Talk2Topics

AI-Powered Podcast Transcription and Topic Segmentation System

Technical Project Documentation

Prepared By

Devi Padmavathi Pediredla

Abstract

Talk2Topics is an end-to-end intelligent system developed to convert long-form podcast audio into structured, searchable knowledge. The system integrates speech recognition, semantic text processing, clustering algorithms, and summarization models into a unified pipeline. Instead of presenting users with long transcripts or requiring them to listen to entire recordings, the system automatically detects topic boundaries, generates concise summaries, extracts keywords, and organizes spoken content into navigable sections. The project demonstrates how artificial intelligence can transform raw audio into structured information that supports efficient understanding, searchability, and analysis.

1.Introduction

Long-form spoken content such as podcasts, lectures, and recorded discussions has become a major source of information. However, audio content is inherently sequential, meaning users must listen continuously to locate relevant information. Even when transcripts are available, they often appear as long unstructured text blocks that are difficult to scan. This limitation reduces accessibility and makes it challenging to extract knowledge efficiently from spoken material. The Talk2Topics system addresses this challenge by transforming spoken audio into structured textual representations. Rather than focusing only on transcription accuracy, the system emphasizes semantic organization. By detecting topic shifts, summarizing discussions, and highlighting keywords, it converts raw speech into meaningful segments that users can quickly browse and understand.

2. Problem Statement

Traditional speech-to-text systems generate transcripts but do not provide structural understanding. Users still need to read or listen through entire recordings to find specific information. This lack of semantic segmentation limits the usability of transcripts and prevents efficient knowledge extraction from long recordings.

3. Objectives

The primary objective of this project is to design a system capable of converting long audio recordings into structured textual knowledge. To achieve this goal, the system must automatically identify topic boundaries, generate concise summaries for each segment, extract relevant keywords, maintain timestamp alignment, and present results in an intuitive format. Another objective is to ensure that each processing stage enhances the quality of the next stage, forming a cohesive and reliable pipeline.

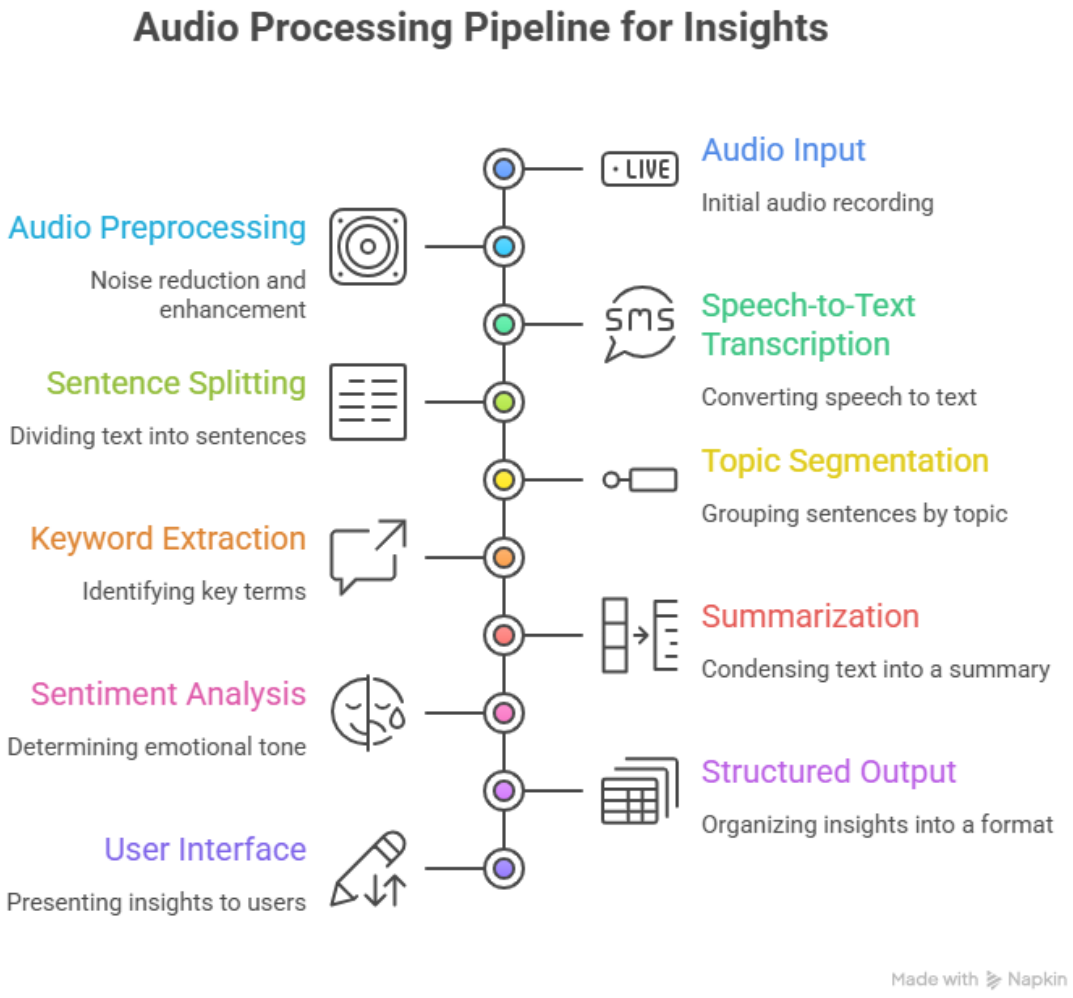
4. Dataset Description

The system was evaluated using a collection of long-form speech recordings derived from the TED Talks dataset. Ten recordings were selected, each ranging from approximately thirty to sixty-three minutes. These recordings include multiple speakers, diverse accents, and natural speaking patterns, providing realistic testing conditions for transcription and

segmentation models. Before analysis, audio files were standardized through preprocessing steps that included format conversion from MP3 to WAV, volume normalization, mono channel conversion, silence trimming, and segmentation into smaller chunks. These preprocessing operations improved transcription accuracy, reduced noise effects, and ensured efficient processing of long recordings.

5. System Architecture

The Talk2Topics architecture follows a sequential processing pipeline. Raw audio input first undergoes preprocessing to clean and normalize the signal. The processed audio is then transcribed into timestamped text using a speech recognition model. The transcript is divided into sentences, and each sentence is converted into a semantic embedding representing its meaning in numerical vector space. A clustering algorithm groups semantically related sentences together, producing coherent topic segments. Subsequent modules generate summaries, extract keywords, and analyze sentiment for each segment. Finally, all outputs are combined into a structured representation suitable for searching, reading, and navigation.



6. Technologies

Used Audio preprocessing and signal analysis were implemented using PyDub and LibROSA. Speech recognition was performed using the Whisper model due to its strong performance across accents and speaking conditions. SentenceTransformers generated semantic embeddings that enabled meaning-based clustering. TF-IDF was used for keyword extraction, the T5 transformer model produced abstractive summaries, and TextBlob performed sentiment classification. The user interface layer was built using HTML and JavaScript to provide interactive navigation between segments.

7. Methodology

7.1 Transcription

Audio recordings were divided into smaller chunks before transcription to ensure stable processing and memory efficiency. Each chunk was processed individually and later merged while preserving timestamps. This approach allows accurate transcription of long recordings without losing temporal alignment.

7.2 Topic Segmentation

Two segmentation strategies were explored during development. The first method compared similarity between consecutive sentences and split topics whenever similarity dropped below a threshold. Although sensitive to topic changes, this approach produced fragmented segments. The second method applied clustering to group sentences according to global semantic similarity. Experimental evaluation demonstrated that the clustering approach produced clearer topic boundaries and more coherent discussions. Therefore, clustering was selected as the final strategy.

7.3 Summarization

Instead of summarizing entire transcripts, the system generates summaries for each topic segment. This design preserves contextual meaning while significantly reducing reading time. Segment-level summarization allows users to preview discussions and decide whether to explore them in detail.

7.4 Sentiment Analysis

Sentiment classification is applied at the segment level rather than to the transcript as a whole. This allows emotional variation across discussions to be observed and provides additional contextual insight into conversational tone.

8. Evaluation and Testing

The system was evaluated using interview-style podcasts, lecture recordings, and conversational discussions. Performance was measured using criteria such as transcription accuracy, segmentation coherence, keyword relevance, summary clarity, interface usability,

and timestamp alignment. Iterative testing led to improvements including threshold tuning, summary compression adjustment, and keyword filtering refinement.

9. Limitations

Despite strong performance, several limitations remain. Transcription accuracy may decrease when recordings contain background noise, overlapping speakers, or strong accents. Topic segmentation requires balancing between overly detailed splitting and overly broad grouping. Output quality also depends on pretrained model performance, which may vary across domains. In addition, generic keywords may occasionally appear, sentiment analysis may miss subtle tones, and long recordings require more processing time.

10. Future Work

Future enhancements will focus on improving analytical capability and scalability. Planned developments include speaker diarization for multi-speaker detection, adaptive clustering algorithms, larger context-aware summarization models, automatic topic title generation, confidence scoring, analytics dashboards, distributed processing support, API integration, and improved robustness to noisy recordings.

11. Engineering

Insight An important lesson learned during development is that architecture design strongly influences output quality. Early experiments showed that weak segmentation logic produced poor summaries even when powerful models were used. After redesigning the system around semantic clustering, output quality improved significantly. This demonstrates that system architecture often has greater impact than individual model selection in applied artificial intelligence systems.

12. Conclusion

Talk2Topics demonstrates how speech recognition and natural language processing can transform long audio recordings into structured, searchable knowledge. By integrating transcription, semantic segmentation, summarization, keyword extraction, and sentiment analysis into a single pipeline, the system enables efficient exploration and understanding of spoken content. The results confirm that embedding-based clustering produces meaningful topic boundaries and significantly improves downstream comprehension.