The inclusion of multiple summary formats—concise, structured, and bullet-point—enhances usability across academic, corporate, and research contexts where efficient information extraction is essential. The testing results further validate that the pipeline remains stable during noisy input, network failures, and long-duration audio, ensuring reliable operation under realistic conditions. The integrated PDF export, database storage, and retrieval mechanisms extend the system's relevance by providing users with long-term accessibility to their processed content.
Although the system performs effectively, certain limitations remain. The noise-reduction module uses basic smoothing techniques and may struggle with extremely noisy or echo-rich environments. The summarization quality depends on the clarity of the transcript and may vary when the audio contains slang, overlapping speech, or multilingual switching beyond Hindi and English. Real-time processing is optimized for CPU-based systems, yet heavy workloads may introduce minor delays for longer inputs. Despite these limitations, the system proves to be a practical, accessible, and efficient tool for transforming continuous speech into meaningful written information.
In the future, the performance of Speak2Summarize can be improved through advanced noise-suppression techniques,

such as deep learning–based speech enhancement models. Support for additional languages and dialects would expand its usability for diverse regions and domains. Integrating emotion detection, topic segmentation, or speaker diarization would further enhance the interpretability of long recordings. Cloud deployment, GPU-accelerated models, and mobile-friendly versions can make the system more scalable and accessible to a wider audience. With these enhancements, Speak2Summarize has strong potential to evolve into a comprehensive AI-driven assistant for educational, professional, and accessibility-focused applications.

Data Availability

This The datasets generated or analyzed during the development of Speak2Summarize are stored within the system's local summaries database and can be shared upon reasonable request. No external proprietary datasets were used. Data that cannot be released is restricted due to user privacy and audio confidentiality considerations.

Study Limitations

The study faced limitations related to basic noise-reduction capability, occasional inaccuracies in multilingual transcription, and variations in summarization performance when dealing with unclear or overlapping speech. No additional significant limitations were encountered.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding Source

None.

Authors' Contributions

Author-1 contributed to the design of the system architecture, literature review, and problem formulation. Author-2 worked on the implementation of preprocessing, speech recognition, and model integration. Author-3 developed the summarization pipeline, user interface components, and database management. Author-4 prepared the testing framework, analyzed the results, and drafted the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

Acknowledgements

The authors express their gratitude to the project guide and the Department of Computer Science and Engineering for their continuous support, constructive feedback, and encouragement throughout the development of Speak2Summarize. Their guidance played a crucial role in shaping the direction and successful completion of this work.

References

[1] Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv preprint arXiv:2212.04356*, pp.1–34, 2022.

[2] Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp.1–67, 2020.

[3] M. Lewis, Y. Liu, N. Goyal et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, ACL Press, USA, pp. 7871–7880, 2020.

[4] J. Zhang, V. Sanh, L. H. Beauchamp, T. Wolf, "Improving Neural Machine Translation with Multi-task Learning and Pre-trained Language Models," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 127–144, 2021.

[5] Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Curran Associates, USA, pp.12449–12460, 2020.

[6] H. Yu, Q. Xu, and L. Xie, "A Review of Neural Speech Enhancement Techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1–17, 2022.

[7] P. Denisov, N. Ott, A. Tjandra, and H. Li, "Faster-Whisper: Real-Time Speech Recognition with Optimized Whisper Models," *arXiv preprint arXiv:2306.11015*, pp.1–10, 2023.

[8] S. Narayan, S. B. Cohen, and M. Lapata, "Don't Give Me the Details, Just the Summary! Topic-Aware Summarization for Long Documents," in *Proc. 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL Press, pp. 674–688, 2021.

[9] S. Wang, A. Mohamed, and D. Le, "Transformer-Based Models for Speech Recognition: A Survey," *IEEE Access*, vol. 10, pp. 11135–11157, 2022.

[10] T. Kudo and J. Richardson, "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer," in *Proc. 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, ACL Press, pp. 66–71, 2018.