SAMI 2021

19th IEEE World Symposium on Applied Machine Intelligence and Informatics.

January 21-23, 2021

Herl'any, Slovakia

A Framework for Lecture Video Segmentation from Extracted Speech Content

Dipesh Chand, Hasan Oğul

Østfold University College

Halden, Norway

Presentation Outline

- Background & Purpose
- Problem
- Dataset, Ground truth and Evaluation Metrics
- Proposed Framework
- Experimental Setup
- Result and Discussion
- Conclusion

Background & Purpose

Content-based Search: Search analyzes the contents of the video rather than the metadata such as keywords, tags, or descriptions associated with the video.

Lecture Video Segmentation: The goal of video segmentation is to divide the video stream into the basic elements of the index into a series of meaningful units.

Objective of this Research: To explore audio extracted from lecture videos to obtain **Textual** and **Acoustic** features and use them to **segment** the lecture video.

Problem

- Various topic contents are often covered in the lecture video.

- Retrieving the desired part of the video is still a very difficult and time-consuming process.

Dataset, Ground truth & Evaluation Metrics

Dataset

- 37 lecture video from Coursera
- Overall duration of lectures 2 hours, 45 minutes and 52 seconds.
- Total size 182.6 MB
- Videos are in MPEG-4 video (.mp4) format.

Ground Truth

- By analyzing Web Video Text tracks (WebVTT).
- Considering start timing of the full sentences.
- Total number of cues from WebVTT file 2568.
- Total number of Sentences 1688.
- Total segments in Groundtruth 616.

Evaluation Metrics

$$Precision = \frac{|S \cap G|}{|S|}$$

$$Recall = \frac{|S \cap G|}{|G|}$$

$$F1\ Score = 2.\frac{Precision.Recall}{Precision + Recall}$$

Proposed Framework

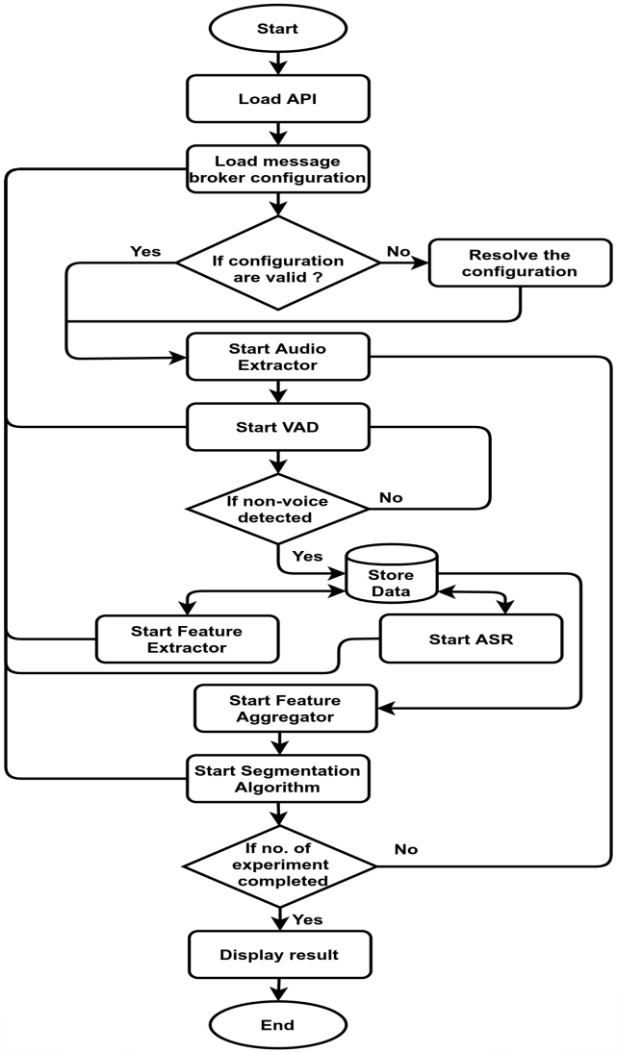

Figure 1: Flowchart of lecture video segmentation model

Proposed Framework, Continued

Feature Extraction Process

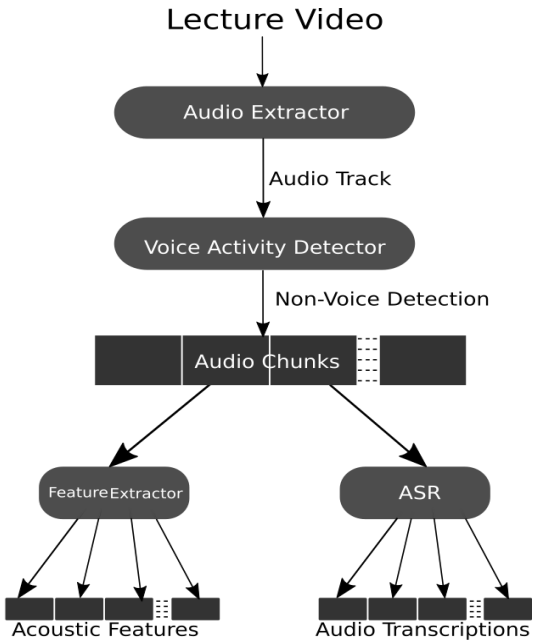

Figure 2: Feature extraction process from lecture video

Proposed Framework, Continued

Segmentation Process

Multi-objective model:

- $Ui = \alpha(F\,i + V\,i) + \beta \cdot P\,i + \gamma \cdot Di$ …….(1)
- $Di = Dcos(i - 1, i) + Dcos(i, i + 1)$…….(2)

$$maxT \sum_{i=1}^{n} U_i \cdot X_i - \sum_{i=1}^{n} X_i \qquad \text{…….(3)}$$

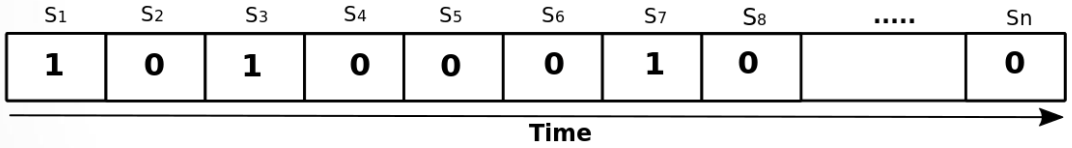

Figure 3: Representation of lecture video segment as a chromosome

Proposed Framework, Continued

Segmentation Process

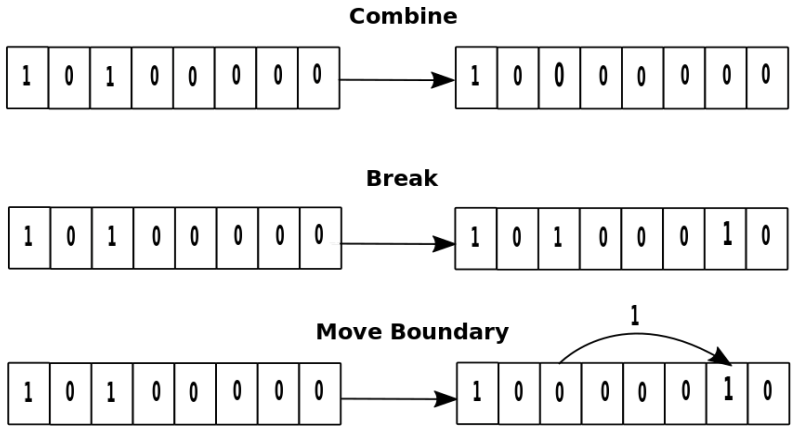

Figure 4: Illustration of local search movement

Experimental Setup

- Used Tools and Technologies: Dockers, FFmpeg, VAD, aubio, Pocketsphinx, Word2Vec, Genetic algorithm.
- Proportional parameters α, β, and γ are set to 0.05, 1, and 10
- For Genetic Algorithm:

- Number of generations = 1000
- Population size = 200 individuals
- Crossover = 35% of the individuals
- Mutation = 7% of population.
- Local Search = 35% of best solutions

Result and Discussion

Number of segments	Number of Matched segments	Precision	Recall	F-Score
518	355	0.69	0.58	0.63

Table 1: Outcome of proposed model

Comparison with Similar Models:

Method	Precision	Recall	F-Score
Our Model	0.690	0.580	0.630
System 1	0.465	0.491	0.477
System 2	0.400	0.480	0.400

Table 2: Comparison between proposed system and other similar systems

Conclusion

- Design and tested a system for Lecture Video Segmentation.
- Designed a system capable of using open source tools and algorithms.
- Proposed framework which can handle multiple number of lecture videos continuously.
- Improvement in performance than other similar models.

THANK YOU

Questions?