A Review on Techniques for Speech-to-Text Conversion and Text Summarization: Progress and Phase-Wise Implementation

Chanchala Tripathi1, Hemashu waghmare2, Dhruv Dalvi3, Rishabh Jain4, Sanket Asole5, Yuvraj Chavan6

1-6Yeshwantrao Chavan College of Engineering, Department of Computer Science and Engineering, Nagpur, Maharashtra, India

Email: {chanchalatripathi, hemanshu.waghmare, dhruvdalvi786, rjain23112002, sanketasole7755, yuvichavan968 [}@gmail.com](mailto:%7d@gmail.com)

***Abstract***—***With the exponential growth of digital audio data, Speech-to-Text (STT) conversion and text summarization have emerged as critical technologies within the domain of natural language processing (NLP). These advancements enable seamless transcription of spoken content into text and distillation of vast amounts of textual data into concise summaries. Applications span various fields, including meeting transcriptions, virtual assistants, podcast analysis, and automated content generation, emphasizing their increasing significance in both academic research and real-world implementation.***

***This paper presents the methodologies and results of the first two phases of a comprehensive five-phase project aimed at integrating STT and summarization into a unified framework. Phase one focuses on audio preprocessing, incorporating advanced noise reduction techniques such as spectral gating and segmentation of long audio files into manageable chunks. Phase two explores speech recognition and speaker diarization, leveraging the powerful WhisperX model to achieve accurate transcription and effective speaker differentiation in multi-speaker environments.***

***Key contributions of this work include the detailed step-by-step implementation of preprocessing and transcription processes, evaluation of the effectiveness of applied methods, and discussion of encountered challenges. These phases provide a robust foundation for subsequent integration of transformer-based text summarization models. This paper not only highlights the current progress but also sets the stage for achieving a seamless pipeline that can process, analyze, and summarize audio data in diverse application domains.***

***Index Terms: Speech-to-Text, WhisperX, Noise Reduction, Audio Segmentation, Diarization, Text Summarization***

1. INTRODUCTION

In the modern era of data-driven decision-making and automation, audio content has become a dominant source of information. From business meetings and online lectures to podcasts and interviews, audio data holds valuable insights that need to be efficiently extracted and utilized. Speech-to-Text (STT) and text summarization systems play a pivotal role in transforming raw audio data into actionable textual formats while ensuring that lengthy information is condensed into meaningful summaries. These technologies have become indispensable tools for improving accessibility, enhancing productivity, and facilitating information retrieval in diverse fields.

1.1 The Role of STT and Text Summarization

STT systems bridge the gap between spoken language and textual representation, enabling applications such as:

* Accessibility Solutions: Assisting individuals with hearing impairments through real-time transcriptions.
* Automated Documentation: Streamlining workflows by transcribing meetings, interviews, and lectures.
* Content Indexing: Facilitating the organization and retrieval of audio and video content in multimedia platforms.

Text summarization complements STT systems by providing concise overviews of large textual data, saving time and effort in understanding the core content. This is especially useful for summarizing transcriptions from podcasts, speeches, or multi-hour recordings into a format that highlights key points.[10]

1.2 Evolution of Technologies

The evolution of STT and summarization systems underscores significant advancements in computational methodologies:

1. Early STT Systems:

* Earlier STT systems relied on statistical models like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs).[1]
* These approaches integrated acoustic, language, and pronunciation models to convert speech signals into text. However, they struggled with variations in accents, languages, and noisy environments, making them less reliable for real-world applications.

1. Modern Approaches in STT:

* The advent of deep learning introduced neural network-based models, including Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks[3][6]
* Contemporary models, such as transformer architectures like OpenAI’s Whisper, have further enhanced transcription accuracy, processing speed, and adaptability to multiple languages and accents.

1. Advances in Text Summarization:

* Early summarization techniques were rule-based, relying on manually crafted heuristics to extract or generate summaries. These methods often lacked contextual understanding and flexibility.
* Modern summarization employs neural architectures like BERT, GPT, and T5, which enable abstractive and extractive summaries with contextual awareness, coherence, and grammatical correctness [8][5].

These advancements have paved the way for more efficient and accurate systems capable of addressing the growing demand for audio processing and information summarization.

1.3 Project Objective and Scope

Recognizing the critical role of STT and text summarization, our project seeks to develop an integrated framework that seamlessly combines these technologies into a unified pipeline. The project is structured into five phases, each focusing on specific aspects of system development. This paper specifically discusses the first two phases:

1. Audio Preprocessing:

* Focused on enhancing audio quality by reducing noise and segmenting long audio files into smaller, manageable chunks.
* This phase ensures that the input audio is optimized for downstream transcription tasks.

1. Speech Recognition:

* Utilized WhisperX, a state-of-the-art model based on transformer architecture, to transcribe audio into text and identify multiple speakers through diarization.
* This phase establishes the foundation for integrating transcription with text summarization.

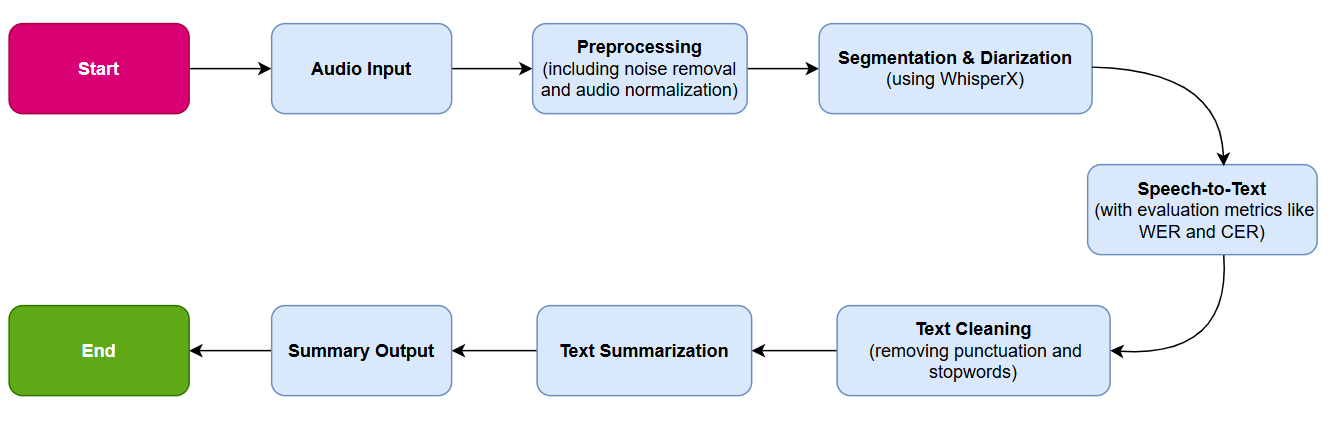
1.4 Future Directions

While this paper highlights progress made in the first two phases, the subsequent phases will explore advanced summarization models like BART and T5. These models will enable the generation of concise and coherent summaries directly from transcribed text, paving the way for seamless audio-to-text-to-summary pipelines. Future work will also focus on enhancing the system’s scalability, adaptability to diverse datasets, and real-world usability in multi-lingual and multi-domain contexts.

By addressing the challenges inherent in each phase and leveraging cutting-edge technologies, this project aims to create a robust, end-to-end system for processing, analyzing, and summarizing audio data, thereby contributing to the growing field of natural language processing[4][6].

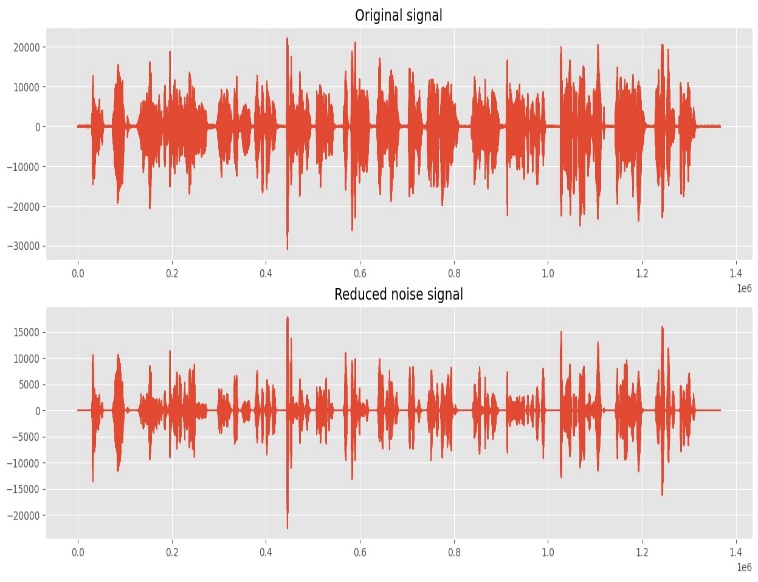
1. METHODOLOGY

The methodology for this project involved two critical phases—audio preprocessing and speech recognition with speaker diarization. These phases were designed to ensure the quality and usability of audio data for accurate transcription and to lay the groundwork for the subsequent integration of text summarization techniques.



*Figure 1: Workflow diagram*

In the first phase, audio preprocessing focused on enhancing the quality of raw audio data through noise reduction and segmentation. Noise reduction aimed to improve the Signal-to-Noise Ratio (SNR) by suppressing unwanted background noise while preserving the essential speech components. This process utilized spectral gating, implemented through the noisereduce library, which analyzes the frequency domain of the audio to identify and attenuate noise frequencies. Audio files were handled using pydub, which simplified operations such as loading, modifying, and saving the data. The procedure involved loading the audio file, performing spectral analysis to detect noise patterns, applying spectral gating to suppress these frequencies, and saving the denoised audio for further processing [2]. Challenges such as overlapping speech and ambient noises in outdoor recordings were mitigated effectively using this approach, which maintained speech clarity without significant distortion.



*Figure 2: Original signal vs Reduced noise signal*

Audio segmentation was another critical aspect of preprocessing. This step divided long audio files into smaller, manageable chunks, optimizing memory usage and enhancing the efficiency of transcription models. The segmentation process relied on detecting periods of silence longer than two seconds using pydub. The audio was split into segments at these points, ensuring that each segment maintained sentence-level coherence. The resulting segmented files were saved and prepared for input into the transcription model. This approach reduced the computational load and improved the handling of real-time transcription tasks, particularly in scenarios involving lengthy recordings or overlapping speech.

The second phase of the project involved speech recognition and speaker diarization, which converted segmented audio files into textual representations while identifying and distinguishing multiple speakers. For this purpose, WhisperX, a state-of-the-art Speech-to-Text (STT) model based on OpenAI's Whisper architecture, was employed. WhisperX leverages transformer-based layers to achieve high transcription accuracy and robust handling of diverse accents, languages, and noise conditions [9]. The segmented audio files were processed through WhisperX, which generated time-stamped transcriptions. These outputs were validated through manual evaluation and cross-referencing with reference datasets to ensure accuracy.

Additionally, speaker diarization was integrated into the WhisperX pipeline to differentiate between speakers in multi-speaker recordings. This process extracted acoustic features from the audio, clustered speech segments based on speaker characteristics, and assigned speaker labels to the transcribed text. Diarization added contextual depth to the transcriptions, making them highly useful for applications such as meeting documentation and podcast analysis [10]. The combined approach of WhisperX for transcription and diarization enabled precise and contextually rich outputs, setting the stage for the subsequent phases of the project.

1. RESULTS AND DISCUSSION

3.1 Effectiveness of Noise Reduction

Noise reduction techniques demonstrated a significant improvement in transcription clarity. While traditional methods often suppressed essential speech components, spectral gating preserved critical frequencies, ensuring high-quality input for STT.

3.2 Audio Segmentation

Segmentation was particularly effective in handling long recordings, reducing memory constraints and enabling efficient processing. This step ensured that WhisperX processed manageable chunks of audio without performance degradation.

3.3 Speech Recognition and Diarization

WhisperX produced accurate transcriptions and successfully differentiated speakers, even in multi-speaker settings. The integration of diarization enhanced the utility of the transcriptions by providing speaker-specific annotations.

1. CHALLENGES AND FUTURE DIRECTIONS

Despite significant progress in the initial phases, several challenges arose during implementation that require further refinement and optimization. One notable challenge was processing noisy outdoor recordings. While techniques like spectral gating enhanced audio clarity, environments with heavy traffic, wind, or overlapping conversations posed significant issues. Noise in these scenarios often overlapped with human speech frequencies, making it difficult to suppress noise without distorting speech signals, thus reducing transcription accuracy, especially in field recordings or public speeches.

Another challenge involved overlapping speech in speaker diarization. Although WhisperX performed well in most scenarios, it struggled when multiple speakers talked simultaneously, leading to segmentation and attribution errors. This issue was particularly evident in dynamic group discussions, panel debates, and informal meetings with frequent interruptions. Additionally, scalability and efficiency presented concerns, as handling large datasets demanded considerable computational resources during preprocessing and transcription. Current methods, optimized for smaller datasets, limit scalability for enterprise-level or real-time applications.

To address these challenges, several future directions have been identified. The integration of transformer-based text summarization models, such as BART and T5, will be prioritized. These models are expected to generate coherent, high-quality summaries from transcribed text by leveraging pre-trained models fine-tuned on domain-specific datasets for enhanced precision and contextual relevance.

Improvements to speaker diarization will focus on resolving overlapping speech issues through advanced clustering techniques and deep learning-based methods like EEND (End-to-End Neural Diarization), which can effectively manage simultaneous speakers [9]. Additionally, testing on larger, more diverse datasets will enhance the system’s generalizability. These datasets will include multiple languages, accents, and audio qualities from real-world scenarios such as call center recordings, conference meetings, and interviews, ensuring robustness across various environments.

The final phase will prioritize real-time processing and deployment on cloud platforms like AWS or Azure, enabling live transcription and summarization capabilities [7]. This will facilitate scalability and accessibility for both enterprise and individual users. These efforts aim to refine system performance and broaden its applicability across diverse use cases and environments, ensuring it meets evolving demands.

1. CONCLUSION

The completion of the first two phases of this project represents a significant step toward building a comprehensive Speech-to-Text (STT) and text summarization pipeline. The findings and implementations thus far highlight the following key contributions:

1. Enhanced Audio Preprocessing:

Noise reduction through spectral gating effectively improved the clarity of input audio, reducing the impact of background disturbances.

Audio segmentation streamlined the transcription process, optimizing memory usage and ensuring manageable processing of lengthy recordings.

1. Reliable Transcription and Diarization:

WhisperX demonstrated robust performance in converting speech to text with high accuracy, even for multi-speaker scenarios.

The integration of speaker diarization provided additional context by distinguishing and annotating individual speakers within the transcriptions.

1. Foundational Framework:

These advancements establish a solid foundation for the upcoming phases, which will incorporate transformer-based summarization models and focus on system scalability, real-time processing, and multi-language support.

Significance and Impact:  
This project addresses critical challenges in audio processing and contributes to the broader field of natural language processing by integrating advanced STT and summarization technologies. Once completed, the system will enable efficient handling of large-scale audio data, supporting applications in accessibility, automation, and content analysis across industries. By addressing the outlined challenges and implementing the proposed future directions, the project aims to deliver a robust, scalable, and versatile solution for audio data processing, transcription, and summarization.

1. REFERENCES

[1] M. Mehta, K. Gupta, S. Tiwari and Anamika, "A Review on Sentiment Analysis of Text, Image and Audio Data," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1660-1667, doi: 10.1109/ICCMC51019.2021.9418360.

[2] V. Maheshwar Reddy, K. Deepika, K. Adithya Surya Prakash, and M. Sanathan, "A survey on audio analysis: Text characterization and summarization," World Journal of Advanced Research and Reviews, vol. 21, no. 03, pp. 1596–1601, 2024, doi: 10.30574/wjarr.2024.21.3.0789.

[3] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," arXiv:1303.5778 [cs.NE], 2013, doi: <https://doi.org/10.48550/arXiv.1303.5778>

[4] Ghadekar, Premanand & Anand, Divsehaj & Gupta, Aryan & Oswal, Preeti & Sharma, Dheeraj & Khare, Shreyas. (2023). Audio Based Text Summarization Using Natural Language Processing. 10.1007/978-981-99-3656-4\_17.

[5] M. -H. Su, C. -H. Wu and H. -T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2061-2072, 2020, doi: 10.1109/TASLP.2020.3006731.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N. (2017). Attention is all you need [J]. Advances in neural information processing systems, 30(1), 261-272.

[7] Kaushal Rajendra Khonde, Dr. Jaimeel Shah, Dr. Pratik Patel, "Audio Transcription and Summarization System using Cloud Computing and Artificial Intelligence" in 2023 International Journal on Recent and Innovation Trends in Computing and Communication, Available: https://ijritcc.org/index.php/jritcc/article/view/8606

[8] Dhumal, Priyanka & Sutar, Sudarshan & Surve, Indraneel & Munawwar, Mirza & Nanaware, Vishal. (2024). Text Summarization Using NLP. International Journal of Advanced Research in Science, Communication and Technology. 319-324. 10.48175/IJARSCT-18650.

[9] Monteiro, R., Pernes, D. (2023). Towards End-to-End Speech-to-Text Summarization. In: Ekštein, K., Pártl, F., Konopík, M. (eds) Text, Speech, and Dialogue. TSD 2023. Lecture Notes in Computer Science(), vol 14102. Springer, Cham. <https://doi.org/10.1007/978-3-031-40498-6_27>

A. Vartakavi, A. Garg and Z. Rafii, "Audio Summarization for Podcasts," 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 431-435, doi: 10.23919/EUSIPCO54536.2021.9615948.