A Project/Dissertation Review-1 Report

on

Youtube Transcript Summarizer

Submitted in partial fulfillment of the

requirement for the award of the degree of

B.Tech – Computer Science and Engineering

Under The Supervision of

Ms. J. Angelin Blessy

Assistant Professor

Submitted By:

Varun Sharma

19021011906

19SCSE1010764

Samarth Singh

19021011408

19SCSE1010219

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA INDIA October,

2021 **Abstract**

Nowadays, the information is widely available in different medias: TV, social networks, newspapers, etc. The main difference in comparison to what we had one or two decades before is that people can access to the videos of social networks in foreign languages. When, the video does not necessitate any understanding, there is no main problem. In the opposite, when the information necessitates to understand the language, a human being is limited in terms of mastering foreign language, even if YouTube proposes a rough translation of some contents.

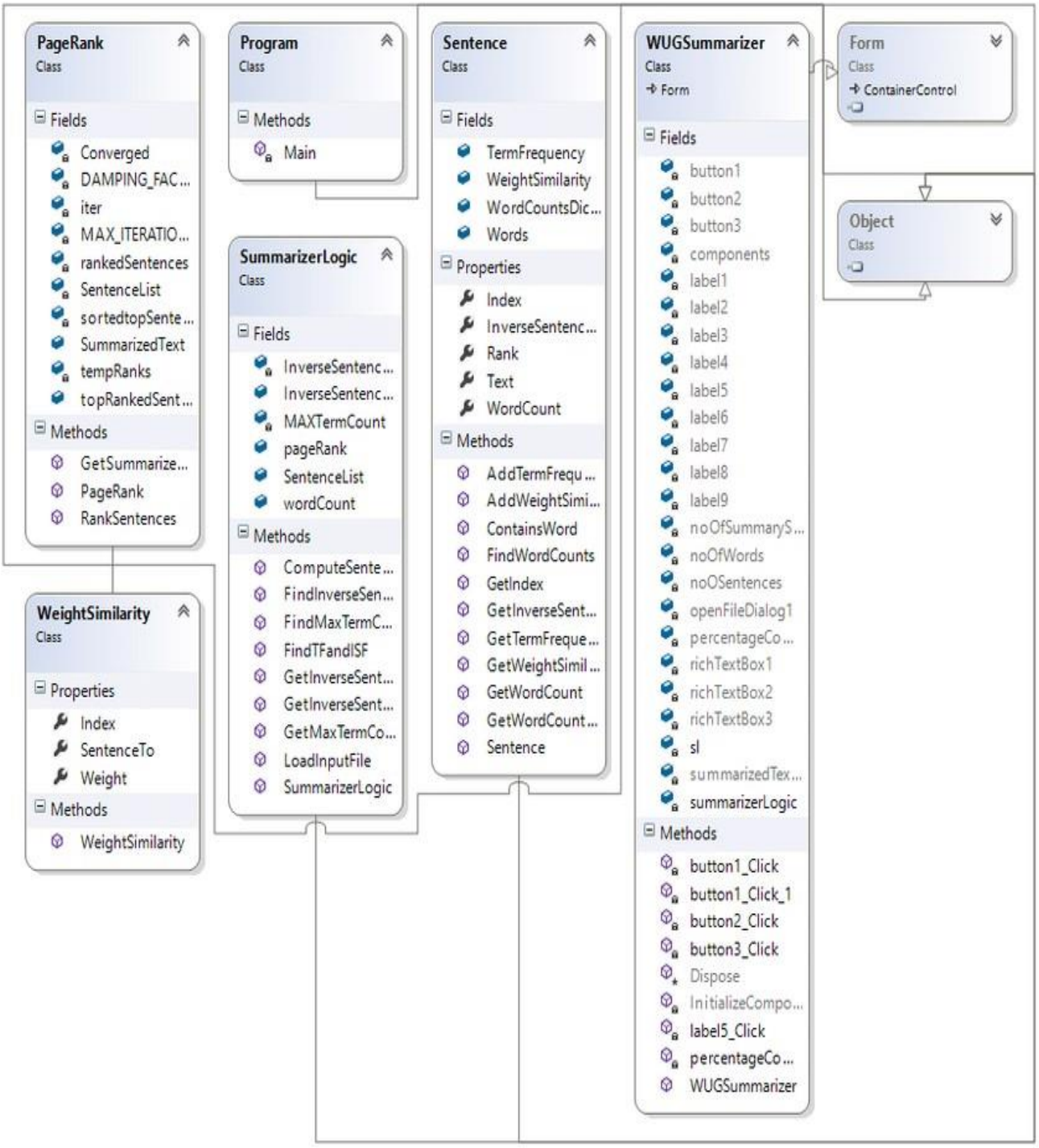
The main goal of this project is to understand the content of a video in a foreign language. In this work, we consider the understanding process, such as the aptitude to capture the most important ideas contained in a media expressed in a foreign language. In other words, the understanding will be approached by the global meaning of the content of a support and not by the meaning of each fragment of a video. Several stumbling points remain before reaching the fixed goal. They concern the following aspects: Video summarization, Speech recognition, Machine translation and Speech segmentation. All these issues will be discussed and the methods used to develop each of these components will be presented. A first implementation is achieved and each component of this system is evaluated on a representative test data.

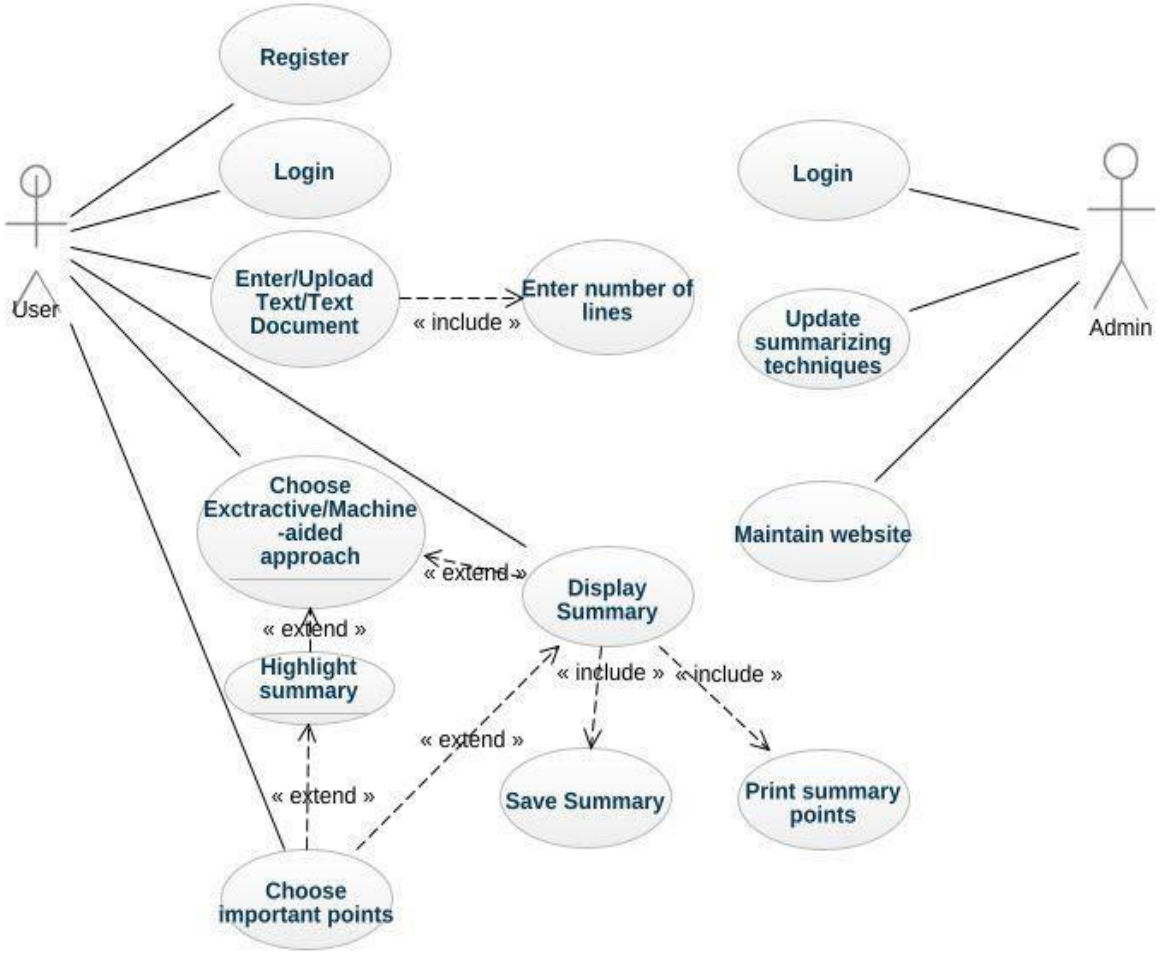
Several methods have been proposed for analyzing video and deriving compact representations for users to browse. One class of these methods is video summarization. The summarization process generally contains two main steps. In the first step, the source video is analyzed and divided into coherent video segments. This is often done by performing shot boundary detection on the video track. These segments are then assigned importance scores by using various combinations of visual, audio, textual, and other features extracted from the video stream. The work reported in1 is a good example of a multi-modal segments evaluation.

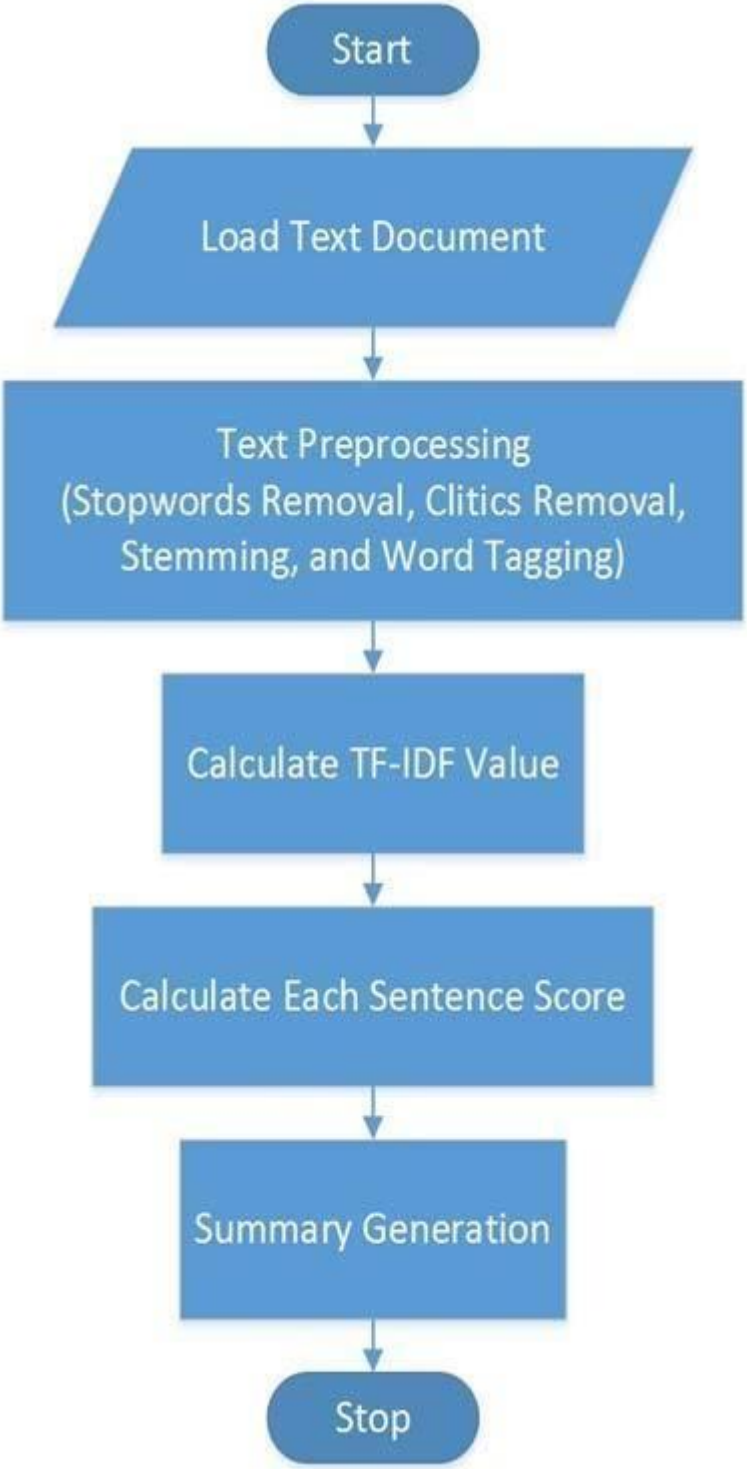
In this paper we propose an algorithm to automatically summarize educational video programs. We use concepts from text summarization, applied to transcripts derived using automatic speech recognition. We also use temporal analysis of pauses between words to detect sentence boundaries. We will see that the dominant word pair selection algorithm works well in identifying main topics in video speech transcripts. We will also develop an experimental design for a user study to judge the quality of the generated summaries using quizzes based on questions about original program content that the subjects answered after watching summaries generated from the programs. We discuss various shortcomings of the quiz method of evaluating summaries. The problem of deriving good evaluation schemes for automatically generated video summaries is still a complex and open problem.

List of Tables

List of Figures




Register
Login
Enter/Upload Text/Text Document
« include »
Enter number of lines
User
Login
Update summarizing techniques
Admin
Choose Exctractive/Machine -aided approach
« extend »
Display Summary
Maintain website
« extend »
Highlight summary
« include »
« include »
« extend »
Save Summary
Print summary points
« extend »
Choose important points


Start
Load Text Document
Text Preprocessing
(Stopwords Removal, Clitics Removal,
Stemming, and Word Tagging)
Calculate TF-IDF Value
Calculate Each Sentence Score
Summary Generation
Stop

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

Table of Contents

CHAPTER-1 Introduction

Due to advances in video streaming and expansion of low-cost storage media, digital video has become an important factor in education, entertainment, and commerce. Consequently, there has been a great interest in designing and building systems that organize and search video data based on its content.

In addition to search capabilities, such systems should be able to derive intuitive and compact data representations so that users may easily and quickly browse through the whole database or through the results of a query. Such representations rapidly provide the user with information about the contents of the particular sequence being examined while preserving the essential message.

Developing efficient representations for video browsing presents some unique algorithmic challenges, as well as new technical challenges. Video is a sequential and information-rich medium. It includes audio and motion, and it carries long temporal contextual relationships between shots and scenes. In contrast to images in an image database, manipulation of video is inherently more complex.

For example, images can be represented as thumbnails and users can easily judge relevance of these images at a glance. The same task is very time consuming for video sequences, where one hour is composed of more than 100,000 frames, divided into hundreds of shots. Additionally, the audio, which often conveys much of the information (e.g., a video of a talking head accompanied by slides), is even harder to browse in an efficient manner.

Chapter-1

Tools and Technology Used

Hardware Requirements:

- Operating system- Windows 7,8,10 •
- Processor- dual core 2.4 GHz (i5 or i7 series Intel processor •
- or equivalent AMD) •
- RAM-4GB

Software Requirements:

- Python
- Pycharm
- PIP 2.7
- Jupyter Notebook
- Flask
- HTML
- JavaScript
- CSS
- Cloud hosting (AWS/Azure/Heroku)

Chapter-2 Literature Survey

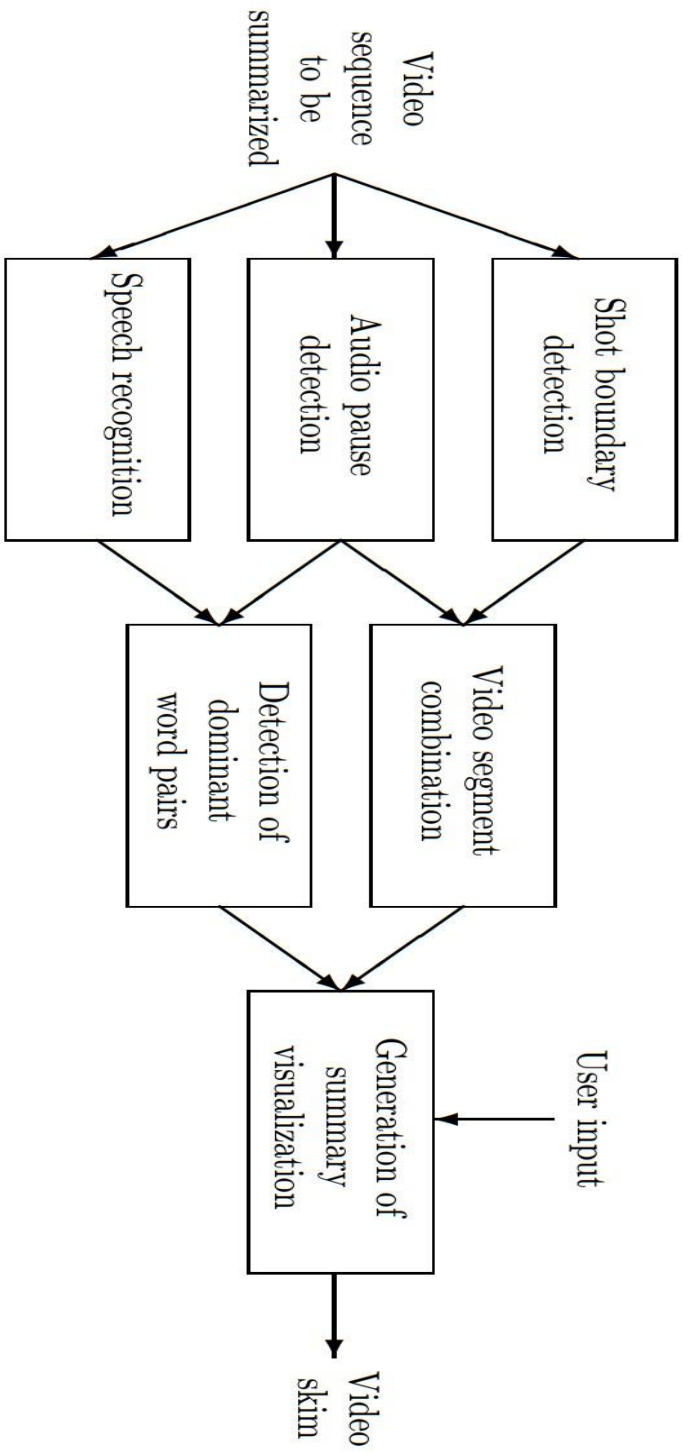
Nowadays, the information is widely available in different medias: TV, social networks, newspapers, etc. The main difference in comparison to what we had one or two decades before is that people can access to the videos of social networks in foreign languages. When, the video does not necessitate any understanding, there is no main problem. In the opposite, when the information necessitates to understand the language, a human being is limited in terms of mastering foreign language, even if YouTube proposes a rough translation of some contents.

Our main objective is to make available a system, helping people to understand the content of a source video by presenting its main ideas in a target understandable language. The understanding process is considered here to be the comprehension of the main ideas of a video. We think that the best way to do that, is to summarize the video for having access to the essential information. Henceforth, we focuses on the most relevant information by summarizing it and by translating it to the user if necessary. As a result, we will permit to have another side of story of an event since we can, for instance, have the Russian version of the war in Syria.

Several skills are necessary to achieve this objective: video summarization, automatic speech recognition, machine translation, text summarization, etc. Each output of a sub-system, we can enrich in upstream or downstream the other modules. That makes us working in such as a workflow where the flow refers to the information necessary for a component. In this article, we will present the first result that works such as a pipeline system connecting the output of each component to the input of the next one.

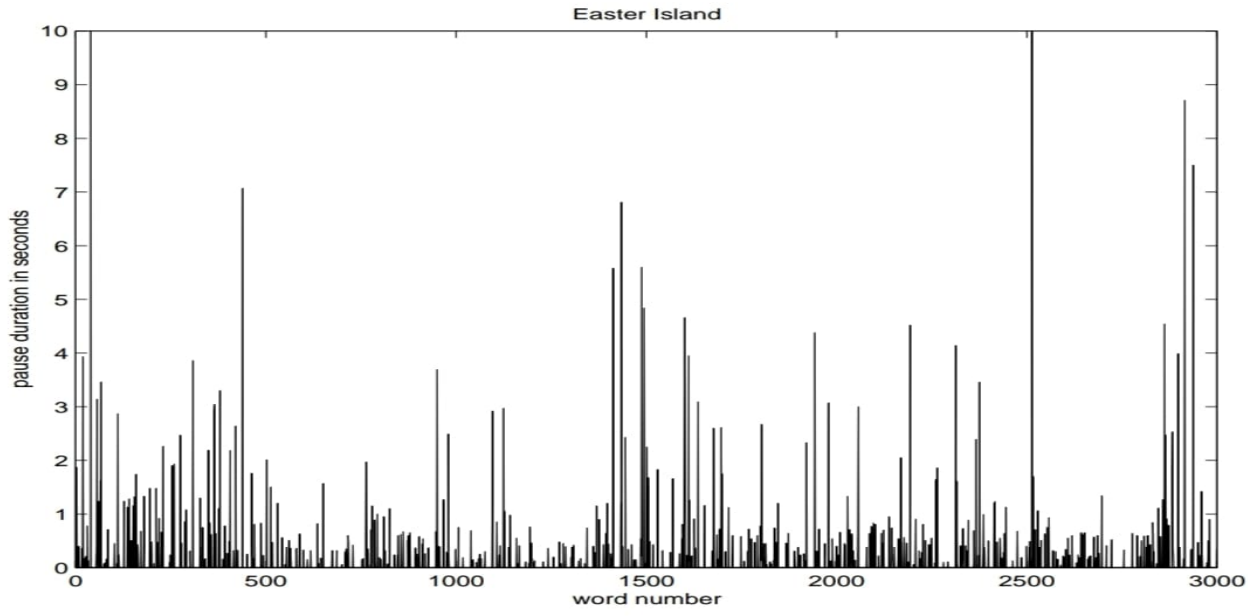
Chapter-2

Video Skimming Diagram



Chapter-3

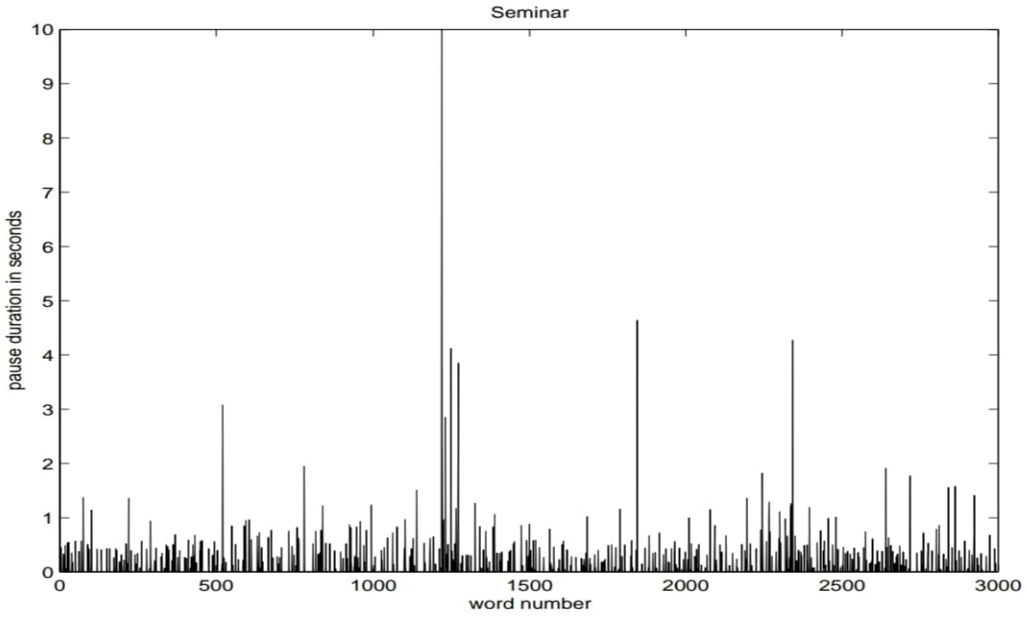
In user studies of summarization algorithms it has been found that users find it very annoying when audio segments in the summary begin in the middle of a phrase. Therefore, instead of using shot boundaries to segment video into segments, we have decided to use audio pause boundaries. In other words, the segments of video frames that the video is divided into are not determined using visual features, as was done in most of the previous studies, but are determined by detecting large interword pauses of the speaker in the audio. An advantage of this method is that it avoids having very long shots in the summary which commonly occur in videos of presentations and meetings where the camera is fixed on the speaker for a long time. The interword pause distributions of two video sequences, a wildlife documentary and a seminar, are shown in Figure 2. The pauses in the documentary are larger since this is an educational program with a well-defined script and the speaker speaks at a relaxed pace for all audience to easily follow. On the other hand, the pauses in the seminar video are generally quite short, corresponding to free-flowing natural speech, with “vocalized pauses” when the speaker is thinking about how to phrase a thought. These plots indicate that the interword pause durations show large variations from speaker to speaker and between video genres. Hence using a global threshold to detect segment boundaries does not work in practice.



There are many techniques available to detect pauses in speech. However, we detect pauses with a simple heuristic using the time stamped speech transcript file generated by ViaVoice at the completion of the speech recognition step of the CueVideo indexing process. This file contains the time stamp for each word recognized by the speech recognition program together with its length. We use a sliding window scheme to detect pauses between segments which is similar to the shot detection method proposed in. First, using the time stamps and word lengths from the speech transcript file, the durations of the pauses between all the words in the video are computed. Then, a symmetric window of size 2m + 1 is placed around the ith pause and a segment boundary is declared between words i and i + 1 whenever

1. the length of the ith pause is the maximum within the window, and

2. it is also n times the value of the second maximum in the window



Conclusion

In this paper we propose an algorithm to automatically summarize educational video programs. We use concepts from text summarization, applied to transcripts derived using automatic speech recognition. We also use temporal analysis of pauses between words to detect sentence boundaries. We used the Transformers Summarization Pipeline system to perform the media processing. We have shown that the dominant word pair selection algorithm works well in identifying main topics in video speech transcripts. We have also developed an experimental design for a user study to judge the quality of the generated summaries using quizzes based on questions about original program content that the subjects answered after watching summaries generated from the programs. We discussed various shortcomings of the quiz method of evaluating summaries. The problem of deriving good evaluation schemes for automatically generated video summaries is still a complex and open problem.