Leveraging Whisper and DeepSeek-V2 for Advanced Voice Text Summarization

Prathamesh Nikam[1][0009−0004−3162−2782]

University of Hyderabad, Hyderabad, India
24mcmb03@uohyd.ac.in

Abstract. The proliferation of audio and video content has created a demand for efficient methods to extract essential information. This paper presents a system for AI-powered voice text summarization that leverages OpenAI's Whisper model for robust speech-to-text transcription and the DeepSeek-V2 model for coherent, reasoning-based text summarization. We explore the architecture of both models, their respective strengths, and the process of fine-tuning them for these specific tasks. The paper details the data preprocessing steps for both audio and text, the datasets used for training and evaluation, and the metrics employed to measure performance, namely Word Error Rate (WER) for transcription and ROUGE and BLEU for summarization. We also discuss the practical implementation of this system, including the integration of the models and the ethical considerations surrounding bias and privacy.

Keywords: Voice Summarization · Speech Recognition · Text Summarization · Whisper · DeepSeek-V2 · Reasoning Models.

1 Introduction

The digital age is characterized by an explosion of audio and video content, leading to information overload from sources like meeting recordings, lectures, podcasts, and video blogs. [1] The ability to distill key insights from this spoken content into concise formats has become a critical need for both individuals and organizations. [2] AI-powered voice text summarization addresses this challenge by automating the conversion of spoken words into written summaries, which saves time, enhances accessibility, and improves the overall utility of spoken information. [3]

The applications of this technology are widespread. In business, it can automate the creation of meeting minutes and analyze sales calls. In education, it can summarize lectures for students. [4] Journalists can use it to quickly condense news broadcasts, and it holds immense potential in the healthcare and legal fields for summarizing patient consultations and legal documents, respectively. [1,5] Furthermore, it enhances accessibility for individuals with hearing impairments and allows content creators to repurpose their audio content into text formats.

Despite its potential, developing robust AI-powered voice text summarization systems presents several challenges. Spoken language is often complex, with errors in speech recognition, disfluencies, and a lack of clear sentence structure. [6] Ensuring the factual accuracy, coherence, and fluency of the generated summaries is paramount. [7] This paper details a system that combines the strengths of two powerful AI models, OpenAI's Whisper and the DeepSeek-V2 reasoning model, to address these challenges.

2 Related Work

The field of speech summarization is interdisciplinary, drawing from automatic speech recognition (ASR) and text summarization. [8,9]

2.1 Speech Recognition

Traditional ASR systems have been surpassed by end-to-end deep learning models. OpenAI's Whisper model, trained on a massive and diverse dataset of 680,000 hours of multilingual and multitask supervised data, has demonstrated significant robustness to accents, background noise, and technical language. [10,11] Its Transformer-based encoder-decoder architecture allows it to handle a wide variety of speech processing tasks. [10,12]

2.2 Text Summarization

Text summarization techniques are broadly categorized as extractive or abstractive. [4,13] Extractive methods select important sentences from the source text, while abstractive methods generate new sentences that capture the essence of the original content. The rise of Large Language Models (LLMs) has led to significant advancements in abstractive summarization. [4,14] Models like DeepSeek-V2, which employ advanced architectures like Mixture-of-Experts (MoE) and focus on general reasoning, have shown strong performance in language understanding and generation tasks. [15]

3 Methodology

Our proposed system for AI-powered voice text summarization consists of a two-stage pipeline: first, transcribing the audio input using Whisper, and second, summarizing the transcribed text using the DeepSeek-V2 model.

3.1 The Whisper Model for Speech-to-Text

At its core, Whisper utilizes a Transformer-based encoder-decoder architecture. The encoder processes a log-Mel spectrogram of the audio input to extract acoustic features. The decoder then autoregressively predicts the corresponding sequence of text tokens, conditioned on the encoded audio features. [10] A key

feature of Whisper is its multitask capability, allowing it to perform multilingual speech recognition, speech translation, and language identification through the use of special tokens that specify the desired task. [12]

Whisper is available in various model sizes, from "tiny" with 39 million parameters to "large" with 1.55 billion parameters, offering a trade-off between accuracy and computational resources. [16]

Table 1. Whisper Model Sizes and Specifications.

Size	**Parameters (M)**	**English-only Model**	**Multilingual Model**	**Approx. VRAM (GB)**	**Relative Speed**
tiny	39	tiny.en	tiny	~1	~10x
base	74	base.en	base	~1	~7x
small	244	small.en	small	~2	~4x
medium	769	medium.en	medium	~5	~2x
large	1550	N/A	large	~10	1x
turbo	809	N/A	turbo	~6	~8x

3.2 The DeepSeek-V2 Model for Reasoning-Based Summarization

For the summarization task, we employ the DeepSeek-V2 model, a powerful, open-source Language Model. Unlike models specialized for coding, DeepSeek-V2 is engineered for strong general reasoning and language comprehension. [15] It utilizes a Mixture-of-Experts (MoE) architecture, which enables it to scale to trillions of parameters while maintaining computational efficiency. Its strength in reasoning allows for a deeper understanding of the context, logical flow, and key arguments within a text, leading to more coherent and factually consistent summaries. Furthermore, its large context window is a significant advantage when processing lengthy transcriptions from audio sources.

3.3 Data Preprocessing

Audio Preprocessing for Whisper. Whisper expects audio input to be sampled at 16kHz. Therefore, resampling is a necessary first step for audio data with different sampling rates. The model's built-in feature extractor automatically converts the audio into a log-Mel spectrogram. For long audio files, segmentation into 30-second chunks is often required.

Text Preprocessing for DeepSeek-V2. The transcribed text from Whisper undergoes several preprocessing steps before being fed into DeepSeek-V2. This includes tokenization, where the text is broken down into smaller units, and text normalization, which involves standardizing the text to a consistent format (e.g., lowercasing, handling punctuation).

3.4 Fine-Tuning

Fine-Tuning Whisper. To enhance performance on specific domains or accents, Whisper can be fine-tuned on relevant datasets. The Hugging Face ‘transformers‘ library provides a comprehensive framework for this process, which involves using a cross-entropy loss function and carefully adjusting hyperparameters like learning rate and batch size.

Fine-Tuning DeepSeek-V2. Similarly, DeepSeek-V2 is fine-tuned on the "samsum" dataset, which consists of dialogue transcripts and their corresponding human-written summaries. This process enables the model to learn the nuances of summarizing conversational text, leveraging its inherent reasoning capabilities.

4 Experiments and Results

4.1 Datasets

For fine-tuning the audio-to-text component, several publicly available datasets can be utilized, such as CommonVoice11, LibriSpeech, VoxPopuli, and CoVoST2. The text-to-summary component is fine-tuned using the "samsum" dataset.

4.2 Evaluation Metrics

Word Error Rate (WER) for Transcription. The primary metric for evaluating the performance of the Whisper model is the Word Error Rate (WER). [17] It is calculated as:

$$WER = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference transcript. A lower WER indicates higher accuracy.

ROUGE and BLEU for Summarization. The performance of DeepSeek-V2 is evaluated using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy). [7,13] ROUGE measures the overlap of n-grams between the model-generated summary and a human-written reference. [18] BLEU assesses the similarity of the generated text to the reference text by calculating the precision of n-grams. [13]

4.3 Workflow Integration: The MAIN.PY Script

The complete workflow is orchestrated by a ‘MAIN.PY‘ script. This script handles the audio input, invokes the appropriate Whisper model for transcription, and then passes the transcribed text to the DeepSeek-V2 model (accessed via

LM Studio) for summarization. The choice of the Whisper model size is configurable, allowing for a balance between transcription accuracy and processing time. The interaction with DeepSeek-V2 involves sending the transcribed text along with a carefully crafted prompt that instructs the model on how to perform the summarization.

5 Discussion

The combination of Whisper and DeepSeek-V2 presents a powerful and flexible system for voice text summarization. The variety of Whisper model sizes allows for adaptation to different resource constraints and accuracy requirements. DeepSeek-V2's advanced reasoning capabilities make it particularly well-suited for untangling the unstructured and often logically complex text produced by speech transcription, resulting in more coherent summaries.

The performance of the system is heavily dependent on the quality of the fine-tuning datasets and the effectiveness of the prompts used for summarization. Further research is needed to explore the impact of different prompting strategies that can better leverage the model's reasoning pathways.

6 Ethical Implications and Bias Considerations

The use of AI for voice text summarization raises important ethical concerns.

6.1 Privacy

These systems process spoken data that can contain sensitive personal information. Ensuring the security and confidentiality of this data is crucial to protect individual privacy.

6.2 Bias

The datasets used to train these models can contain biases that are then reflected in their performance. Whisper's training data has a known bias towards English, which can affect its accuracy on other languages and accents. [12] Similarly, the "samsum" dataset may contain demographic biases that could lead to less accurate summaries for certain conversational styles. It is essential to use diverse and representative datasets and to implement techniques for bias detection and mitigation.

6.3 Accuracy and Misuse

Inaccurate transcriptions or summaries can have serious consequences, particularly in high-stakes domains like healthcare and law. The potential for these systems to be used to create misleading or false summaries also needs to be addressed through robust evaluation and human oversight.

7 Conclusion and Future Work

This paper has outlined a comprehensive system for AI-powered voice text summarization using OpenAI's Whisper and the DeepSeek-V2 reasoning model. The approach demonstrates significant potential for efficiently extracting key information from spoken content.

Future work will focus on several areas. Advanced fine-tuning strategies for both models will be explored to improve accuracy and efficiency. The impact of different prompting techniques that explicitly engage chain-of-thought reasoning in DeepSeek-V2 will be investigated. Furthermore, the integration of other complementary AI models, such as for noise reduction or topic modeling, will be explored to enhance the overall system. As this technology continues to evolve, ongoing attention to the ethical implications is paramount to ensure its responsible and beneficial deployment.

References

1. Sonix: AI Text Summarization Tool: A Helpful Technology. (2025). https://sonix.ai/resources/ai-text-summarization/
2. Nowigence Inc.: Importance & Benefits of Auto Text Summarization. (2025). https://www.nowigence.com/importance-benefits-of-auto-text-summarization/
3. Author, A.: VOICE TO TEXT SUMMARIZATION USING NLP. ResearchGate (2025). https://www.researchgate.net/publication/390260605_VOICE_TO_TEXT_SUMMARIZATION_USING_NLP
4. IJISRT: Meeting Insights Summarisation Using Speech Recognition. (2023). https://ijisrt.com/assets/upload/files/IJISRT23APR2036.pdf
5. Frase: 20 Applications Of Automatic Summarization In The Enterprise. (2025). https://www.frase.io/blog/20-applications-of-automatic-summarization-in-the-enterprise/
6. Dialpad: AI Summarization: Use Cases, Challenges, & Solutions. (2025). https://www.dialpad.com/blog/why-ai-summarization-is-hard/
7. Google Cloud: AI summarization. (2025). https://cloud.google.com/use-cases/ai-summarization
8. Retkowski, F., et al.: Summarizing Speech: A Comprehensive Survey. arXiv preprint arXiv:2504.08024 (2025)
9. Graphcore: How to use OpenAI's Whisper for speech recognition. (2023). https://www.graphcore.ai/posts/how-to-use-openais-whisper-for-speech-recognition
10. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356 (2022)
11. OpenAI: Introducing Whisper. (2022). https://openai.com/research/whisper
12. Speechmatics: Whisper Speech to Text Deep-Dive. (2022). https://www.speechmatics.com/blog/whisper-speech-to-text-deep-dive/
13. Goyal, T., Li, J.J., Durrett, G.: News Summarization and Exploration. arXiv preprint arXiv:2206.01755 (2022)

14. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461 (2020)
15. DeepSeek-AI Team: DeepSeek-V2: A Strong, Economical, and Open-Source Language Model. arXiv preprint arXiv:2405.04434 (2024)
16. Hugging Face: OpenAI Whisper. https://huggingface.co/openai/whisper-large-v3
17. Author, A.: A Tiny Whisper-SER: Unifying Automatic Speech Recognition and Multi-label Speech Emotion Recognition Tasks. In: APSIPA ASC (2024)
18. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv preprint arXiv:1912.08777 (2020)