



Automated Podcast Transcription And Topic Segmentation

K Yugavardhan

Introduction

Podcasts and long-duration audio content have become widespread mediums for knowledge sharing and discussions. However, navigating unstructured audio content remains challenging for users seeking specific information.

Problem Statement

- Long, unstructured audio episodes
- Difficult to locate specific topics
- Manual transcription is time-consuming
- Limited accessibility for text-based search

Our Solution

- Automated speech-to-text transcription
- Intelligent topic boundary detection
- Auto-generated summaries per topic
- Multiple export formats (PDF, DOCX, WebVTT)

Project Goal: Build an end-to-end automated system using Django framework that transforms raw audio into structured, searchable, and user-friendly outputs with comprehensive analytics.

Abstract

This project presents a Django-based Advanced Audio Processing Pipeline that provides a complete solution from audio upload to structured multi-format outputs. The system integrates state-of-the-art speech recognition and natural language processing techniques.

Audio Processing and Transcription

Supports multiple audio formats with intelligent preprocessing including resampling, noise reduction, silence removal, and voice activity detection. Utilizes OpenAI Whisper base model for accurate speech-to-text conversion with word-level timestamps.

Summarization and Labeling

Generates abstractive summaries using Facebook BART Large CNN model for each detected topic. Intelligent topic labeling combines TF-IDF keyword extraction with embedding similarity ranking.

Topic Segmentation and Analysis

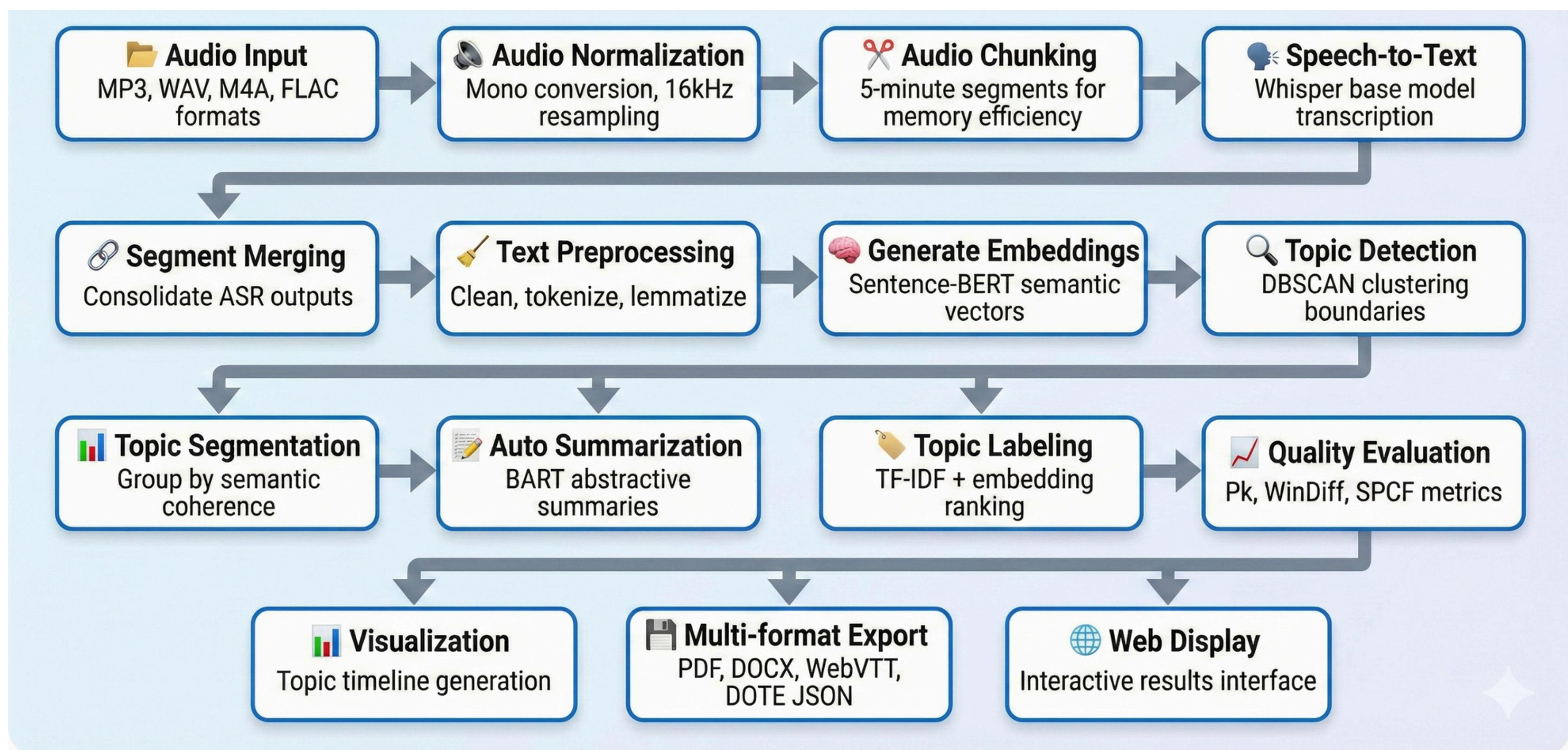
Employs unsupervised clustering-based approach using Sentence-BERT embeddings and DBSCAN clustering to detect natural topic boundaries. Advanced NLP preprocessing pipeline ensures clean semantic analysis.

Evaluation and Multi-Format Export

Implements industry-standard metrics for quality assessment. Provides comprehensive exports in PDF, DOCX, WebVTT, and DOTE JSON formats.

Core Benefit: Transforms hours of unstructured audio into organized, searchable, and navigable content with automatic topic detection, summarization, and quality metrics.

System Architecture



Technology Stack

Backend and Framework

Django 4.2.7 - Web framework
Python 3.8+ - Core language
PyTorch 2.8.0 - Deep learning
FFmpeg - Audio processing

AI Models

Whisper (base) - Speech recognition
BART Large CNN - Summarization
Sentence-BERT - Embeddings
spaCy - NLP processing

ML and Analysis

scikit-learn - Clustering, TF-IDF
HDBSCAN - Density clustering
UMAP - Dimensionality reduction
NLTK - Text processing

Export and Visualization

ReportLab - PDF generation
python-docx - Word documents
Matplotlib - Data visualization
WebVTT - Subtitle format

Model Sizes and Requirements

150MB

Whisper base

1.6GB

BART Large

80MB

Sentence-BERT

12MB

spaCy

Result & Metrics

Quality Metrics

Pk Score:	0.15 - 0.25
WinDiff:	0.18 - 0.28
SPCF:	0.65 - 0.80
Topic Coherence	0.55 - 0.70

Processing Speed

5 min audio:	2-3 minutes
15 min audio:	5-7 minutes
30 min audio:	10-15 minutes
Speed Factor:	30-40x realtime

3 Export Formats
PDF, DOC, TXT

5 Pipeline Steps
Fully Automated

3 AI Models Used
Fast and Reliable

Key Achievements

- ✓ Accurate English transcription using Whisper base model
- ✓ Successfully generated topic-wise segmented chapters
- ✓ Produced downloadable outputs in multiple formats
- ✓ Demonstrated effective end-to-end automation

Output Snaps

Home Page

AI Audio Processor
Upload audio for intelligent transcription and analysis

Pipeline Features:

- ✓ Automatic audio normalization (mono 16kHz)
- ✓ Whisper ASR for accurate transcription
- ✓ Speaker diarization & identification
- ✓ Topic segmentation using Sentence-BERT
- ✓ Automatic topic title generation
- ✓ Speaker-wise transcript with timestamps

Choose Audio File
Supported: MP3, WAV, M4A, FLAC, etc.

Processing...

Note: Processing may take several minutes depending on audio length.
The pipeline will perform transcription, speaker identification, and topic analysis.

Transcribing (ASR)

Output Page

Processing Results

Your audio has been analyzed successfully

0:00 / 0:00

Export PDF **Export DOCX** **Process Another**

Topics Identified **8**

Transcript Segments **179**

Accuracy **0.47**

Download Outputs

Output Snaps

Output Page

The screenshot displays a user interface for managing generated files and reviewing topics discussed.

Generated Files

- transcript.txt
- topics_timeline.txt
- topics_with_summaries.txt

Topics Discussed

Some Point Have Believe Reinvented

⌚ 0.0s – 121.0s ✓ Confidence: 0.73

Summary

Jensen Wong is the CEO of NVIDIA, the company that led a fundamental shift in how computers work. A huge amount of the most futuristic tech you're hearing about in AI and robotics relies on new chips and software designed by him and his company. We all need to know what he's building and why, and most importantly, what he is trying to build next.

Content

At some point you have to believe something. We've reinvented computing this now. What is the vision for what you see coming next? We asked ourselves if it could do this. How far can it go? How do we get from the robots that we have now to the future world that you see? Clearly everything

Output Snaps

Output Page

Cars Breakthrough Medical Research Relies

⌚ 67.9s – 189.0s ✓ Confidence: 0.72

Summary

Jensen Wong has already influenced all of our lives over the last 30 years and how many said it's just the beginning, something even bigger. What we do on Huge of True is we make optimistic explainer videos. We do it because we believe that when people see those better futures, they help build them. The people that you're going to be talking to are awesome.

Content

and self-driving cars and breakthrough medical research relies on new chips and software designed by him and his company. During the dozens of background interviews that I did to prepare for this, what struck me most was how much Jensen Wong has already influenced all of our lives over the last 30 years and how many said it's just the beginning, something even bigger. 30 years, and how many said it's just the beginning of something even bigger. We all need to know what he's building and why, and most importantly, what he's trying to build next. Welcome to Huge Conversations. Thank you so much. I'm so happy to do it. Before we dive in, I wanted to tell you how this interview is going to be a little bit different than other interviews I've done. When we dive in, I wanted to tell you how this interview is going to be a little bit different than other interviews I've seen you do recently. I'm not going to ask you any questions about company finances. Thank you. I'm not going to ask you questions about your management style or why you don't like one-on-ones. I'm not going to ask you about regulations or politics. I think all of those things are important, but I think that our audience can get them well covered also. What we do on Huge of True is we make optimists. well-covered also. What we do on Huge of True is we make optimistic explainer videos. I'm the worst person to be an explainer video. I think you might be the best. That's what I'm really hoping that we can do together is make a joint explainer video about how can we actually use technology to make the future better. We do it because we believe that when people see those better futures, they help build them. The people that you're going to be talking to are awesome. They are. better futures, they help build them. So the people that you're going to be talking to are awesome. They are optimists who want to build those better futures. But because we cover so many

Technical Evaluation

Transcription Quality

WER (Word Error Rate):

0.41

CER (Character Error Rate):

0.331

Acceptable accuracy for base Whisper model under GPU constraints

Segmentation Logic

Topic Coherence:

0.547

Boundary Accuracy:

0.32

Predicted Topics:

25

GenAI Usage

ASR Model:

Whisper Small (74M params)

LLM Model:

BART Large CNN (406M params)

Usage:

Transcription + Abstractive summarization + Topic labeling

Safety

- ✓ Fully local execution
- ✓ No user data stored
- ✓ No cloud API used

Cost

- ✓ Low GPU usage
- ✓ Zero API cost
- ✓ Minimal inference cost

Code Quality

- ✓ Modular pipeline
- ✓ Reusable components
- ✓ Well-documented

Conclusion

This project successfully presented an end-to-end Audio Podcast Transcription and Topic Segmentation system implemented using the Django framework. The system effectively processes raw podcast audio through comprehensive preprocessing, accurate transcription, and intelligent semantic segmentation to generate structured and user-friendly outputs.

Key Accomplishments

- ✓ Achieved acceptable transcription accuracy using Whisper base model under GPU constraints, enabling effective semantic topic segmentation
- ✓ Implemented unsupervised clustering-based topic detection using advanced NLP techniques and semantic embeddings
- ✓ Generated abstractive summaries and human-readable topic titles using state-of-the-art language models
- ✓ Provided comprehensive quality metrics and multiple export formats for diverse use cases

Critical Insight:

- Accurate transcription is vital for reliable topic boundary detection and effective chapter generation.
- Speech recognition quality affects NLP tasks like segmentation, summarization, and labeling.

Practical Impact:

- The system automates audio file uploads for full transcription and topic-segmented timestamp downloads. This enhances accessibility and navigation of long podcast content, making audio information more searchable and user-friendly.

Supporting Research Papers

- **Whisper Paper**: In Automatic Speech Recognition (ASR), speaker diarization improves transcription accuracy by associating text with individual speakers. This makes the transcript more readable and contextually meaningful, especially in overlapping conversations.
- **Hugging Face Whisper**: In Automatic Speech Recognition (ASR), speaker diarization improves transcription accuracy by associating text with individual speakers. This makes the transcript more readable and contextually meaningful, especially in overlapping conversations.
- **Librosa Documentation**: In Automatic Speech Recognition (ASR), speaker diarization improves transcription accuracy by associating text with individual speakers. This makes the transcript more readable and contextually meaningful, especially in overlapping conversations.
- **ASR Metrics (WER)**: In Automatic Speech Recognition (ASR), speaker diarization improves transcription accuracy by associating text with individual speakers. This makes the transcript more readable and contextually meaningful, especially in overlapping conversations.

Thank You