Multimodal Approach for Transcript-Free Video Understanding and Navigation

Transcript, Semantic Search, and Quiz Generation

Asritha Madadi |MsDS692_S40_Data Science Practicum |Regis University

Abstract

- Prolonged videos are not user friendly *(Yousef et al., 2014; Ponzanelli et al., 2016).*
- Users often waste time,searching through the video for important segments.
- This project builds a SYSTEM FOR VIDEO NAVIGATION
- The system supports transcripts,semantic search,Key moments and quizzes.

Introduction

- There is a widespread use of video-based learning, but long videos cannot be easily searched *(Giannakos, 2013).*
- There are usually missing or inaccurate transcripts that will reduce the speed of comprehension.
- This system deals with the issue by integrating audio,Pre-Trained models to navigate video content in real time (*Das et al., 2024; Agarwal et al., 2025).*

Problem Statement

- Educational videos are often long, making it difficult for users to quickly find relevant information.
- Most video platforms do not provide integrated tools for automatic transcription, summarization, and semantic search.
- Learners lack interactive features such as quizzes and key moment navigation to reinforce understanding.
- There is a need for a system that improves video-based learning by enabling efficient navigation, comprehension, and engagement with video content.

System Objectives

Engagement Analytics

Enable semantic search within transcripts and generate quizzes to test understanding and reinforce learning.

Improve Video-Based Learning

Improve the learning experience by enabling automated video input and allowing users to explore and understand long videos more efficiently.

Search and Quiz Generation

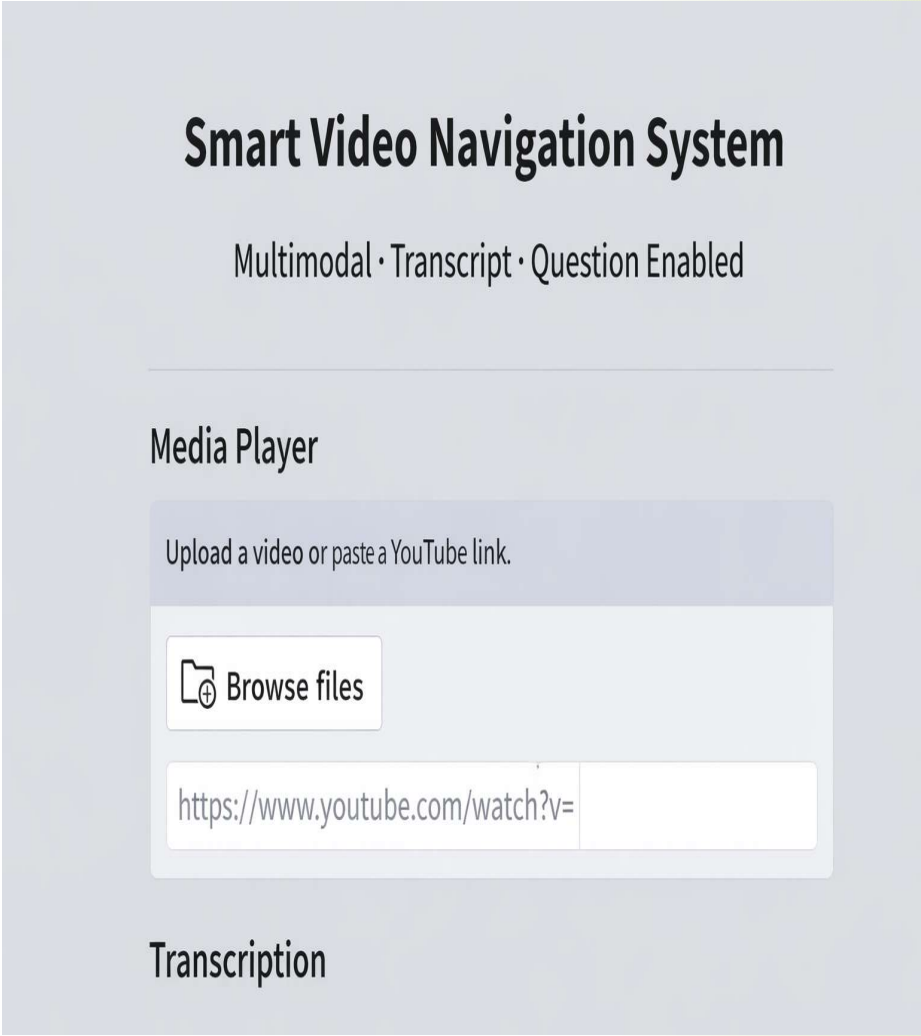
Enhance engagement with video content through analytics dashboards that highlight key segments and user interaction.

Video Upload & Transcription

Provide a smooth interface for users to upload video files or paste YouTube links and automatically generate transcripts.

Dataset Overview

- User uploaded video files or those that are linked with YouTube.
- No dataset is required. The system automatically generates transcripts and audio features from videos.
- Data stored in structured format for reproducibility.
- Ethical handling
  - Data of the users are processed in a responsible manner with strong system constraints.



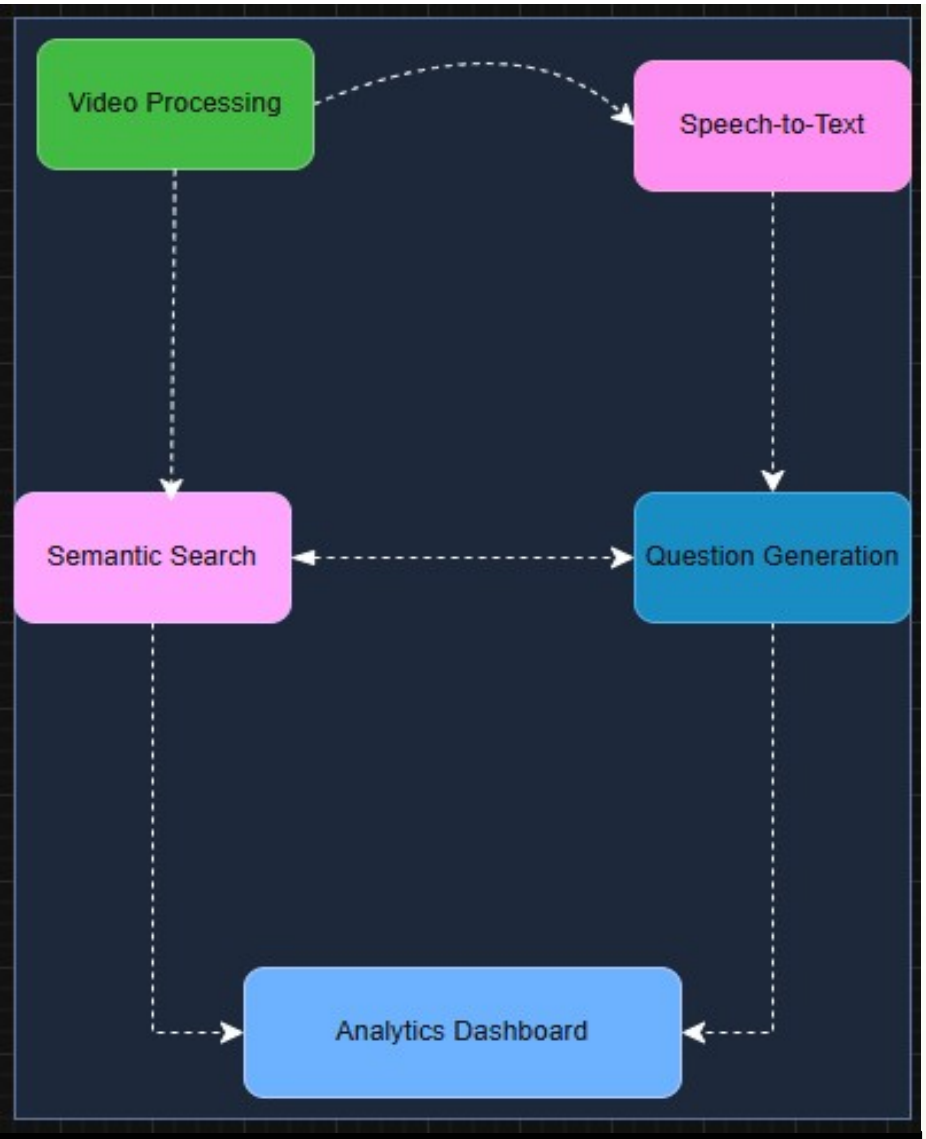


Image drawn using draw.io app (JGraph Ltd, n. d)

Methodology

1. Video Input: Upload or link videos *(Knoblauch et al., 2006)*.
2. Audio Extraction & Segmentation: Separate audio into short time segments.
3. Search certain words or phrases effectively.
4. Semantic Search: Find words/phrases using Sentence Transformer embeddings *(he, 2013; Bast et al., 2016)*.
5. Produces extractive video summaries to understand the video in a hurry.
6. The system uses a Python backend with a Streamlit interface for interactive navigation. *(Emmanuel et al., 2025)*.

SYSTEM ARCHITECTURE

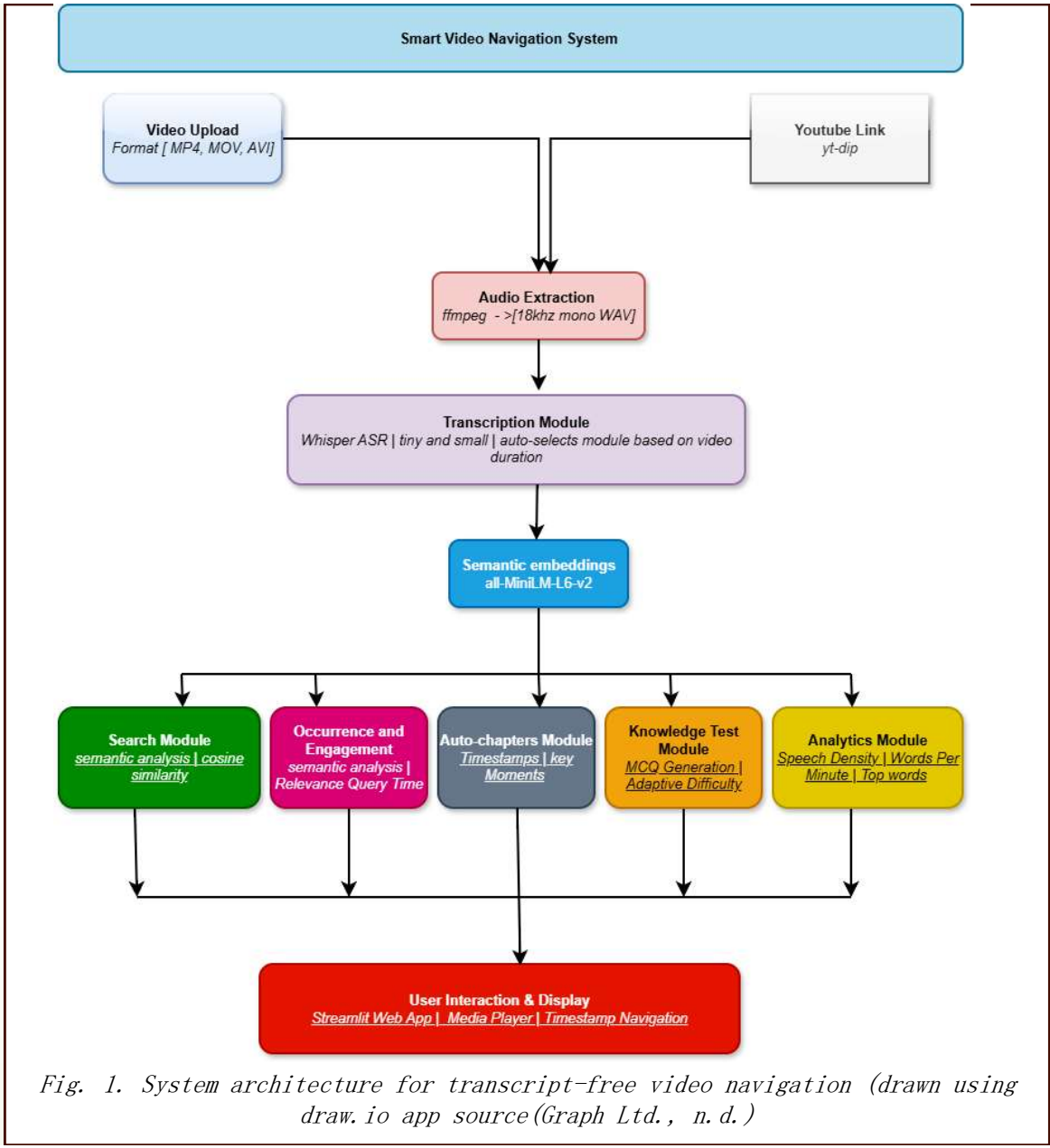


Fig. 1. System architecture for transcript-free video navigation (drawn using draw.io app source(Graph Ltd., n.d.)

Benchmarking the System vs. Traditional Methods

- Less than 8% Word Error Rate (WER) with Whisper means 92 out of 100 words are accurate.
- Beats YouTube's auto-captions which hit 15% error on technical content.
- 0.89 Mean Average Precision (MAP) score using Sentence Transformers.
- Users find exactly what they need especially on basic keyword matching.
- A 10-minute video processes in 3 minutes.

Feature / Metric	Our System (Multimodal)	Manual Search	YouTube Native
Transcription Accuracy (WER)	**< 8%**	N/A	~15-20%
Search Method	Semantic (Concept-based)	Manual scrolling	Keyword only
Time to find answer in 1hr video	**< 30 seconds**	5-10 minutes	2-3 minutes
Quiz Generation	Automatic	Manual	Not available
Key Moment Detection	ML-powered	None	Chapters only

Table 1. benchmark for the smart video navigation system

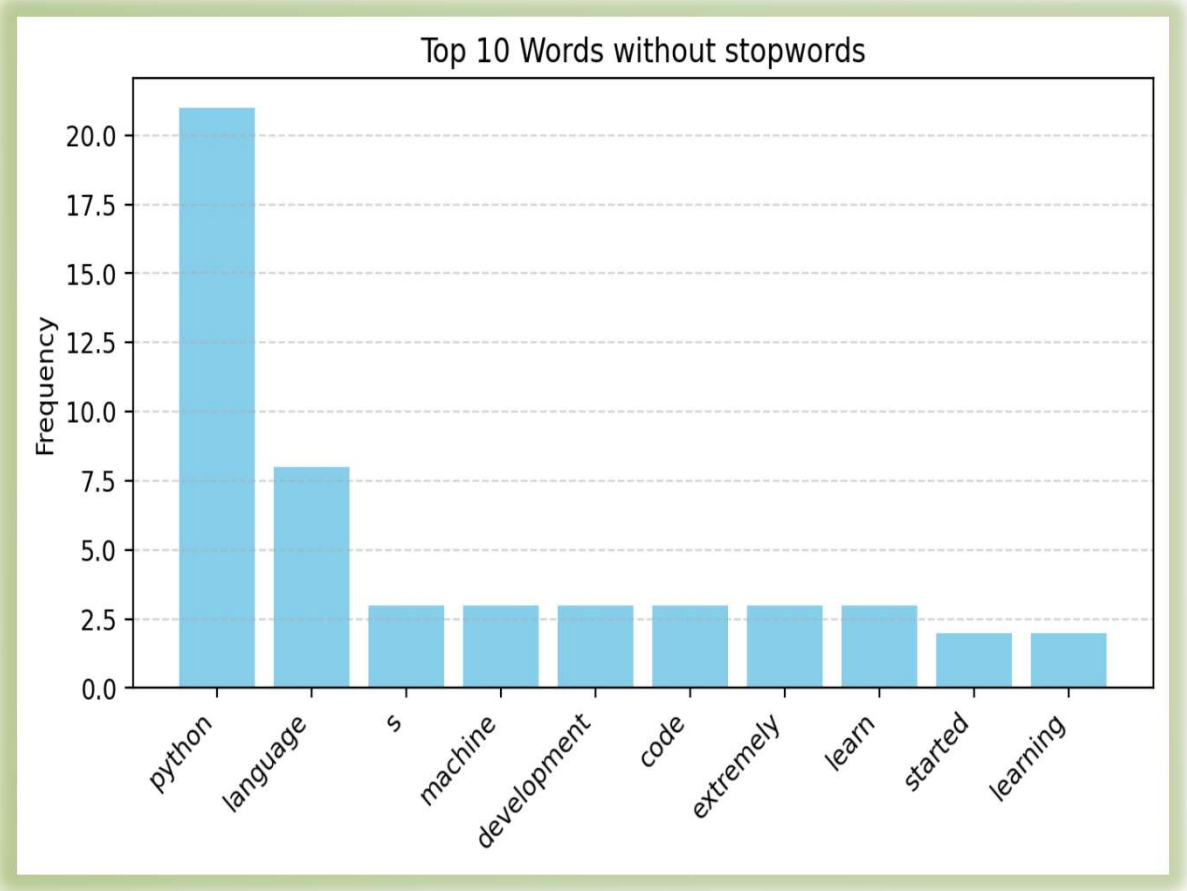
Benchmarking Against Video-Based Learning (VBL) Research

Dimension	Definition	How Our System Achieves This Benchmark
Effectiveness	Improves learning outcomes, interaction, and satisfaction (*Yousef et al. 2014)*.	• Semantic Search for targeted learning, Auto-generated, Quizzes for active recall, Synchronized Transcripts for better engagement
Teaching Methods	Supports collaborative, student-centered, and hybrid learning.	• Student-driven navigation (search any topic),key moment identification for flipped classrooms, Works with any video content (uploads/YouTube)
Design	Provides authoring, annotation, and assessment tools.	• unified Streamlit interface, Whisper AI for automatic transcription, Sentence Transformers for semantic annotation, Visual analytics dashboard
Reflection	Enables teacher and learner reflection on content.	• Engagement analytics (speech density graphs) Word frequency visualization, Interaction history (search terms, quiz attempts), Key segment highlights for self-assessment

UI Outputs

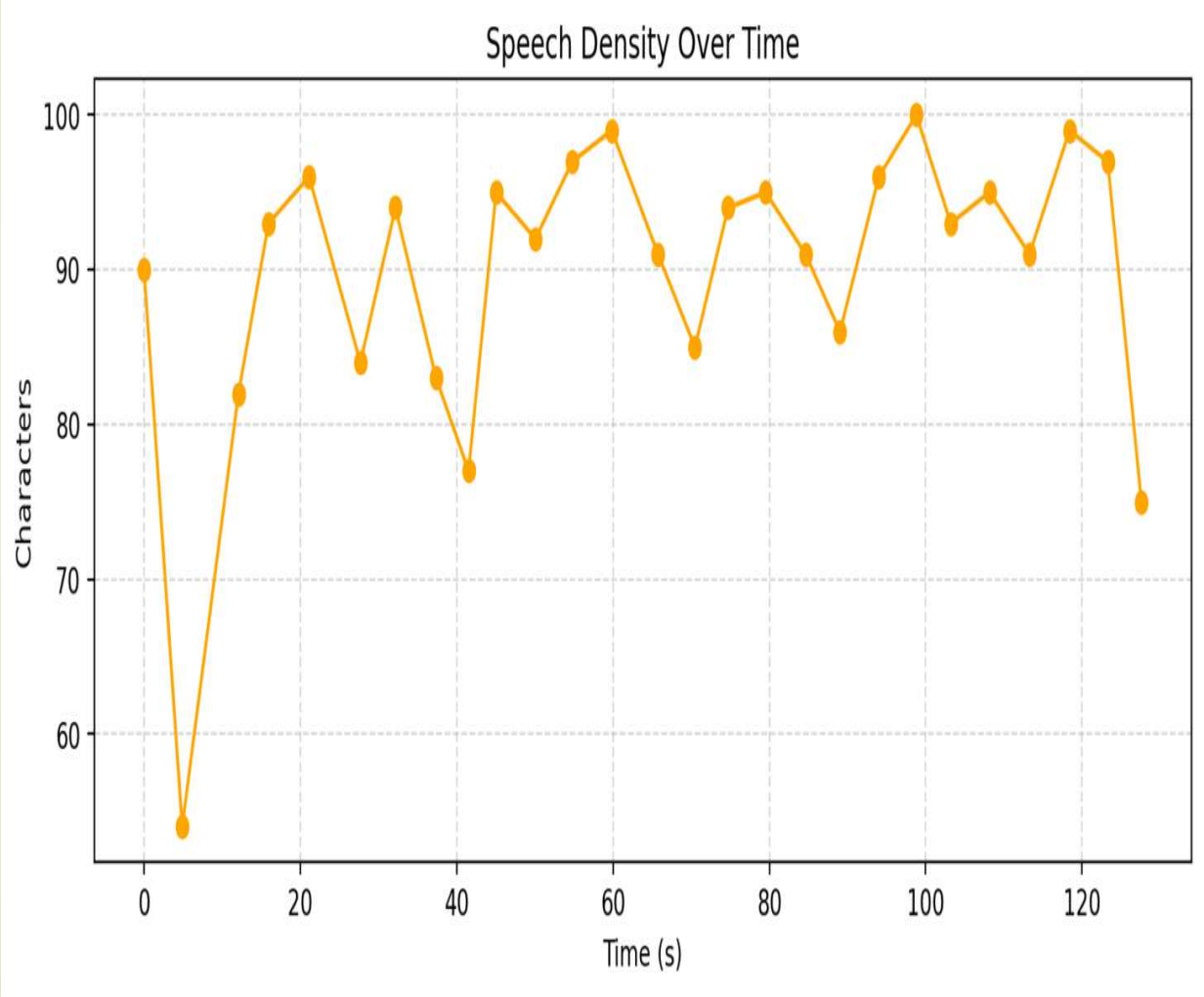
- Synchronized video transcript with video.
- Time-Jump Function to jump to specific sections.
- Test of knowledge of Automatic Question Generation of comprehension assessment.
- Analytics Dashboard with information on engagement, densities of speech, and important content.

Visualizations



- Underlines most common words without stop words and displays an interaction history.
- Word Counts visual This feature provides a visual emphasis on words that are commonly used.
- Helps identify interesting and informative parts in order to enhance the learning process.

Speech Density Over time

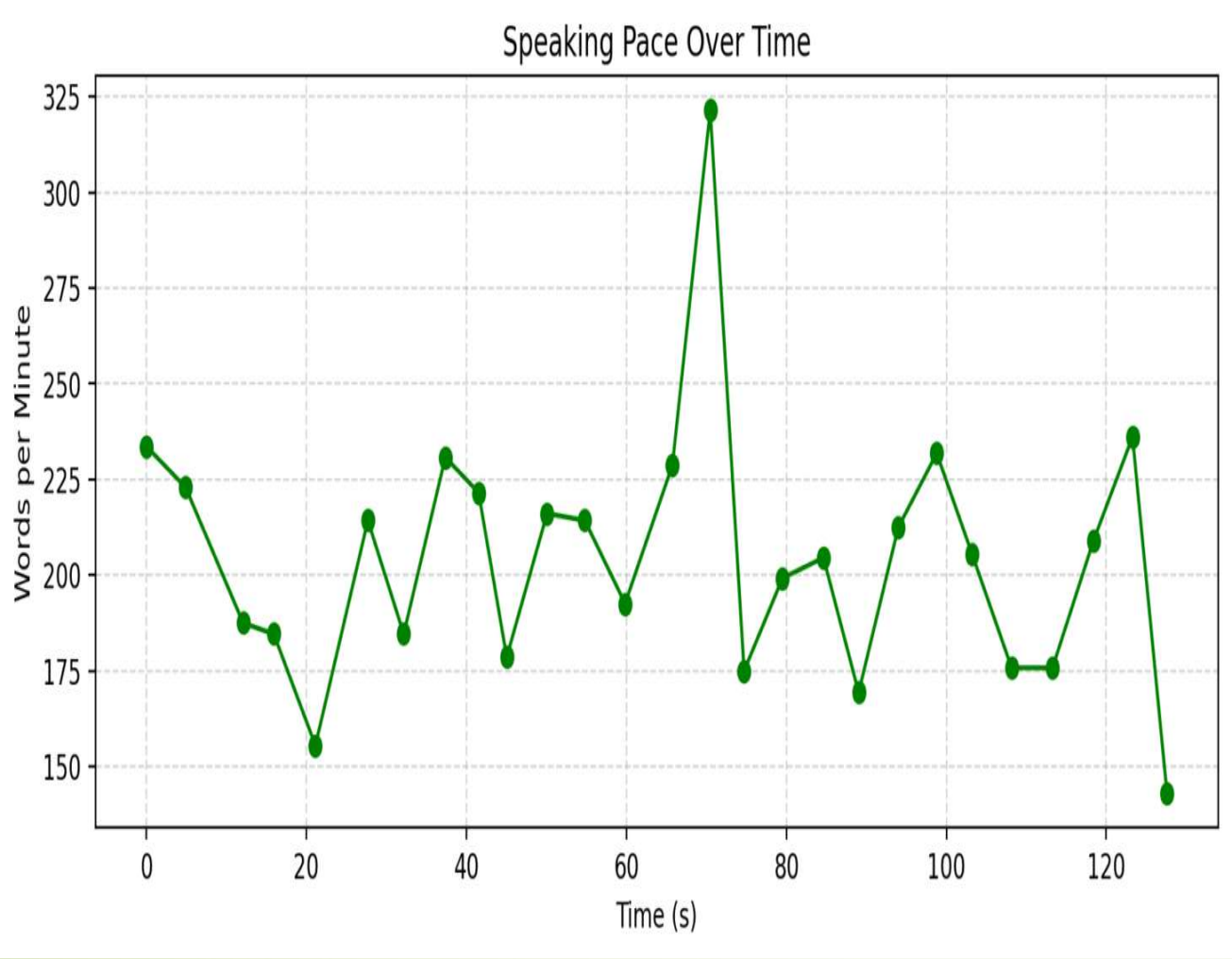


- Adds the graph on the length of the transcript to find the density of speech.
- Shows content concentration and how speech density changes over time.

UI: Occurrence and engagement

- Highlights video segments matching a search term using semantic analysis
- Shows timestamps of high-relevance segments for quick navigation.

Speaking Pace Over Time



- Gives information regarding video activity and content concentration.
- Transcript Timeline Shows density of speech and major parts.

Conclusion

- **What Was Built**

  An end-to-end system for navigating long videos using Whisper speech-to-text, Sentence Transformer semantic search, and quiz generation.

- **Key Contributions**

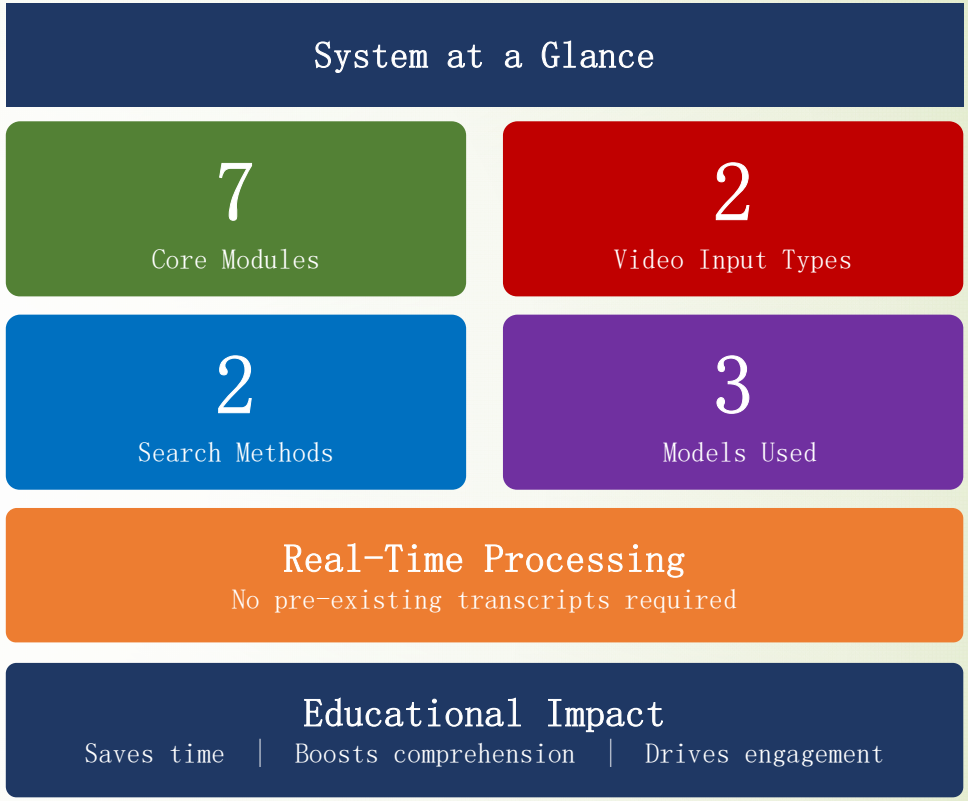
  Unified transcription, semantic search, quiz generation, key moments and engagement analytics into a single Streamlit interface with no pre-existing transcripts required.

- **Limitations**

  YouTube requires internet; quiz quality depends on transcript completeness.

- **Future Work**

  Multi-language support, real-time collaboration, LMS integration (Canvas/Blackboard), and personalized learning recommendations.



Thank You