AI-Driven Speaker Recognition and Audio Summary Generation

Dr. B. Muthusenthil[1], B. Aditya Bharathi[2], C.P. Harish Raaj[3], Jai Harish Satheshkumar[4]
[1,2,3,4] *(Artificial Intelligence and Data Science, SRM Valliammai Engineering College, India)*

Abstract: **This project harnesses advanced artificial intelligence (AI) and natural language processing (NLP) to develop a comprehensive voice analysis system that generates accurate captions and concise summaries from audio inputs. By analyzing key voice parameters—such as tone, pitch, and modulation—the system effectively captures the nuances of speech. Additionally, speaker diarization is integrated to distinguish between multiple speakers, ensuring precise attribution of dialogue.**

The system seamlessly converts spoken content into readable text, enhanced by an advanced summarization algorithm that distills essential information while preserving context. Addressing the growing demand for efficient voice-to-text conversion, this solution proves valuable for transcription services, meeting documentation, and accessibility applications.

A key objective of this project is to deliver a functional prototype capable of handling complex audio inputs, correctly attributing speech, and generating high-quality transcriptions and summaries. With applications across media, legal, and education sectors, this technology offers a powerful tool for managing and interpreting spoken content. By demonstrating how AI and deep learning can revolutionize voice analysis and text generation, this project contributes to making spoken information more accessible, structured, and insightful.

Keywords**: AI-driven transcription, deep learning, natural language processing, speaker diarization, speech summarization.**

1. INTRODUCTION

The rapid advancements in artificial intelligence (AI) and natural language processing (NLP) are revolutionizing speech processing, particularly in transcription, interpretation, and summarization. Human speech is inherently complex, influenced by tone, pitch, and modulation, making accurate processing challenging, especially in multi-speaker conversations. Existing voice-to-text systems often struggle with precision and contextual understanding, limiting their effectiveness in applications requiring high accuracy.

To address these challenges, this project introduces an AI-driven voice analysis system designed to generate accurate captions and concise summaries from audio inputs. By leveraging deep learning and NLP techniques, the system captures the nuances of speech and employs speaker diarization to differentiate between multiple speakers, ensuring proper attribution of dialogue. The summarization component extracts key insights, preserving the original context while distilling essential information.

This technology is particularly valuable in transcription services, accessibility solutions, and documentation across various industries, including media, law, and education. With its ability to handle complex audio inputs and deliver structured, meaningful text outputs, this system enhances efficiency, reduces manual effort, and ensures the integrity of spoken communication. By automating transcription and summarization, it offers a scalable, adaptable solution for managing and comprehending spoken content in an increasingly digital world.

1.2 Objectives:

1. To develop an Accurate Voice Analysis System: Our primary objective is to create an AI-powered voice analysis system capable of generating highly accurate captions and summaries from audio inputs. This includes developing a deep learning model that effectively processes voice parameters such as tone, pitch, and modulation to ensure high-quality transcription.

2.To enable Multi-Speaker Differentiation: We aim to facilitate accurate speaker attribution in multi-person conversations by integrating advanced speaker diarization techniques. This will allow the system to distinguish between different speakers in real-time, ensuring that dialogue is attributed correctly in multi-speaker environments like meetings and interviews.

3. Accurate User Voice Recognition System : Our primary objective is to create an AI-powered voice recognition system capable of accurately identifying users based on their voice characteristics. This includes developing a deep learning model that effectively processes voice parameters such as tone, pitch, and modulation to ensure precise speaker identification. The system will enable robust authentication and personalization by leveraging advanced speech processing techniques.

4. To streamline the Summarization Process: Our system seeks to enhance productivity by providing concise and contextually appropriate summaries of long-form audio content. By distilling essential information from conversations, it will allow users to quickly grasp key points without sifting through entire transcripts, thereby improving efficiency in industries such as media, education, and legal.

5. To increase Accessibility to Voice-to-Text Technology: This project aims to make high-quality voice-to-text conversion more accessible to a broad range of users, including those with hearing impairments. By offering precise and easy-to-read transcriptions, the system can serve as a valuable tool for individuals who require assistance with understanding spoken content.

2. ARCHITECTURE

2.1 Architecture Diagram:
The overall flow of the program is diagrammatically represented below:

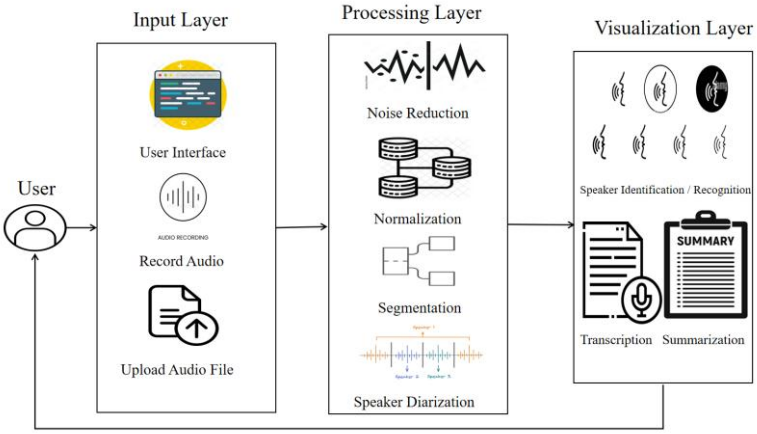

Fig 4.1 - Architecture Diagram

The entire program is segmented into three distinct layers, each serving a crucial role in facilitating interactions and operations within the system:

2.2 Architecture Layers:

2.3 Input Layer:

The Input Layer is the entry point where users provide audio data for analysis. This layer includes the following components:

User Interface:

This is the main interface that users interact with. It could be a web or mobile application that enables easy access to all functions. The interface allows users to navigate between different features, control the recording process, and upload files. It may also include buttons, sliders, and visual elements for controlling the quality of the audio input and selecting processing options.

Record Audio:

Provides a feature for users to record audio directly within the application. This option could use a built-in microphone or any connected recording device. The recording feature may include options for stopping, pausing, and replaying the audio before submission.

Upload Audio File:

Users can also choose to upload an existing audio file. This is helpful when users have prerecorded audio that they want to analyze. The system may support multiple audio formats (e.g., MP3, WAV) to provide flexibility. Upload functionality could also involve file size limitations, duration limits, and real-time checks for compatibility with the system.

2.4 Processing Layer:

The Processing Layer is the core of the system where all the technical operations on the audio file take place. Each component performs a specific function to prepare and analyze the audio:

Noise Reduction:

This step uses algorithms to filter out background noise, making the audio clearer and enhancing the accuracy of subsequent analysis. Common techniques include spectral subtraction, wavelet-based denoising, and adaptive filtering.

This step is crucial for recordings made in noisy environments, as it improves transcription quality and speaker recognition accuracy.

Normalization:

Adjusts the audio volume to a consistent level, ensuring that variations in sound intensity are balanced. This step improves the performance of speech processing algorithms by ensuring that loud and quiet parts of the audio are uniformly scaled. Normalization techniques may include peak normalization (adjusting to the highest peak) and loudness normalization (adjusting based on perceived loudness).

Segmentation:

Divides the audio file into smaller, manageable segments, which may be based on silence detection, time intervals, or speaker changes. Segmentation helps in organizing the data, making it easier to analyze and apply additional processes like diarization or transcription. This step is especially useful for long audio files, as it reduces processing time and ensures that each segment is optimally prepared for speaker analysis.

Speaker Diarization:

Diarization is the process of identifying distinct speakers within the audio and labeling their corresponding segments. This component uses machine learning models or clustering techniques to distinguish between voices, applying speaker labels (e.g., Speaker 1, Speaker 2) to each segment. Diarization is vital for scenarios where multiple people are speaking, as it attributes the content correctly to each speaker. It also enables more accurate summarization and transcription.

Speaker Recognition:

Speaker recognition is the process of identifying and verifying individuals based on their unique vocal characteristics. This system utilizes deep learning models to analyse voice parameters such as pitch, tone, and modulation, distinguishing one speaker from another. By leveraging speaker embeddings and similarity measurements, the model accurately associates audio segments with specific users. Speaker recognition plays a crucial role in authentication, security, and personalized user experiences, ensuring precise identification in applications like voice-based logins and customized AI interactions.

2.5 Visualization Layer

The Visualization Layer presents the analyzed results in a user-friendly format. It emphasizes clarity and accessibility, allowing users to understand the content briefly.

Speaker Identification:

This feature visualizes the speakers detected in the audio, displaying speaker labels along with their corresponding segments. It could use icons, colors, or waveforms to represent each speaker, making it clear when each speaker is active. Speaker identification may be shown in a timeline format, allowing users to see the flow of conversation and transitions between speakers over time.

Summarization:

The summarization component generates a condensed version of the audio content, highlighting the main points and omitting redundant or unimportant information. This summary is especially valuable for lengthy recordings, as it provides users with an overview of the content without needing to listen to the entire audio. Summarization techniques could involve natural language processing (NLP) models that detect essential information, keywords, and themes, presenting them in a clear and coherent format. The summarization may also incorporate speaker context, ensuring that key points are attributed to the correct speaker, adding depth to the summary.

3. CONCLUSION

In conclusion, our project marks a substantial advancement in voice analysis technology, providing an efficient, AI-powered system for

speaker diarization, caption generation, and content summarization. By incorporating state-of-the-art deep learning models, we have developed a solution that not only transcribes audio accurately but also attributes dialogue to the correct speakers, creating an organized and easy-to-follow narrative of any audio content. The system's iterative learning and feedback integration enhance its accuracy, making it more responsive to diverse audio inputs, including multiple speakers and complex speech characteristics. This adaptability results in improved transcription precision and summarization relevance, ensuring users receive concise yet comprehensive overviews of lengthy audio files. Our approach offers significant value across fields like media, education, and business, where time-efficient, accurate transcription and summarization can streamline workflows and boost accessibility.

By providing a user-friendly interface, we empower users to engage with audio content more effectively, removing barriers to information access and enhancing overall productivity. The deep integration of AI techniques, including advanced NLP for nuanced transcription and a robust summarization framework, enables our system to handle complex audio data with ease, making it an invaluable tool for users worldwide. In summary, our project showcases how AI can revolutionize audio processing, setting a new standard for efficient, accessible voice-to-text solutions. The homepage of our website serves as a central access point for users to record audio directly by clicking the 'Start Recording' button, which facilitates local storage of the recorded files. Upon completion of the recording, users can view the generated WAV file displayed on the left side of the interface, enabling them to initiate speaker diarization and conversation summarization processes.

The interface is designed with user-friendliness in mind, featuring a clean and minimalist layout that promotes seamless interaction. The speaker diarization module utilizes the pyannote/speaker-diarization-3.1 framework to analyze user inputs and deliver real-time transcriptions. Additionally, the summarization module employs a fine-tuned flan-t5-base model to provide effective summaries of the conversations. With its intuitive design and comprehensive functionalities, the homepage serves as a singular solution for efficiently handling audio input, generating transcripts, and offering summarization capabilities.

REFERENCES

Journal Papers:

[1] Agostinelli, F., Anderson, M. R., & Singh, S. (2023). "Speaker-Consistent End-to-End Speech Recognition via Self-Supervised Learning."

[2] Chen, W., Huang, Z., Li, B., & Dong, Z. (2024). "Few-shot Learning for Multi-Speaker Recognition Using LLMs in Conversational AI."

[3] Deb, A., & Kang, W. (2024). "Prompting Large Language Models with Audio for General-Purpose Speech Summarization."

[4] Dupont, D., & Leroy, L. (2023). "Fine-Tuning LLMs for Real-Time Multi-Participant Speaker Tracking in Dialogues."

[5] Gao, H., Wang, C., & Zhang, P. (2022). "Automatic Speaker Identity Tracking in Group Conversations using Transformers."

[6] Hinton, G., Osorio, M., & Yuan, K. (2023). "Multi-Channel Speaker Embeddings for AI Agent-based Dialogues."

[7] Kang, W., & Roy, D. (2024). "Prompting Large Language Models with Audio for General-Purpose Speech Summarization."

[8] Lee, S., Kim, D., & Park, J. (2017). "Utilizing Deep Learning for Mental Health Diagnosis Through Chatbot Interaction: A Preliminary Study." Frontiers in Psychiatry.

[9] Li, Z., Luo, X., & Jiang, T. (2023). "Voice Activity Detection for Speaker Change Detection in Multi-Turn Conversations."

[10] Liu, Y., & He, L. (2022). "Towards Human-Like Spoken Dialogue Generation Between AI Agents from Written Dialogue."

[11] Nori, A., Mohan, S., & Iyyer, M. (2023). "From LLM to Conversational Agent: A Memory Enhanced Architecture for Summarization."

[12] Park, Y., & Gupta, A. (2023). "End-to-End Speaker Recognition Using LLM-Integrated Speech Summarization."

[13] Ravi, K., & Sharma, P. (2023). "Dialogue Summarization Using Advanced Speech Recognition Techniques in Large Models."

[14] Wang, L., Zhang, Y., & Li, P. (2023). "Automatic Speech-to-Text Conversion and Summarization Using Language Models."

[15] Zhang, J., & Chen, H. (2022). "Real-Time Speaker Recognition and Summarization in Conversational AI with Transformer Models."