# EVALUATION OF SPEECH AND TEXT-BASED INDEXING FOR CLASSROOM LECTURE VIDEOS

A Thesis Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Mahima Joshi

December 2014

# EVALUATION OF SPEECH AND TEXT-BASED INDEXING FOR CLASSROOM LECTURE VIDEOS

**Mahima Joshi**

APPROVED:

**Dr. Jaspal Subhlok, Advisor**

**Dr. Olin Johnson**

**Dr. Chris Barr**

**Dean, College of Natural Sciences and Mathematics**

# Acknowledgements

I would like to express my deep gratitude to my advisor, Dr. Jaspal Subhlok, for his guidance and support throughout the course of this thesis. I also want to thank my other committee members, Dr. Olin Johnson and Dr. Chris Barr, for being a part of my thesis and constantly encouraging and supporting me. I would like to thank Dr. Nauhad Rizk and Dr. Johnson for providing indispensable input for this thesis. I would also like to thank Tayfun Tuna for his technical help and suggestions, without which this thesis would not have been successful. Lastly, my greatest gratitude goes to my husband, Dila R Banjade, and my family for their love, encouragement, and support for making this thesis a success.

# EVALUATION OF SPEECH AND TEXT-BASED INDEXING FOR CLASSROOM LECTURE VIDEOS

An Abstract of a Thesis

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Mahima Joshi

December 2014

# Abstract

Lecture videos are useful and great learning resources. At the University of Houston, videos are widely used throughout departments within the College of Natural Sciences and Mathematics such as Computer Science, Biology and Biochemistry, Earth and Atmospheric Sciences, etc. Since most videos are very long, it is difficult to directly access the required topic within a video. The ICS (indexed, captioned, and searchable) videos project provides students direct access to a topic within video lectures by providing index points representing the topic. These index points are generated using text from the extracted images using OCR (optical character recognition) technology. Index points are assigned with the assistance of an indexing algorithm that determines topic change based on text similarity.

We present a topic-based lecture video segmentation using speech text/captions. The purpose of this thesis is to utilize the spoken text of a lecture video to assign index points using an underlying text-based indexing algorithm. To achieve this goal, a set of twenty-five lecture videos was taken from various departments at the University of Houston and Coursera website. The captions were produced with the assistance of the YouTube Speech Recognition System. The performances and limitations of OCR text, uncorrected/original speech text, and corrected speech text-based indexing was analyzed. The results indicate that slide text-based indexing yields 4% better results than spoken text-based indexing. The corrected speech text/caption provides better indexing results (11%) where OCR text fails to perform and the results closely matched the ground truth.

The error analysis done on speech texts and slide texts prove that poor OCR text and caption quality are some of the main issues that hamper indexing accuracy.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1. Introduction

Recently, a lot of universities and research institutions are recording and publishing their lecture videos online for student access. Online courses have become a popular source of learning due to the recent advancements in technology. Information technology has a big role to play in audiovisual recording and availability. E-lecturing has become a common trend among universities, such as the University of Houston, MIT Open Courseware, and Stanford University. Massively open online courses (MOOC) are popular worldwide for providing online lectures in different fields and are a great source of learning. Students can easily access these online materials anytime, anywhere, without being physically present in class [1].

The lecture videos that capture overall classroom interaction provide actual classroom experience to students who are not able to attend class. However, for long lecture videos, students may have to spend a long time accessing the specific information within that lecture video. Time constraints associated with lecture video retrieval can reduce learning efficiency for a student. Therefore, video indexing could be useful for the advancement of online learning. The main problem thus becomes the efficient retrieval of the appropriate information in a long lecture video.

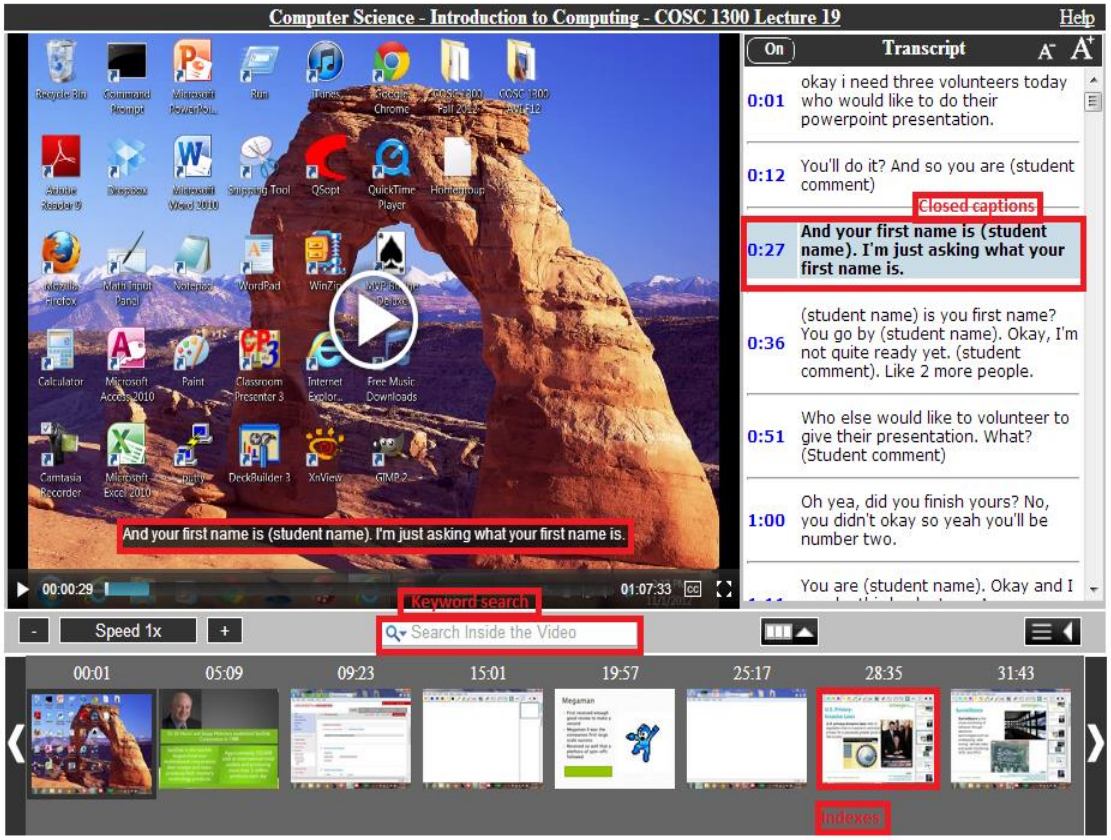
## 1.1 Motivation

Online video lectures have become a day-to-day educational resource for students in higher education [2]. A major weakness of the video formats is the size and continuity of the video. Lecture videos present a challenge, as it is difficult to access the content or a topic quickly from a long video. The segmentation and indexing of a video lecture is an important step in making sure specific content is easily retrieved [3]. The benefit of lecture videos is that the students have an opportunity to watch the content as many times as they want [1]. These lectures, if available online, can be accessed regardless of time and location. Lecture videos contain useful information, such as text in the image of each slide and the audio content. The keywords in such cases can provide a small description of the lecture and can be used for the information retrieval process [4]. Utilizing the texts in a lecture video has been proven theoretically more efficient in the indexing of a video lecture, according to a survey conducted at the University of Houston [5], for the usage of ICS (indexed, captioned, and searchable) videos to determine the effectiveness of the indexing. The results indicate that the indexing process produced acceptable results, but there is room for improvement. The indexing currently done on the ICS project uses text extracted from the images in a lecture; however, speech text is also an important element in a video lecture [6]. Indexing allows the captions to be utilized effectively. Indexing and captioning can improve the learning process in two ways. First, it can make video contents searchable, since audio of a video lecture contains useful information that sometimes might not be present in the text extracted from the images of view graphs. Secondly, captions are helpful to the students whose first language is not English. The speech to text information is “well suited for content based lecture video retrieval” [6].

Therefore, the captions of a video lecture could be utilized for content-based indexing or topic-wise segmentation.

## 1.2 Background: ICS Video Player

At the University of Houston, recording online lectures and making them accessible to students is a common trend. Many instructors are using ICS videos. The survey results at the University of Houston have indicated the efficiency of the ICS video player in student learning. The main goal of the ICS player is to make online content as widely available as possible and enhance the student learning process [1] [5] [7] by providing indexes, a keyword search capability, and captions of the instructor's audio. The work presented in this thesis is a part of the ICS videos project indexing. This thesis aims to create and discuss the development of an audio-based text indexing system as an extension of the current indexing systems. The idea for the audio-based text indexing system comes from the caption section within the videos project. The video lecture recorded by the instructor contains presentations in the form of view graphs like PowerPoint and audio recordings as well. This is achieved by recording the computer screen and placing a microphone to record audio. The recorded video is uploaded to the ICS server framework for further processing. The ICS video player is a customized player and the screenshot is presented in Figure 1.1.



0:01 okay i need three volunteers today who would like to do their powerpoint presentation.

0:12 You'll do it? And so you are (student comment)

0:27 And your first name is (student name). I'm just asking what your first name is.

0:36 (student name) is you first name? You go by (student name). Okay, I'm not quite ready yet. (student comment). Like 2 more people.

0:51 Who else would like to volunteer to give their presentation. What? (Student comment)

1:00 Oh yea, did you finish yours? No, you didn't okay so yeah you'll be number two.

You are (student name). Okay and I

And your first name is (student name). I'm just asking what your first name is.

**Figure 1.1: An overview of ICS player and its components [8]**

### 1.2.1 Indexing

The indexes in a video contain the topics in the lecture so that students can easily access the point or topic of interest within a video [5]. Each video lecture consists of different sub-topics, and indexing is the process of identifying sub-topics within a lecture video. Significant scene changes are detected in a video, and these segments are marked as transition points. Subsets of these transition points are the index points that represent a different sub-topic. The selection of index points is done using a text-based indexing algorithm. The main criteria for index point selection are based on the image difference between previous and successive transition points. The transition point and index point are represented in Figure 1.1.

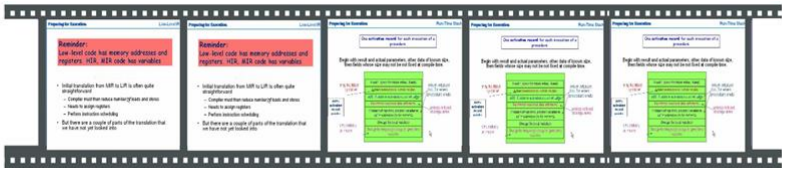


**Figure 1.2: Transition point in a video, third frame is an index point (new transition point) [5]**

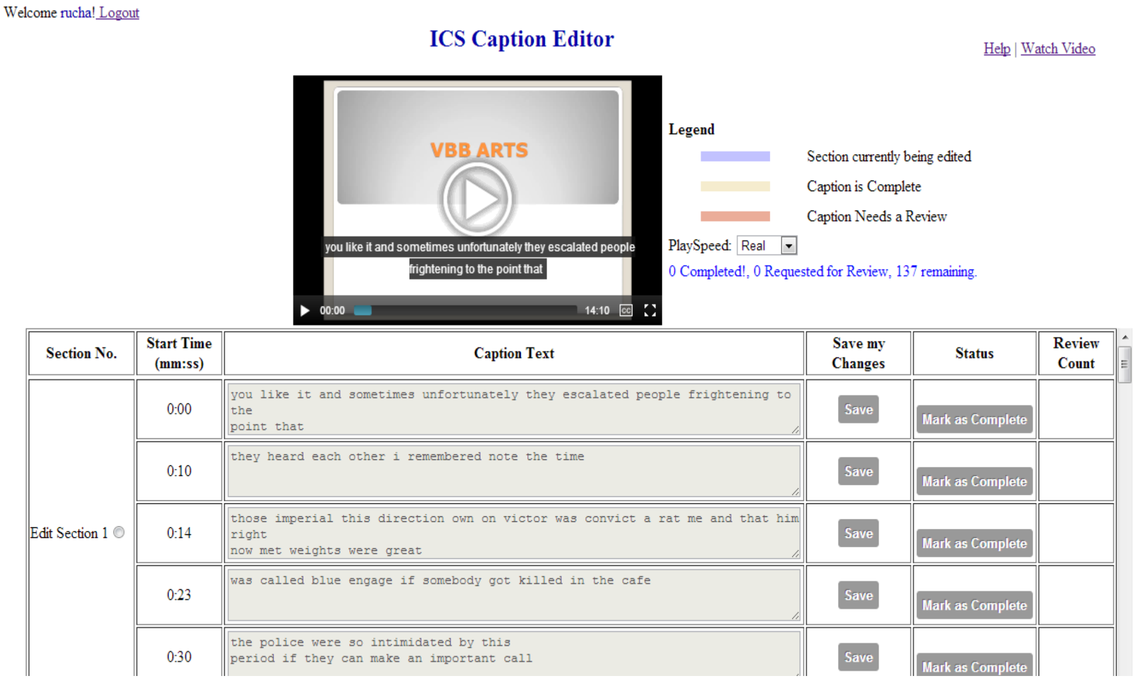
Evaluations from previous work showed that index points were accurate most of the time but do not always represent the topic [5]. The indexing option in ICS player is significant for the students, as they can identify their topic of interest.

### 1.2.2 Captioning

As seen from Figure 1.1, the ICS video player provides a closed captioning option on the right-hand side panel of the video player. Captioning is done to enhance the student accessibility of the video lectures. The students have the option to turn off/on these captions. Captions are displayed on the video screen along with the view graphs, and the complete transcript is displayed on the right-hand side in the ICS video player. The complete transcript is accessible via scrolling through the transcript panel.

#### 1.2.2.1 ICS Caption Editor

Through the use of caption editing, video lecture captions are corrected. For example, YouTube gave erroneous captions and the text accuracy was not 100%; therefore, it was decided to correct the captions of some of the video lectures out of a set of twenty-five. For this purpose, the ICS videos project's caption editor tool was used. It is a custombuilt web-based tool that assists in editing captions [9]. To begin, there is a link that has been given to the users of the caption editor and the user needs to login to be able to start the editing process. The caption editor provides an easy to use and efficient method to caption the audio-text. Caption editing was really helpful in the research. A snapshot of the ICS caption editor is given in Figure 1.3.



Welcome rucha! Logout

ICS Caption Editor

Help | Watch Video

you like it and sometimes unfortunately they escalated people frightening to the point that

Legend

🟦 Section currently being edited

🟨 Caption is Complete

🟥 Caption Needs a Review

PlaySpeed: Real

0 Completed!, 0 Requested for Review, 137 remaining.

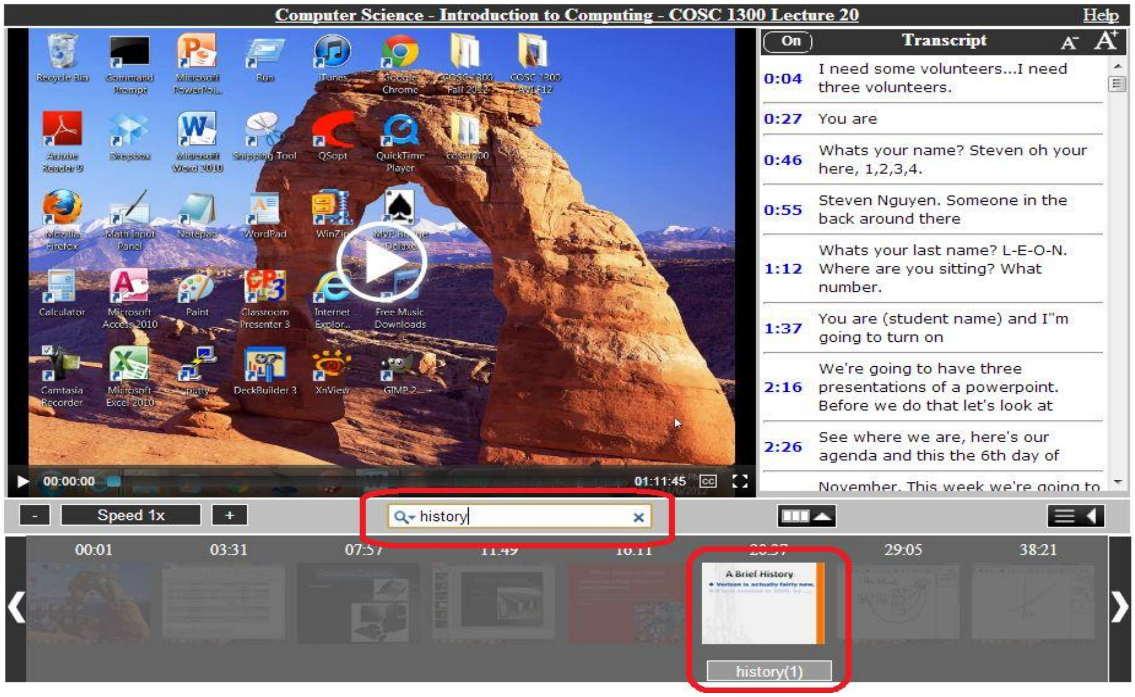
| Section No. | Start Time (mm:ss) | Caption Text | Save my Changes | Status | Review Count |
|---|---|---|---|---|---|
| Edit Section 1 ○ | 0:00 | you like it and sometimes unfortunately they escalated people frightening to the point that | Save | Mark as Complete | |
| | 0:10 | they heard each other i remembered note the time | Save | Mark as Complete | |
| | 0:14 | those imperial this direction own on victor was convict a rat me and that him right now met weights were great | Save | Mark as Complete | |
| | 0:23 | was called blue engage if somebody got killed in the cafe | Save | Mark as Complete | |
| | 0:30 | the police were so intimidated by this period if they can make an important call | Save | Mark as Complete | |

**Figure 1.3: ICS Caption Editor [9]**

After logging into the caption editor interface, captions were corrected by playing the video. On the left side of the panel is the start time and in the center is the original caption text/uncorrected speech text generated by YouTube. These were corrected manually according to the start and end times of each section. The changes were saved, and afterward the entire corrected caption text was downloaded from the server. Correcting captions is a time-consuming process; however, the ICS caption editor has a crowd source, captioning feature. The corrected captions were used for evaluation as discussed in Chapter 5.

### 1.2.3 Keyword Search

In Figure 1.1, there is text box at the center of the video panel for a keyword search where students can enter the terms to search among the entire video database. The search option enables the search functionality inside the video. The steps to achieve a keyword are as follows. First, the indexer identifies the segments in a video and the transition points in the form of images. The texts on video frames that are extracted and stored in a database are identified by OCR technology [7].  Second, when a student searches for a keyword from the ICS interface, the related segments are identified and presented as navigable search results [9]. Figure 1.4 represents the search functionality along with the highlighted corresponding index point.



0:04 I need some volunteers...I need three volunteers.

0:27 You are

0:46 Whats your name? Steven oh your here, 1,2,3,4.

0:55 Steven Nguyen. Someone in the back around there

1:12 Whats your last name? L-E-O-N. Where are you sitting? What number.

1:37 You are (student name) and I"m going to turn on

2:16 We're going to have three presentations of a powerpoint. Before we do that let's look at

2:26 See where we are, here's our agenda and this the 6th day of

November. This week we're going to

**Figure 1.4: ICS video player with Search functionality**

## 1.3 Goal and Summary of Research

The main goal of this thesis is to develop a speech text-based video indexing method for the topic-wise segmentation of lecture videos. Every index point represents a topic inside a lecture video. To identify the topics, a video is split into smaller segments where the scene changes. Speech text-based indexing is based on the text similarity in the video segments. An index point represents the start of a new topic. The previous work done is based on the text extracted from the images in a video using OCR technology. The text from the images was utilized to achieve indexing of a video lecture. This thesis proposes the development of a text-based indexing system from the audio content of a video lecture that contains significant information. The captions or text collected for this project were collected using an open-source speech recognition system (i.e., YouTube) [10].

Initially, a text-based indexing algorithm was chosen out of five different algorithms. The performance of image- and speech-based indexing was evaluated for each of the five algorithms. Captions were corrected manually to see whether indexing results were improved. Error analysis of the speech-based indexing provides insight into the major causes of problems with speech and OCR text-based indexing and the limitations of the text-based approach. The differences between corrected and uncorrected speech, as well as OCR and speech-based text indexing, were analyzed, and future work in this direction is proposed.

## 1.4 Thesis Outline

This thesis is organized as follows: Chapter 1 explains the ICS (indexed, captioned, and searchable) videos project. Chapter 2 describes the work done in this field or the fields closely related to captioning and speech text-based indexing. Chapter 3 describes the hypothesis proposed for this thesis and various speeches and text-based indexing algorithms. Chapter 4 gives a detailed explanation of the methodology adopted for the experimentation purpose of different text types in lecture videos. The evaluation of speech-based and OCR text-based indexing, hybrid text, and manually corrected text-based indexing are explained in Chapter 5. Chapter 6 discusses the reasons for errors in indexing for various text types and the percentages of their occurrences. Finally, Chapter 6 summarizes the overall thesis with the derived conclusion and possibilities for future works.

# Chapter 2. Related Work

The capture and distribution of lectures online are extremely useful for students with diverse backgrounds mainly from underdeveloped or developing countries [11] and improve student involvement. However, students have difficulty finding specific topics or information from a long lecture video, and there is a need to automatically index the contents of the lecture videos [12]. The main characteristics of videos that make indexing and retrieval difficult are richer content, large amounts of raw data, and poor structure [13]. The increase in the number of videos has led researchers to attempt to automatically index the videos in order to create a digital library for easy information retrieval [14] [15].

There has been similar work done in the lecture video indexing field with OCR text and speech text as presented in this thesis. The purpose of video indexing is to be able to detect the main key frames that indicate content change in the videos [16]. The work by Caüosnon [3] provided a model that extracts features from video images, allowing the researchers to be able to label the video frames and achieve video indexing with an accuracy of 95%. The work by Yang [4] discussed the development of an automatic method for the extraction of segments and keywords from both OCR and ASR (Automatic Speech Recognition) methods. They proposed a new method for ranking the keywords extracted. Various methods have been developed that use both OCR and ASR data for content-based video retrieval. Some of the work is in the area of semantic multimedia retrieval by applying the techniques of OCR, ASR, etc. for metadata generation [17]. They also came up with an entity recognition algorithm that is used to extract entities from the textual metadata. In [18], the authors proposed a video visual analysis framework that consists of video segmented of a slide, OCR engines, and an automatic lecture overview extraction technique. In [19], the authors introduced a solution to improve the ASR results of German lecture videos. The work above focuses more on OCR and ASR keyword extraction without any filtering methods or focus on the difference between both types of text, spoken text filtering and the difference between slide text and spoken text, and the indexing precision achieved by them individually.

This thesis mainly focuses on video indexing using speech text or captions. The work done by Zang [20] proposed a natural language approach to video indexing for content-based retrieval in order to identify the user topic searched for by entity extraction, frame based indexing, and other techniques for information retrieval. In [21] recorded lectures were transcribed and the speech recognition software generated a time stamp for each word and divided these into clusters. This was done so that the search engine could find the exact location of a topic of interest in order for the user to find an explanation, example, or a repetition of a particular word or topic inside a lecture. In [22], the authors investigated the use of online text resources to improve speech recognition performance for identifying keywords and applied various keyword-filtering methods. We are using a fixed duration-based algorithm based on text similarity applied to slide text and spoken text or speech text. The work done by Cooper [23] presented a video retrieval mechanism using slide and spoken text and provided a method to extract only common key terms from slide text and spoken text and performed analysis using the ground truths for both. They also have combined both the modalities and experiment indexing results. They used Talk Miner for indexing purposes. In [24], the authors have addressed "language model adaptation" for automatic lecture transcription by utilizing the slide information by adding local preferences of the keywords using a "cache model" by referring the slides used during each speech of utterance. They achieved significant accuracy on the detection rate of content keywords. The authors of [25] presented a method to correct the transcripts of lecture videos automatically using text from the slides, and they constructed a "sequential Hidden Markov Model for the observed phonemes that follows slide word order" placed with the words or texts not present on the slide. They showed that there was improvement in the accuracy of the transcripts and the alignment with the words in the slide. We have used manual correction of captions in this thesis. In [26] the authors followed two audio indexing approaches. The first one is based on bilingual automatic speech recognition, and the other one is used after speech "diarization" for the purpose of selecting the corresponding monolingual speech recognizer in order to decode the speech. In addition, they combined both the approaches and evaluated the audio indexing system from an information or topic retrieval point of view. We have also utilized the text from OCR and speech in order to perform hybrid text-based video indexing.

Most related work focuses on either speech or text forms for document retrieval, which refines words before processing. This thesis differs from the previous research, as the video indexing is done using speech-based texts that contain all the words and characters.

We have used YouTube as a speech recognition tool to generate captions and did manual correction of captions utilizing the ICS caption editor. The work has been done on both corrected and uncorrected speech, and the error analysis of OCR text versus uncorrected speech and uncorrected speech text versus corrected speech text is being done. The hybrid text indexing algorithm is a new concept in itself. Despite years of steady progress on performance, perfect or nearly perfect indexing accuracy is still a challenge, and we have proposed a variety of text-based indexing in order to achieve the best accuracy in different video scenarios.

# Chapter 3. Text- and Speech-based Indexing

In this section, we will discuss the hypothesis of the thesis and the indexing algorithm used. The idea for this thesis comes from utilizing the audio and slide text of video lectures. We will discuss the advantages of using slide text, as well as spoken text. We believed and proposed that a video lecture is not complete without speech as well as images. Both contain important information about a topic and can be utilized to achieve index points for ease of access to a topic for the students. We also proposed the correction of raw speech text in order to achieve better indexing accuracy. In another section we will discuss the underlying text-based indexing algorithm for video indexing. The primary goal of the text-based indexing algorithm is to identify the index points out of a group of consecutive transition points within a lecture video. The input to these indexing algorithms is a video consisting of several transition frames.

## 3.1 Advantages of Using Different Text Types for Video Indexing

We proposed that utilizing all the different text types that are part of video lectures and that give topic information could give us better indexing results. These text types include slide text and speech text, which is the instructor audio. In the subsections next we will discuss the proposed advantages of various text types used for video indexing of classroom lecture videos.

### 3.1.1 Slide/ OCR Text

Slides from view graphs (e.g., PowerPoint slides) represent information related to a topic in video lectures for different courses. Slides also contain precise and important information for a keyword search. Slides contain information in the form of text and images, etc. We have text extracted from the slides of video lectures with OCR technology [7]. It also determined the transition points, a point where the scene in a lecture video changes, also known as slide transition. Transition points contain text from the slides with the start and end duration in a lecture video. The text from slides represents main topic information and provides content-based retrieval, however OCR's recognition is erroneous and the indexing accuracy achieved by OCR text-based indexing is not perfect [29]. The lecture style differs in various courses and depends on the instructor's method of organization of content. In this thesis, we proposed a solution to this problem by utilizing the audio content of videos as discussed in the next section.

### 3.1.2 Audio/ Speech Text

Spoken text is one of the main information resources in a lecture video. The instructor speaks or discusses a topic in a lecture with a characteristic vocabulary. It is spontaneous and abundant. Although the speech is not improvised or completely related to a topic every time, it still proves to be one of the important factors in content-based retrieval of a topic in a long, continuous video lecture. We proposed video indexing using the speech content of a lecture video. Using YouTube [10] as speech recognition tool in our experiment, we gained captions/speech transcripts of lecture videos in order to utilize them for indexing purposes. Speech content may differ completely and is more dependent on the topic taught in a class and the instructor’s technique of explanation. For example, one instructor may speak accurately due to more preparation and give a better and more topic-related speech, and the other may speak with more random content and many grammatical errors, for example. However, we strongly believe that speech text contains important topic information and can be used to achieve topic segmentation and index point generation.

### 3.1.3 Hybrid Text

In order to utilize the strengths and topic-related keywords from both speech and slide text, we proposed a hybrid text type for video indexing purposes. Here, we proposed combining the slide text and speech text for video indexing as it contains more topic information. Utilizing the text from slides enables concise and accurate topic information while text from the speech gives in depth and extra information mentioned in the class. This may prove better when similar keywords related to a topic are being searched for indexing in a way that will result in more words for each transition frame.

### 3.1.4 Corrected Speech Text

An analysis demonstrated that the speech recognition tool yields errors while captioning lecture videos. YouTube generates an initial transcription, which is sometimes not suitable for indexing purposes. There are various reasons for errors, such as a heavy accent, out of vocabulary words, uncaptioned speech, etc., [9]. We proposed manually correcting the raw text generated using YouTube and using the corrected text for video indexing purposes. The correction includes non-vocabulary word correction, spelling mistakes, correct recognition of spoken text, etc. This will assist with better text similarity across video segments and will help achieve better indexing accuracy than uncorrected raw speech text.

In the next sections, we will discuss the definitions of transitions, index points, and the underlying algorithm used in order to achieve video indexing for slide and spoken text.

## 3.2 Definitions and Similarity Metric

This section provides the definitions of a transition point and index point. These two terms are very important in understanding the indexing algorithm and are used extensively in this thesis.

### 3.2.1 Transition Point

A transition point is a point in a video where a significant image change occurs. Image changes occur when there is a scene change within a lecture video. This is determined on the basis of image differences between consecutive frames in a video [5]. The scene change occurs when an instructor moves from one viewgraph (e.g., PowerPoint slide) to another viewgraph, etc. In this case, the previous frame is significantly different from the current frame or the current transition frame. A transition point usually presents a starting point of a transition segment. A segment is a section between the starting and end point of a transition point. The transition point has its start and end time at the end of a segment. In [29], it is proposed that the text content in a segment can be represented as a vector **s = (tf$_1$, tf$_2$,….,tf$_n$)** where **tf$_n$** is the term frequency. Figure 3.1 provides the transition point and transition segment view. Index points will be discussed in the next section.

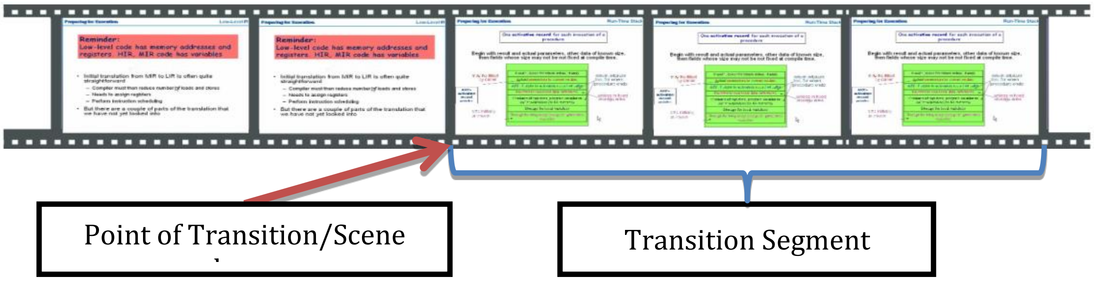
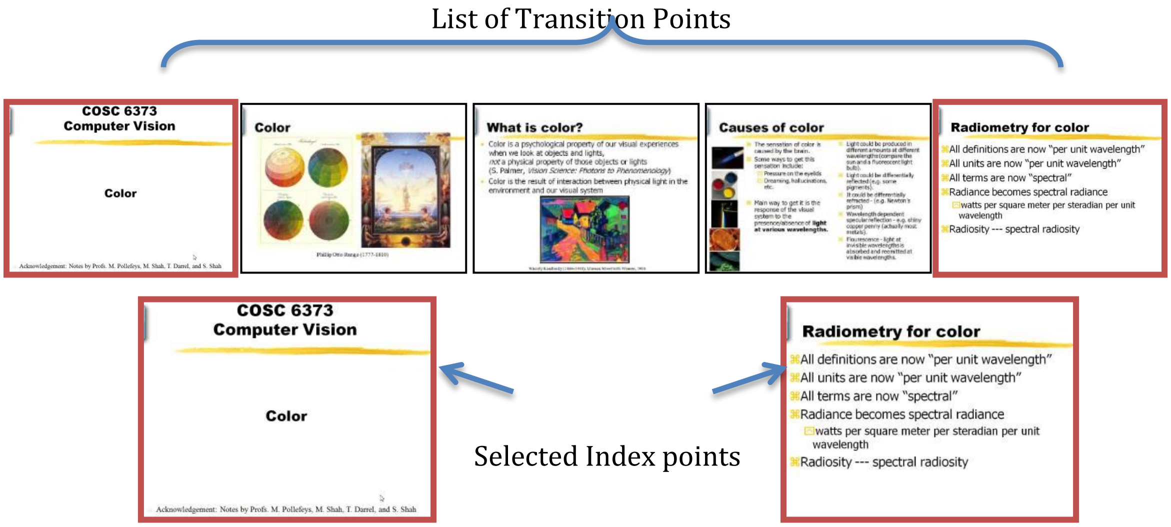


**Figure 3.1: Transition point and transition segment [29]**

### 3.2.2 Index Point

An index point is a subset of transition points in a video lecture, and it represents the start of a new topic of discussion in a video. Transition segments that belong to the same topic are grouped together based on text similarity, which is discussed later, and are represented by the index point. Figure 3.2 provides the selected index points from a list of transition points.



COSC 6373
Computer Vision

Color

Acknowledgement: Notes by Profs. M. Pollefeys, M. Shah, T. Darrel, and S. Shah

What is color?

- Color is a psychological property of our visual experiences when we look at objects and lights, *not* a physical property of those objects or lights (S. Palmer, *Vision Science: Photons to Phenomenology*)
- Color is the result of interaction between physical light in the environment and our visual system

Causes of color

- The sensation of color is caused by the brain.
- Some ways to get this sensation include:
  - Pressure on the eyelids
  - Dreaming, hallucinations, etc.
- Main way to get it is the response of the visual system to the presence/absence of **light at various wavelengths.**
- Light could be produced in different amounts at different wavelengths (compare the sun and a fluorescent light bulb).
- Light could be differentially reflected (e.g. some pigments).
- It could be differentially refracted - (e.g. Newton's prism)
- Wavelength dependent specular reflection - e.g. shiny copper penny (actually most metals).
- Flourescence - light at invisible wavelengths is absorbed and reemitted at visible wavelengths.

Radiometry for color

- All definitions are now "per unit wavelength"
- All units are now "per unit wavelength"
- All terms are now "spectral"
- Radiance becomes spectral radiance
  - watts per square meter per steradian per unit wavelength
- Radiosity --- spectral radiosity

**Figure 3.2: Index points example [29]**

### 3.2.3 Text Similarity Metric: Cosine Similarity

The general text similarity metric used in various algorithms as discussed in the latter part of this thesis is the cosine similarity metric. The value ranges from zero to one and is normalized.

Conditions:

- If two vectors are the same, the angle between them is zero; the cosine similarity value is one.
- If the two vectors are completely different, the cosine similarity value would be zero.

The frequency of term occurrence is used in the cosine similarity calculation and each text block is represented as a vector of the term frequency count. Cosine similarity is given by the following formula:

$$Cosine\ Similarity = \frac{P \cdot Q}{\|P\|\ \|Q\|},$$

where $P$ and $Q$ are the two term frequency vectors. The numerator $P \cdot Q$ is the dot product of the two given vectors. The denominator $\|P\|\ \|Q\|$ is the product of the modulus of the two given vectors [29].

The main factor in cosine similarity is the number of similar terms along with their frequencies. If there are two text segments having similar terms and the frequency is higher, then they are given a high cosine similarity coefficient. Hence, cosine similarity is really useful in topic segmentation in a video lecture.

## 3.3 Underlying Text-based Algorithm

The main purpose of indexing algorithms is to index the entire lecture video so that each index point represents a topic. Before the indexing phase, the whole lecture video is divided into various transition segments. The basic concept of indexing is that it compares the text similarity of a segment with its right and left segment depending on which segment has greater similarity. In the sections below, we will discuss the indexing algorithms taken from [29]. The following five algorithms are being used in the research section of the ICS videos projects: uniform, fixed grouping duration, linear weighted, nonlinear weighted, and boundary-based algorithms. The main algorithm that we have used in our evaluation of audio and speech text is the fixed grouping duration algorithm.

### 3.3.1 Fixed Grouping Indexing Algorithm

We have used the fixed grouping indexing algorithm in the entire procedure of this thesis. The decision to merge is based on the text similarity between the segments. This algorithm compares the text of the smallest segment with a group of segments on its right as well as on the left side. The number of segments to be grouped is determined by an empirically selected grouping duration so that the combined duration of the group should not exceed the grouping duration. The segments that are grouped are considered a single segment. The text in the individual segments is added to form the group. For the similarity comparison against a group, the text of a given segment is compared with the combined text of the group. The explanation of the algorithm is as follows.

For the input, a video with a list of transition points is taken along with the required number of index points and grouping duration in seconds. The grouping duration is empirically selected in seconds. Initially, a transition segment (point **K)** is selected with the least duration. The text similarity of the current segment is compared with the text similarity of the left or right group of the segment. If the similarity toward the left is greater, it is merged with the immediate transition segment on the left. Otherwise, it will be merged with the right segment. This is done until the required number of index points equals the remaining transition points. Figure 3.3 summarizes the fixed grouping algorithm.

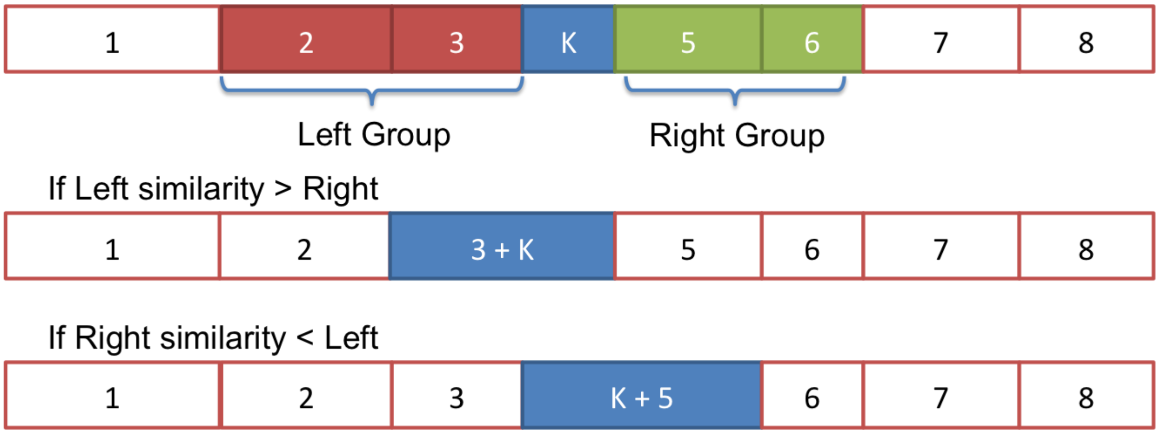


**Figure 3.3: Fixed grouping algorithm example.**

# Chapter 4. Methodology

After the discussion of the proposed methods and different texts for video indexing, in this chapter we will discuss the implementation methodology adopted for the purpose of evaluation. We had OCR text extracted from the slides of lecture videos along with the time frame for each transition point. Each frame contains text. We proposed and implemented spoken, hybrid, and corrected text in each corresponding transition frame of a video lecture that contains slide text. The overall methodology will be discussed in the sections below.

## 4.1 Files for Evaluation

For the purposes of our experiment, we had the text from the images that were extracted using OCR [7] in the Extensible Markup Language (XML) file format. These XML files contain all the transition points, segment durations, and extracted text. The files containing speech text generated while attaining captions from YouTube were available in SubRip Text (srt) file format and needed to be in the XML file format with specific transition points and time durations. These files, along with the XML file containing speech text, were used for analysis and achieving index points when run against text-based indexing algorithm.

### 4.1.1 Placing Speech Text to Corresponding Transition Points

The next step was to convert the caption files generated from YouTube into the srt file format to XML file format. This was done via a simple JAVA program and the text in

XML file was replaced with the speech text present in the srt file corresponding to the respective transition point duration.

An example of the srt and XML file replacement of text is given below:

- Example .srt file**:** The start and the end time format is as: hh:mm:ss:msec.

```
1

00:00:01,520 --> 00:00:10,960
```

*with these different representations numbers you standard written addition track pretty much thing how wind up doing example here's a binary example like warm 10 reading based and aerial in Europe one one euro carry one one here binary addition pretty much the same thing Harry attraction at the same back but we're not going to really talk about that name mainly were you get into signed me express negative numbers*

- Example XML file: The start time and duration is taken into consideration for replacement. In this case the transition point contains text starting at 00.01 seconds until 00.10 seconds

```
<tp>
  <tpNo>1</tpNo>
  <isIndex>1</isIndex>
  <inNo>1</inNo>
  <time>00:01</time>
  <duration>0:10</duration>
  <title> Index 1</title>
```

<img>i_0002.jpg</img>

<txt> *what to do with representations of numbers just what we do with numbers add them subtract them multiply them divide them coitipbye them nu nn nn exampie 10 17 so simple to add in binary that we can build circuits to do lt subtraction just as you would in decimal comparison how do you tell if* </txt>

</tp>

- Replacement of XML file text with speech text in srt file: In this phase, the text or words in the <txt></txt> tags are replaced and filled with speech text present in the srt file according to the time duration of the transition point. This is the original speech text (uncorrected speech text).

<tp>

<tpNo>1</tpNo>

<isIndex>1</isIndex>

<inNo>1</inNo>

<time>00:01</time>

<duration>0:10</duration>

<title> Index 1</title>

<img>i_0002.jpg</img>

<txt> *with these different representations numbers you standard written addition track pretty much thing how wind up doing example here's a binary example like warm 10 reading based and aerial in Europe one one euro carry one one here binary addition*

*pretty much the same thing Harry attraction at the same back but we're not going to really talk about that name mainly were you get into signed me express negative numbers*

</txt>

 </tp>

### 4.1.2 Hybrid Text: Union of OCR Text and Speech Text

The example below represents the hybrid text combination method used for the evaluation taken from Section 4.1  In this case, OCR text is placed initially followed by speech text inside a <text> </text> tag for a particular transition frame.

For example:

<tp>

  <tpNo>1</tpNo>

  <isIndex>1</isIndex>

  <inNo>1</inNo>

  <time>00:01</time>

  <duration>0:3</duration>

  <title> Index 1</title>

  <img>i_0002.jpg</img>

  <txt> *what to do with representations of numbers just what we do with numbers add them subtract them multiply them divide them coitipbye them nu nn nn exampie 10 17 so simple to add in binary that we can build circuits to do lt subtraction just as you would in decimal comparison how do you tell if with these representations of numbers, you can*

```
use numbers standard arithmetic, which is addition, subtraction, multiplication, division, pretty much the same thing wind up doing it, for example there's a binary fission example just like when we get to ten and we're adding base ten, you just carry over and plug into so you have zero plus 1 one and one plus one is zero carry the one, one plus one zero carry the one, so for binary addition pretty much the same thing except the carry here, subtraction has the same exact effect there but we're not going to really talk about that today mainly we're getting into assigned  representation, so how can we express negative numbers </txt>
 </tp>
 <tp>
```

### 4.1.2 Placing Corrected Speech Text

As discussed previously, we have corrected captions available for few of the lectures. The corrected captions are also used in order to determine indexing accuracy and find out if it gives better accuracy than the OCR-based text indexing. These corrected captions were also in the srt file format, and we needed to make them available in the XML file format similarly to the previous section.

- Replacements of XML file text with corrected speech text in srt file.

```
<tp>
  <tpNo>1</tpNo>
  <isIndex>1</isIndex>
  <inNo>1</inNo>
  <time>00:01</time>
  <duration>0:3</duration>
  <title> Index 1</title>
  <img>i_0002.jpg</img>
  <txt> with these representations of numbers, you can use numbers standard arithmetic, which is addition, subtraction, multiplication, division, pretty much the same thing wind up doing it, for example there's a binary fission example just like when we get to ten and we're adding base ten, you just carry over and plug into so you have zero plus 1 one and one plus one is zero carry the one, one plus one zero carry the one, so for binary addition pretty much the same thing except the carry here, subtraction has the same exact effect there but we're not going to really talk about that today mainly we're getting into assigned  representation, so how can we express negative numbers </txt>
 </tp>
 <tp>
```

The different collection of texts in the same input format is required in order to evaluate their performance against the underlying indexing algorithm. Collection of ground truth is the initial step, which is discussed in the next section, in evaluating these files containing different kinds of texts.

## 4.2 List of Lectures

The uncorrected speech text consists of sixteen video lectures from the University of Houston (UH) and nine from Coursera [27]. Later on, some of the lectures speech text were corrected and evaluated again. The evaluation helped to understand the advantages, disadvantages, strengths, and weakness of the text type used for indexing the video lectures. This prompted further enhancements and also provided insight for future developments. The videos were selected from the Computer Science, Biology and Biochemistry, and Earth and Atmospheric Sciences departments as shown in Table 4.1. The main reason for selecting these video lectures was due to their presentation format (PowerPoint slides that contain text), ability of caption generation via YouTube, and the consent of the instructor to provide ground truths.

| **Source** | **Major** | **Course Name** | **Lecture Count** |
|---|---|---|---|
| UH | Computer Science | Introduction to Computing | 4 |
| UH | Computer Science | Computer Organization and Programming | 5 |
| UH | Computer Science | Digital Image Processing | 2 |
| UH | Computer Science | Computer Architecture | 2 |
| UH | Biology and Biochemistry | Human Physiology | 3 |
| Coursera | Computer Science | Compilers | 3 |
| Coursera | Computer Science | Cryptography | 2 |
| Coursera | Computer Science | Machine Learning | 2 |
| Coursera | Computer Science | Probabilistic Graphical Models | 2 |
| Total | | | 25 |

**Table 4.1: List of source of courses used for evaluation.**

## 4.3 Ground Truth Collection

As discussed previously, a lecture video consists of a number of transition points representing a scene. An index point in a lecture video is a subset of the transition points that generally occurs when a scene in a video changes. These index points are provided by the instructor and entirely depend on the instructor approach and perspective of selecting the ground truth. Ground truth heavily impacts the indexing score calculation. Therefore, it is a very critical step toward the video indexing concept as a whole.

### 4.3.1 Ground Truth of Lectures for Evaluation

During evaluation, the output of the indexing algorithms was evaluated against the ground truth of lectures provided by the respective instructors. This is true for the University of Houston's lectures where each lecture is a continuous single classroom lecture. The Coursera lectures consist of sub-topics and separate recordings assembled together to form a single lecture. That is why these were observed and analyzed separately. The transition points representing these sub recordings or individual segments were marked as an index point. The ground truth for video IDs 569-593 were taken from previous work in [29] and the ground truth for video ids 180-341 were collected separately from the instructors teaching that course particularly for this thesis. Table 4.2 and Table 4.3 represent the ground truth for lecture videos from University of Houston and Coursera lecture videos, respectively.

| Video ID | True Index Points | Total Index Points Count | Total Transition Points in a Lecture | Total Video Duration in Minutes |
|---|---|---|---|---|
| 180 | 1 3 17 37 38 39 43 49 57 | 9 | 87 | 65 |
| 184 | 19 21 26 54 66 72 74 75 78 83 92 114 119 121 123 127 128 134 137 138 141 142 143 | 23 | 152 | 40 |
| 186 | 32 34 72 74 75 78 79 80 81 83 84 87 89 92 93 94 95 96 97 98 99 100 101 102 | 24 | 107 | 64 |
| 260 | 1 14 18 41 47 59 66 69 70 72 73 76 77 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 119 120 127 128 129 136 139 141 143 144 | 40 | 160 | 48 |
| 336 | 1 3 6 13 | 4 | 20 | 18 |
| 337 | 1 4 6 9 12 17 19 26 28 32 35 38 43 44 45 | 15 | 48 | 27 |
| 338 | 1 5 6 11 14 15 18 20 21 22 25 33 36 55 58 | 15 | 59 | 30 |
| 339 | 1 6 7 9 12 13 14 15 17 24 30 32 39 40 50 52 | 16 | 54 | 19 |
| 341 | 1 7 9 13 19 33 39 49 51 77 83 | 11 | 91 | 24 |
| 569 | 1 2 3 4 5 6 8 9 10 11 12 13 14 23 27 30 31 32 33 34 35 36 37 38 39 40 | 26 | 41 | 80 |
| 570 | 1 2 3 4 5 6 7 8 9 10 11 12 13 16 19 20 21 23 25 26 27 28 29 30 31 32 33 37 38 39 40 41 | 32 | 41 | 82 |
| 571 | 1 2 3 4 6 7 8 9 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 37 38 42 44 45 | 33 | 45 | 81 |

| Video ID | True Index Points | Total Index Points Count | Total Transition Points in a Lecture | Total Video Duration in Minutes |
|---|---|---|---|---|
| 572 | 1 9 12 15 18 22 25 28 29 | 9 | 31 | 48 |
| 573 | 1 6 12 43 61 | 5 | 83 | 54 |
| 574 | 1 3 5 8 20 23 24 26 35 36 39 40 42 43 58 62 | 16 | 64 | 77 |
| 575 | 1 2 4 18 19 25 27 44 52 55 62 64 65 66 67 69 70 73 74 77 94 95 | 22 | 100 | 77 |
| 576 | 1 6 7 10 11 13 14 22 | 8 | 32 | 85 |
| 577 | 1 2 7 22 33 42 53 58 64 65 66 67 68 | 13 | 71 | 83 |
| 578 | 1 12 13 16 18 23 28 29 31 | 9 | 40 | 72 |
| 579 | 1 2 9 16 20 23 | 6 | 28 | 76 |
| 580 | 1 7 9 10 11 14 | 6 | 20 | 72 |
| 583 | 1 3 5 22 28 32 53 68 78 | 9 | 108 | 82 |

**Table 4.2: Ground Truth For UH Lecture Videos**

| Video ID | True Index Points | Total Index Points | Total Transition Points | Total Video Duration in Minutes |
|---|---|---|---|---|
| 584 | 1 20 25 30 40 | 5 | 46 | 46 |
| 585 | 1 19 26 34 40 49 | 6 | 52 | 80 |
| 586 | 1 5 17 25 35 48 61 70 79 | 9 | 81 | 81 |
| 587 | 1 17 44 63 79 98 | 6 | 105 | 99 |
| 588 | 1 15 24 33 42 52 | 6 | 55 | 60 |
| 589 | 1 17 27 40 54 67 83 | 7 | 85 | 92 |
| 590 | 1 19 93 105 | 4 | 114 | 36 |
| 591 | 1 22 40 60 86 108 | 6 | 118 | 81 |
| 592 | 1 8 21 36 45 54 63 | 7 | 66 | 63 |
| 593 | 1 10 17 25 42 58 65 76 | 8 | 79 | 75 |

**Table 4.3: Ground Truth for Coursera Lecture Videos**

The most challenging issue regarding evaluating these videos is that individual videos could have different index points. Therefore, a custom evaluation metric [29] was used for the calculation as discussed in the later sections.

### 4.3.2 Metric for Ground Truth Calculation

A scoring metric that was devised for the ease of evaluation and relative comparison of the indexing output was used for the purpose of this thesis. An indexing accuracy score is calculated for each video lecture that is based on the output of the indexing algorithm. We used the indexing accuracy criteria in order to evaluate the performance of different algorithms and text types used for indexing. The whole scoring process takes place in two phases: calculation of the theoretical score for the lecture videos and the calculation of the indexing score. As discussed in the previous section, the ratings for the University of Houston videos are provided by the instructors and the ratings of the Coursera videos is based on the different topic segmentations provided in the Coursera website. The ground truth data provided by the instructor is used to calculate the theoretical score, which is the maximum possible attainable score. On the other hand, the indexing score is calculated based on the indexing algorithm output. The following formula is used to calculate indexing accuracy:

$$Accuracy = \frac{Total\ Indexing\ Score}{Total\ Theoretical\ Score}.$$

In the next sections, we will discuss how the ground truth ratings for transition points are provided and the calculation of the theoretical score and indexing score.

### 4.3.3 Ground Truth Rating for Transition Points

Each lecture video consists of various transition points, and these transition points are given a rating in the range of zero to three. The rating is an indication of whether a video has good index points and is based on the following criteria:

- *Rating of three for definitely an index point:* If a transition point is definitely an index point that is a start of a new topic, then it is given a rating of three.
- *Rating of two for probably an index point:* There are some transition points where the instructor is not sure if it is an index point, for example, an outline slide or a

sub-topic that is a part of the main topic, etc. A rating of two is given for such transition points where the probability of index point is higher but not definite.

- *Rating of one for probably not an index point:* Similarly, a rating of one is given when a transition point is probably not an index point.
- *Rating of zero for definitely not an index point:* If the transition point is definitely not an index point, a rating of zero is given.

### 4.3.4 Theoretical Score Calculation

The following condition and score is used to calculate the theoretical score that is the sum of total scores of the transition points for each video lecture.

- For a rating of three or zero, a theoretical score of +ve 2(+2) is given to that transition point.
- For a rating of two or one, a theoretical score of +ve 1(+1) is given to that transition point.

The reason behind this scoring scheme is that the definite and probable index points will always be marked as index points by the algorithm.

The total theoretical score is calculated using the following formula:

$$Total\ Theoretical\ Score = \sum_{1}^{n} Transition\ Point\ Theoretical\ Score,$$

where *n* is the total number of transition frames in the video.

### 4.3.5 Indexing Score Calculation

The indexing score is based on the following two factors: the output of the indexing algorithm and the ground truth. The indexing score is calculated as described in Table 4.4.

| | | Ground Truth Rating | | | |
|---|---|---|---|---|---|
| | | **0**<br>**Definitely Not Index** | **1**<br>**Probably Not Index** | **2**<br>**Probably an Index** | **3**<br>**Definitely an Index** |
| **Indexer Output** | **Not Index (0)** | +2 | +1 | -1 | -2 |
| | **Is Index (1)** | -2 | -1 | +1 | +2 |

**Table 4.4: Indexing Scores for Transition Points [29]**

The detailed description of Table 4.4 is given below:

- If a transition point is recognized as an index point by the algorithm and is marked as definitely an index point in the ground truth (i.e., a rating of three is given), an indexing score of two is assigned to it. Also, if the indexing algorithm marks the transition point as, not an index point, an indexing score of –ve 2(-2) is assigned to that transition point.
- If a transition point is recognized as an index point by the indexing algorithm and is marked as probably an index point in the ground truth (i.e., a rating of two is given), an indexing score of one is assigned to it. Also, if the indexing algorithm

marks the transition point as not an index point, an indexing score of –ve 1(-1) is assigned to that transition point.

- If a transition point is found to not be an index point by the indexing algorithm and is marked as probably not an index point in the ground truth (i.e., a rating of one is given), an indexing score of one is assigned to it. Also, if the indexing algorithm marks the transition point as an index point, an indexing score of –ve 1(-1) is assigned to that transition point.
- If a transition point is found to not be an index point by the indexing algorithm and is marked as definitely not an index point in the ground truth (i.e., a rating of zero is given), an indexing score of two is assigned to it. Also, if the indexing algorithm marks the transition point as an index point, an indexing score of –ve 1(-1) is assigned to that transition point.

These indexing scores are used to determine the accuracy for analysis using different algorithms that we will discuss in the next section.

## 4.4 Captions Generation

In order to perform speech indexing of video lectures, it was first necessary to generate captions of all the video lectures. It was also necessary to determine the best speech recognition technology that yields accurate captions for these video lectures. An automated speech recognition tool is used to extract the spoken text and information from the video lecture. The accuracy of captions greatly affects the overall indexing of video lectures. This will be discussed in the later sections. Video lectures from various departments of the University of Houston available at the ICS videos website [8] and Coursera were used [27].

### 4.4.1 List of Lectures

A total of 25 video lectures were taken from different departments. All the lecture videos were of different time durations. Table 4.5 is a list of lectures teaching different topics with different time durations and instructor’s name.

| Source | Video ID | Topic | Duration in (min) | Instructor Name |
|---|---|---|---|---|
| UH | 180 | Introduction to computing: Singularity | 65 | Dr. Johnson UH |
| UH | 184 | Hardware and software | 40 | |
| UH | 186 | Internet Technologies | 64 | |
| UH | 260 | How to make a presentation | 48 | |
| UH | 336 | Representation of numbers | 18 | Dr. Rizk UH |
| UH | 337 | Floating point numbers | 27 | |
| UH | 338 | Boolean algebra | 30 | |
| UH | 339 | Logic gates and circuits | 19 | |
| UH | 341 | Combinational circuits | 24 | |
| UH | 569 | Cell to cell communication | 80 | Dr. Wayne UH |
| UH | 570 | Hormone classification | 82 | |
| UH | 571 | Changes in membrane | 81 | |
| UH | 572 | Instruction set architecture | 48 | Dr. Gabriel UH |
| UH | 573 | Hardware-based speculation | 54 | |
| UH | 578 | Binary image processing | 72 | Dr. Shah UH |
| UH | 580 | Linear Systems & Linear Image Filtering | 72 | |
| Coursera | 584 | Error Handling | 46 | Coursera |
| Coursera | 585 | Predictive Parsing | 80 | Coursera |
| Coursera | 586 | Semantic Analysis | 81 | Coursera |
| Coursera | 588 | Collision Resistance | 99 | Coursera |
| Coursera | 589 | Authentic Encryption | 60 | Coursera |
| Coursera | 590 | Probabilistic Graphical Model | 36 | Coursera |
| Coursera | 591 | Conditional probability queries | 81 | Coursera |
| Coursera | 592 | Advice to use ML | 63 | Coursera |
| Coursera | 593 | Linear regression with one variable | 75 | Coursera |

**Table 4.5: List of videos with duration, topic and instructor's name.**

### 4.4.2 YouTube Captioning

Since YouTube gave the most accurate results [9] over Dragon Naturally Speaking Preferred 10 (DNS) and Windows Speech Recognition (WSR), we decided to utilize it as an automatic speech recognition tool to generate captions. All the lecture videos were uploaded to YouTube by signing in with a user ID and password. Initially, twenty-eight video lectures were uploaded, but some did not obtain captions at all. The reasons could be the audio recording tool at the time of classroom lecture, the accent of instructor, etc. YouTube gave captions in the form of a srt file within a few days of uploading for most of the lectures. Some of the lecture videos that were not captioned at all were the longer lecture videos. Therefore, we decided to divide the longer videos into smaller videos using an open source tool [28] and then upload these to YouTube. Finally, all the captions of the videos were achieved with the exception of three of the videos for the reasons discussed above. Although there were some errors in the captioning done by the YouTube Speech Recognition System, these errors were corrected later manually for the evaluation portion of this thesis as discussed in later sections.

### 4.4.3 Text Accuracy Calculation

This section explains the analysis of errors that occurred during the captioning process, the accuracy of the caption text achieved, and the reasons for it. For the analysis, all twenty-five lecture videos were reviewed and their total number of captioned text words was calculated.

The text accuracy calculation for each lecture was calculated using the following formula [9]:

Accuracy % = (100 – (Number of errors / Total number of words in ground truth)) * 100.

The total number of words in the ground truth is the total number of words in the original caption text for each video lecture. The average number of words in the set of lecture videos was calculated to be 7,726 and the average number of characters was calculated to be 44,525. The number of errors was calculated by listening to each video lecture for at least thirty minutes each. The text accuracy for each lecture along with the overall average accuracy is given in Table 4.6.

| Video ID | YouTube Text Accuracy | No. of words | No. of characters |
|---|---|---|---|
| 180 | 62.14% | 2,467 | 24203 |
| 184 | | 3,189 | 19828 |
| 186 | | 2,755 | 27033 |
| 260 | | 4,736 | 21634 |
| 336 | 80.09% | 3,605 | 8493 |
| 337 | | 1,713 | 10731 |
| 338 | | 2,194 | 13499 |
| 339 | | 1,581 | 9476 |
| 341 | | 2,469 | 14945 |
| 569 | 81.66% | 12,186 | 72524 |
| 570 | | 12,552 | 74097 |
| 571 | | 12,436 | 72661 |
| 572 | 79.12% | 6,615 | 39327 |
| 573 | | 7,172 | 42023 |
| 578 | 70.39% | 8,498 | 47710 |
| 580 | | 7,931 | 45277 |
| 584 | 82.18% | 7,951 | 42389 |
| 585 | 87.75% | 13,731 | 74643 |
| 586 | 89.58% | 14,027 | 77794 |
| 588 | 80.14% | 11,437 | 60714 |
| 589 | 84.1% | 17,023 | 96301 |
| 590 | 87.63% | 5,651 | 34903 |
| 591 | 84.48% | 10,897 | 64770 |
| 592 | 65.1% | 9,459 | 56238 |
| 593 | 69.83% | 10,873 | 61904 |

**Table 4.6: Text Accuracy for each video lecture.**

The overall average text accuracy achieved for the set of twenty-five video lectures by YouTube is 78%. As observed, some of the video lectures were poorly captioned and some of them obtained really good text accuracy. For some of the lectures (e.g., video IDs 180-260), the average of the total accuracy of all the lectures where the instructor was the same was calculated. This text accuracy completely depends on the number of factors as discussed in the next section. The indexing accuracy evaluation and experiment results will be discussed in the next chapter.

# Chapter 5. Evaluation

In this section, the OCR text, uncorrected speech text, and the corrected speech text were evaluated for their indexing accuracy. The text-based indexing algorithm was applied on twenty-five different video texts from OCR and uncorrected speech text.

## 5.1 Indexing Accuracy Calculation for Algorithms

For further analysis of OCR text, uncorrected, and corrected speech text-based indexing there was a need to select a single algorithm. The algorithm selected assisted in determining the cases where speech-based text indexing was better than OCR text-based indexing and also in determining the factors leading to those results. In order to select an algorithm out of all the algorithms mentioned in the previous chapter, the indexing accuracy of each was calculated on the basis of ground truths and the output of all the algorithms. For the selection process, the uncorrected speech-based text indexing results for all five algorithms were chosen. As discussed, the uncorrected speech text was used for selecting the algorithm on the basis of indexing accuracy for further analysis. Each of the selected twenty-five videos with known ground truths were provided as inputs and the respective indexing score was calculated. The indexing accuracy scores for these video lectures were averaged in order to compare the relative performance of the algorithms. The result was used for selection of an algorithm. Each lecture video has a different time duration and a different number of index points from the ground truth. The indexing algorithms generate the required number of index points. Based on these index points, the indexing accuracy was calculated for each video lecture and algorithm. These were averaged later for ease of comparison. The result of the test for all the algorithms is summarized in Figure 5.2. For this evaluation, the grouping duration of 480 seconds and a half-life of four minutes were chosen.

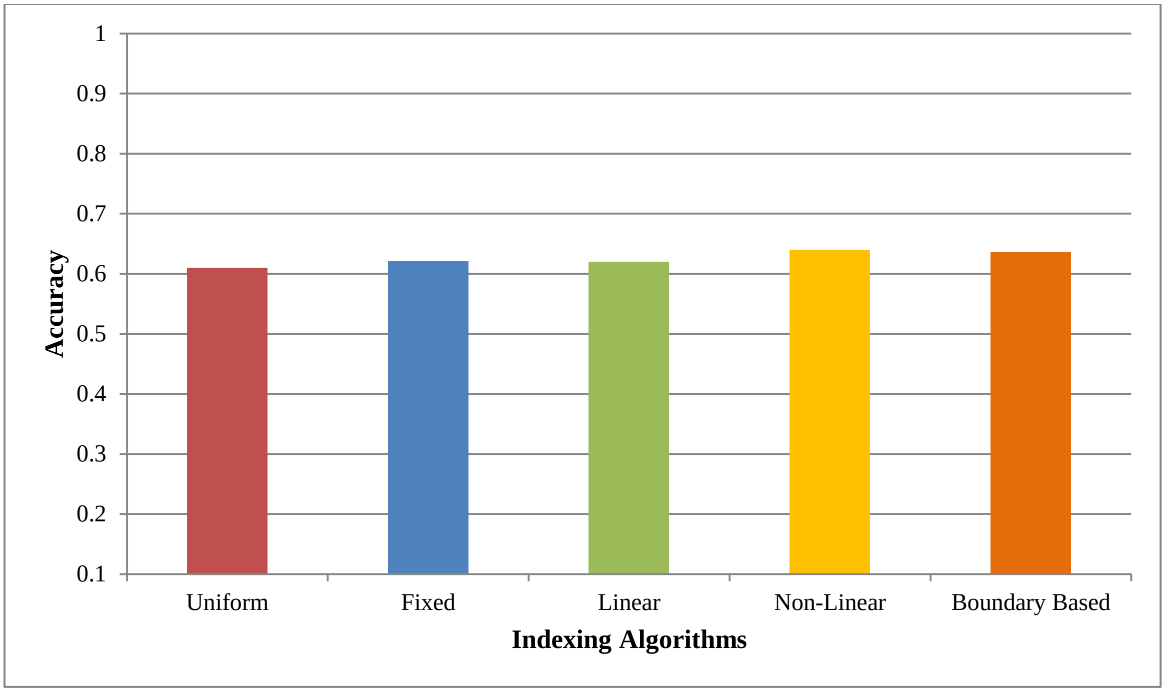


**Figure 5.1: Indexing accuracy for different algorithms on speech text.**

### 5.1.1 Analysis and Selection

As observed from Figure 5.1, there is a marginal difference between the performances of the algorithms for uncorrected speech text-based indexing. The average fixed grouping duration indexing accuracy for these lecture videos was 0.621. The fixed grouping duration algorithm was selected for further analysis due to its ease of use and understanding while analyzing each index point in the latter section of this thesis. For analyzing each index point given by the indexing algorithm against the ground truth, there was a need to select a single algorithm and move forward with further experiments, analysis of errors, and the evaluation of the best possible results.

## 5.2 Indexing Accuracy Evaluation for Different Text Types

In this section, we will discuss the average accuracy calculation for each of the twenty-five video lectures based on text extracted by OCR and the text provided by captions of the instructor audio. These captions are for uncorrected speech texts. There are four main reasons for conducting this test.

- To evaluate the performances of different text types,
- To determine the improvements achieved and a comparison,
- To investigate cases where one text type is better than the other text type, and
- To analyze the reasons that leads a certain text type to perform in a certain way.

### 5.2.1 Analysis of Indexing Accuracy for OCR Text and Uncorrected Speech Text

The analysis is based on the fixed grouping duration algorithm with a maximum grouping duration of 480 seconds. The indexing accuracy for OCR text and uncorrected speech text was calculated with the help of ground truth and the input XML files for all the lecture videos. This analysis is useful in determining which text type results in better indexing accuracy and under what circumstances. Additionally, the overall average indexing score is calculated in this section and the difference in terms of percentage is determined. The indexing accuracy calculation for OCR text and uncorrected speech text helps to determine the scenarios where one is better than another, and that could be useful for different types of lectures when deciding which text type should be used. The result of this test is summarized in Figure 5.2.

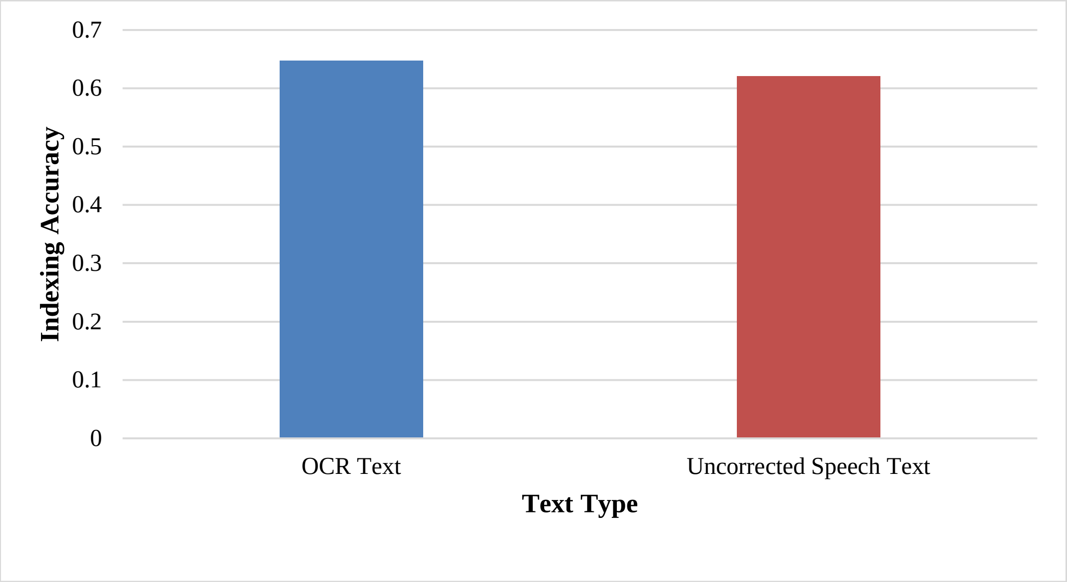


**Figure 5.2: Indexing accuracy for OCR and uncorrected speech text for all the video IDs.**

The indexing accuracy varies for different lectures, and there could be various possible reasons for it that, which will be discussed later. The overall average accuracy for both text types (i.e., OCR text and uncorrected speech text) was determined, and the difference in terms of percentage was calculated. This result is summarized in Table 5.1.

| OCR text | Uncorrected Speech text | Difference in Percentage |
|---|---|---|
| 0.647 | 0.621 | 4.18 |

**Table 5.1: Indexing accuracy percentage difference between OCR text and uncorrected speech text.**

The difference in percentage is calculated as follows:

$$\frac{OCR\ Text - Uncorrected\ Speech\ Text}{Uncorrected\ Speech\ Text} * 100.$$

The indexing accuracy difference in percentage between both text types seems to be marginal. However, the overall average analysis is not enough and not justifiable for deciding which text type obtains a better result and would continue to have better results in the future. There are some video lectures where the uncorrected speech text gives better results than OCR-based texts and vice versa. Hence, we took different video lectures for analysis of different cases. These cases will be discussed in the later sections.

### 5.2.2 Hybrid Text-based Analysis

Hybrid text means the union or the combination of text derived from OCR and text derived from the captions/uncorrected speech text. The idea to evaluate indexing accuracy with hybrid text comes from the fact that both text types contain important keywords that surely impact text similarity and thus can affect indexing accuracy. The text from OCR and the text from captions are both important in a text-based indexing calculation. For this purpose, nine video lectures (569, 570, 571, 572, 580, 590, 591, 592, and 593) were randomly chosen and their speech-based text and OCR text were combined. This was done for each transition frame and loaded into a file that was given as an input to the indexing algorithm in order to determine indexing accuracy. The indexing accuracies of hybrid text, OCR text, and uncorrected speech text were evaluated for this purpose. The result of the hybrid test is summarized in Figure 5.3. Please note that the average for hybrid experiment is being calculated for the selected nine video lectures.

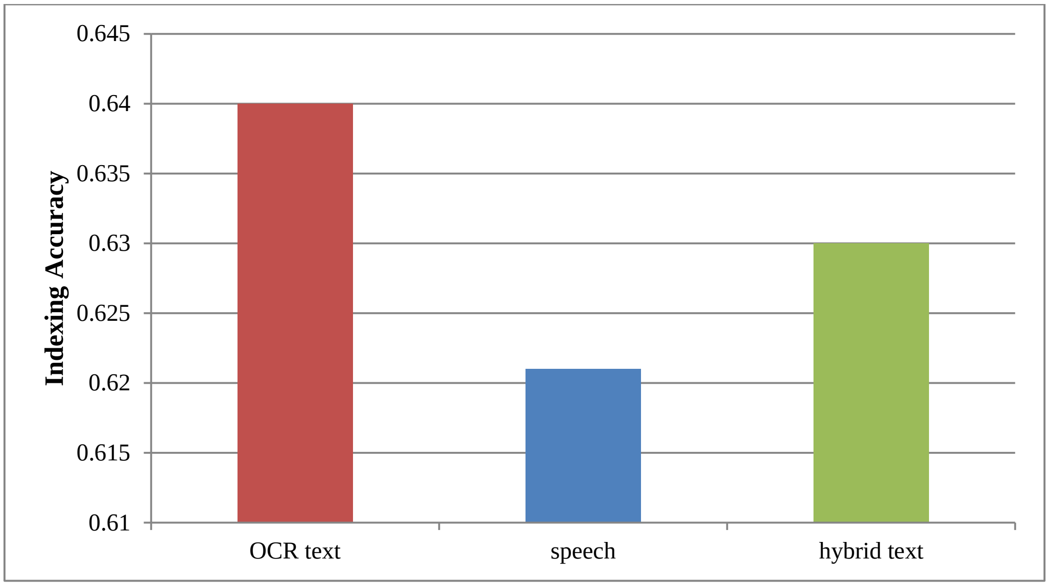


**Figure 5.3: Indexing accuracy with hybrid text of OCR and uncorrected speech text.**

As seen in Figure 5.3, the difference between the performances of all the three text types is marginal, and hybrid text-based indexing does give better indexing accuracy results than uncorrected speech-based indexing, but OCR text-based indexing still performs the best. This means that the text similarity among segments, when the text OCR is combined with the uncorrected speech text, increases since the number of keywords related to topic information increases and thus gains better indexing accuracy. This could also be true due to better text similarity among segments for OCR-based text. However, the hybrid analysis done with nine lectures is not enough to reach any conclusion, and further analysis is required in order to investigate reasons behind the performances of each text type, which is discussed in further sections.

We also performed a hybrid text analysis on OCR speech text and corrected speech text for a different set of lecture videos with IDs 180, 184, 186, 260, 338, 339, 341, 572, and 584. The indexing accuracies for OCR text, uncorrected, and hybrid speech text are summarized in Figure 5.4.

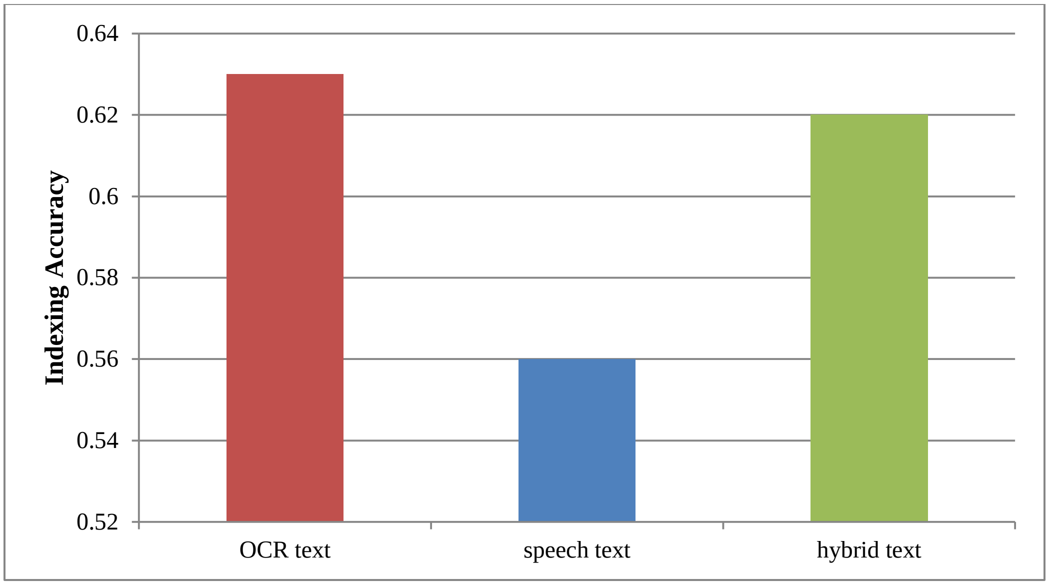


**Figure 5.4: Indexing accuracy with hybrid text of OCR and corrected speech text.**

The evaluation of OCR text and corrected speech text demonstrates that hybrid text indexing accuracy obtains better results than uncorrected/original speech text, but the difference between it and OCR text-based indexing is marginal. The OCR text-based indexing accuracy still performs better.

## 5.3 Test Cases for Evaluation

The analysis of the overall average performances for OCR-based text indexing and uncorrected speech-based text indexing is not enough for a conclusion. The main reason for this is the need to analyze each and every lecture video closely in order to determine the factors causing the index points in a particular text type and not in another. Each lecture video is different in itself. Moreover, there is a need to find out the probable reasons for the selected index points to differ from the ground truth, impacting the overall indexing accuracy performance. For the analysis portion, the fixed grouping algorithm with a maximum grouping duration of 480 seconds was used and the required number of index points for each lecture video was set to the number of index points in the ground truth. The results from all the twenty-five video lectures were used and certain videos were selected for different test cases, depending on the accuracy of the results they obtained.

There are two cases in this scenario:

- *Case 1: Where OCR text performs better than uncorrected speech text.*
- *Case 2: Where uncorrected speech text performs better than OCR text.*

### 5.3.1 Case 1: When OCR Text Obtains Better Indexing Accuracy than Uncorrected Speech Text

For the analysis, the output of the selected video lectures out of twenty-five video lectures was chosen where the indexing algorithm gave better accuracy for the OCR text type than the uncorrected speech text. This was done in order to closely analyze each index point in each video lecture and to determine the probable causes for the selection of a particular point as an index point. The lectures that gave similar indexing accuracy as with uncorrected speech text were not considered for this analysis. Figure 5.5 summarizes the results of the selected video lectures where OCR text-based indexing performed better.

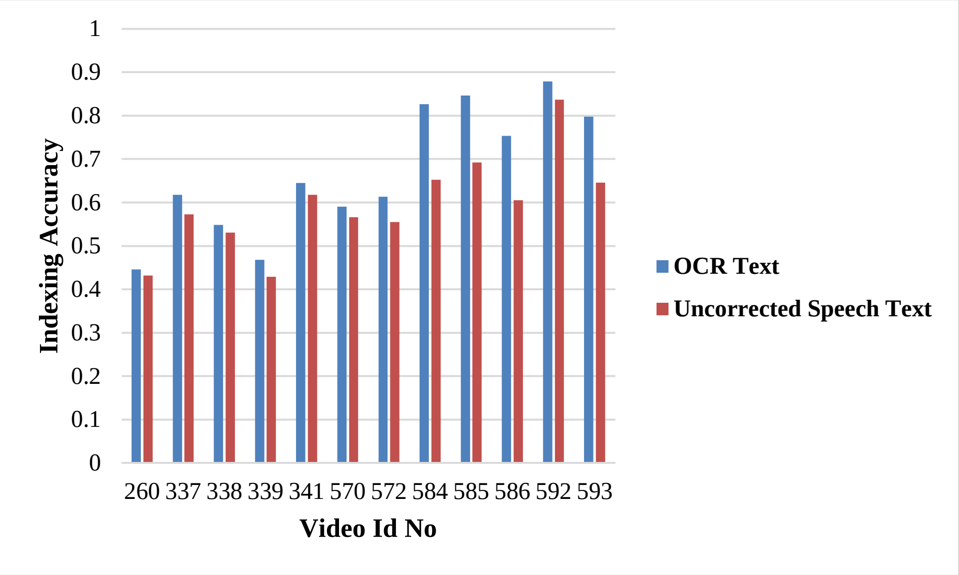


**Figure 5.5: Indexing accuracy for video lectures where OCR text performed better than uncorrected speech text.**

Here, we have not considered video IDs where OCR text and uncorrected speech text resulted in the same indexing accuracy.

### 5.3.2 Case 2: When Uncorrected Speech Text Obtains Better Indexing Accuracy than OCR Text

For the analysis, the output of the selected video lectures out of twenty-five video lectures was chosen where the indexing algorithm resulted in better accuracy for the uncorrected speech text than the OCR text type. This was done in order to closely analyze each index point in each video lecture and determine the probable causes for the selection of a particular point as an index point. The lectures that gave similar indexing accuracy as with OCR text type were not considered for this analysis. Figure 5.6 summarizes the results of the selected video lectures where OCR text-based indexing performed better.

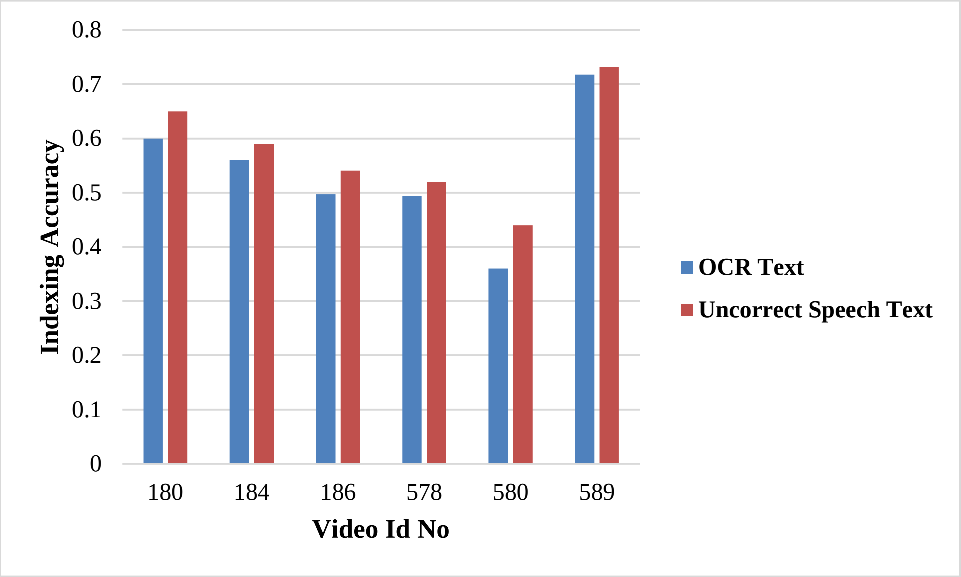


**Figure 5.6: Indexing accuracy for video lectures where uncorrected speech text performed better than OCR text.**

## 5.4 Effect of Speech Text Quality on Indexing Accuracy

The captions generated by YouTube were not perfect and gave different text quality for different video lectures. A scale rate was generated ranging from zero to five and each of the lecture videos was given a rating within this range. For the analysis, each video lecture was heard for 10-15 minutes by randomly playing a single video in between a video lecture and was given a scale rating. The scale rating is as follows:

5- Excellent

4- Very Good

3- Good

2- Average

1- Poor

0-No Text

For the analysis, the lecture videos were categorized based on the scale rating and the overall average indexing accuracy of each category of the scale as was calculated in Table 5.2.

| **Scale** | **No. of Lectures** | **Video IDs** | **Average Accuracy of Uncorrected speech text** |
|---|---|---|---|
| 0 | 0 | NA | NA |
| 1 | 4 | 180,184,186,260 | 0.55325 |
| 2 | 4 | 578,580,592,593 | 0.598 |
| 3 | 9 | 336,337,338,339,341,569,572,573,588 | 0.6 |
| 4 | 8 | 570, 571,584, 585,586,589,590,591 | 0.69 |
| 5 | 0 | NA | NA |

**Table 5.2: Average indexing accuracy for uncorrected speech text of each category of scale.**

The result is summarized in Figure 5.7.

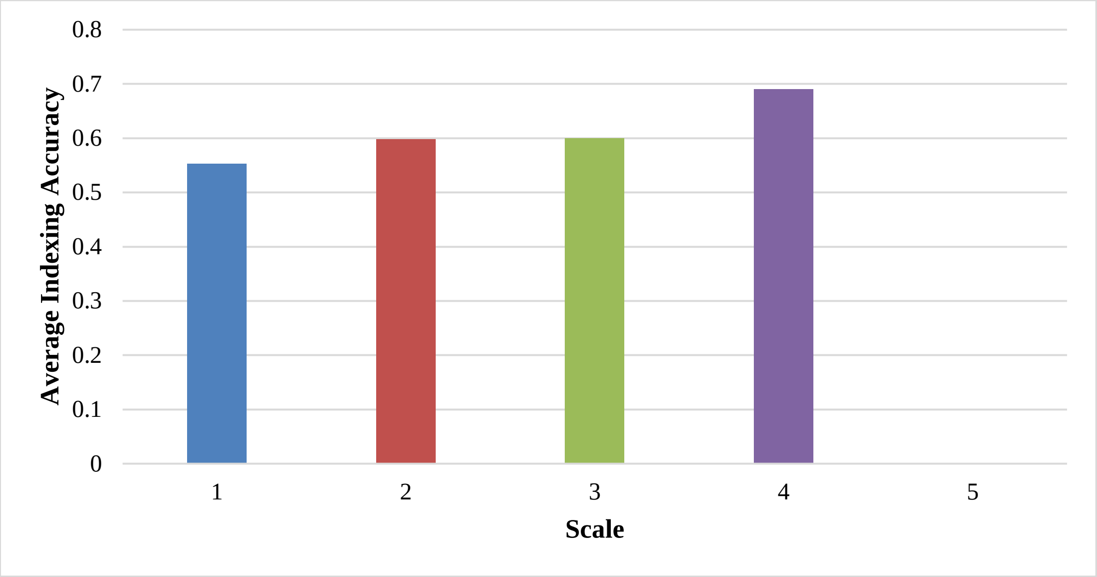


**Figure 5.7: Average indexing accuracy based on the quality of speech text.**

As is evident, it is true that caption quality highly impacts the overall indexing accuracy, and it increases linearly with the increase in the scale rate of the video lectures. The main reason the words that are not recognized properly by the speech recognition give poor indexing results is that these words affect text similarity among segments. It also provided us a way to find ideas to solve this problem, and correcting the captions of video lectures manually seems to be one of the solutions of the problem of poor quality captions. The results and analysis of corrected speech text are discussed in further sections.

## 5.5 Evaluation with Corrected Speech Text

It is evident from Figure 5.5 that the major error impacting the uncorrected speech text indexing accuracy is poor caption quality. To further explore the relationship between the word errors and retrieval effectiveness, we performed the evaluation using manually corrected captions. In order to solve this problem, captions of eleven lecture videos with video IDs 180, 184, 186, 260, 336, 337, 338, 339, 341, 572, and 584 were manually corrected with the help of the ICS captioning tool. The correction of the captions ensures that there is no word missing and captioned inaccurately. Corrected captions get a scale rating of five. The text files for corrected captions were evaluated, in the same way, as uncorrected speech text and OCR text. The ground truth used for the experiments was the same, and the algorithm used was the fixed grouping duration indexing algorithm. The output helped us to determine whether corrected captions gave better performance than uncorrected and text-based indexing. The result of the test is summarized in Figure 5.8.

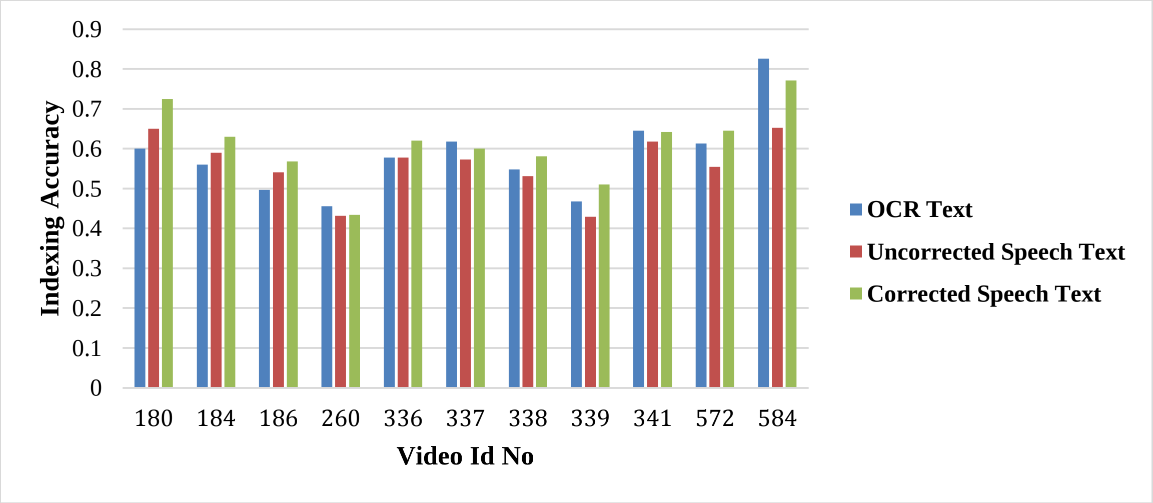


**Figure 5.8: Indexing accuracy with corrected speech text for selected lecture videos.**

It is evident from the output of the test that the correction improves words or text and increases text similarity. The overall difference in the percentage between corrected speech-based text indexing, OCR, and uncorrected speech-based text indexing is summarized in Table 5.3: and its respective graph for overall average indexing accuracy for corrected speech text is shown in Figure 5.9.

| OCR text | Uncorrected speech text | Corrected speech text | Difference in % OCR & corrected | Difference in % uncorrected & corrected |
|---|---|---|---|---|
| 0.582 | 0.55 | 0.611 | 4.98 | 11.09 |

**Table 5.3: Average indexing accuracy for OCR text, uncorrected speech text, and corrected speech text and percentage difference.**

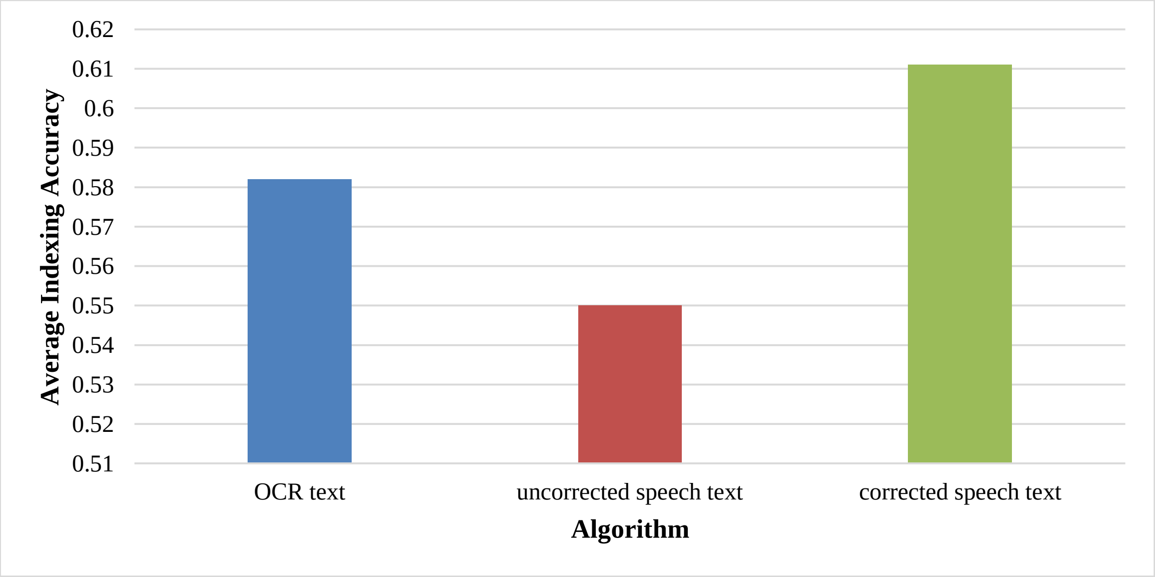


**Figure 5.9: Average indexing accuracy with corrected speech text for selected videos.**

## 5.6 Summary

OCR text-based indexing accuracy results in roughly 4% better performance than uncorrected speech-based text performance. When evaluated separately, it was evident that each lecture is different in its own way (e.g., the duration of lecture, lecture organization, different teaching styles, etc.). The hybrid text type was proposed and evaluated for the speech and OCR text combined together for video indexing purposes. The experiment was performed on nine selected video lectures, which gave better indexing accuracy than individual result for uncorrected speech text type. The results produced by hybrid text on a small number of videos are not enough to come to a decisive conclusion that hybrid text is the best text type for indexing video lectures. It was found that speech words are larger in number than slide words. The errors in speech text are at the word level while they are at the character level for OCR text. Both of these errors have different impacts on indexing video lectures. We have proposed the solutions to this problem. Captions were corrected as a part of one possible solution and results showed an improvement in indexing accuracy. The performance of corrected text-based indexing accuracy increases after lectures were corrected manually. The corrected captions showed 4.9% better results than OCR text and 11% better than uncorrected speech text indexing for selected sets of corrected video lectures.

# Chapter 6. Analysis of Errors

In this chapter, we will discuss the causes of errors in determining an accurate index point compared to the ground truth. After evaluating the results from the OCR text-based, uncorrected speech-based, hybrid, and corrected speech-based indexing, we investigated the probable reasons or causes for the errors in index point selection by the algorithm for a particular text type. In this section, we will discuss three scenarios:

- When OCR text resulted in better indexing accuracy than uncorrected speech text.
- When uncorrected speech text resulted in better indexing accuracy than OCR text.
- When corrected speech text resulted in better indexing accuracy than OCR and uncorrected speech text.

## 6.1 Analysis of Errors: When Uncorrected Speech Text Results in Indexing Errors

The analysis of errors as discussed in Section 5.3.1 investigates and describes various reasons causing errors in indexing for uncorrected speech text. The output of each of the videos in Figure 5.5: Indexing accuracy for video lectures where OCR text performed better than uncorrected speech text that contain index points was analyzed and the reasons causing the errors were investigated closely. Also, the number of occurrences of errors was analyzed. The analysis of errors includes the pattern when ground truth and OCR text mark a point as an index point but the uncorrected speech text is not able to identify it as an index point and hence, marks a wrong transition point as an index point, thus causing poor accuracy and failing to perform better than OCR text. Figure 6.1 summarizes the reasons for the error in indexing and the percentage of occurrences.

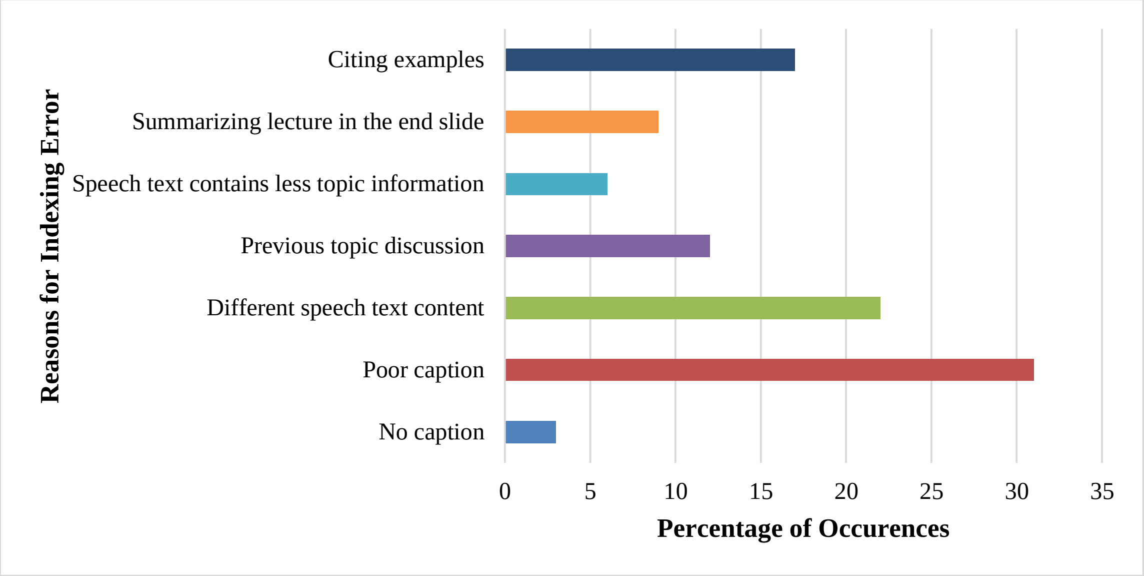


**Figure 6.1: Issues causing indexing errors in uncorrected speech text.**

### 6.1.1 No Caption

Occasionally, an instructor may not speak for a while during a lecture. This may lead to no text data for a particular transition frame or its neighboring frames. Therefore, there is no text for the algorithm to determine text similarity and merge with other transition frames; hence the algorithm is not able to interpret an index point. However, this slide may be marked by an instructor as an index point in the ground truth, thus causing errors in indexing accuracy calculations.

### 6.1.2 Poor Captions

The main reason causing erroneous results in most of the lecture videos for uncorrected speech text-based indexing is the poor quality of caption text that leads to unrecognized text, incomplete sentences, poor captioning of technical words representing topic information, etc. Reduced audio quality predictably degrades the caption quality and indexing accuracy as well. It also depends on the tool dictionary since some of the technical terms that are topic specific might not be present in the speech recognition tool vocabulary. Poor captions may lead to low text similarity among neighboring segments, thus greatly impacting the selection of correct or incorrect index points. This causes errors and lowers the indexing accuracy for uncorrected speech text-based indexing. The results are highly unpredictable for poorly captioned speech. One possible solution could be to manually correct the original caption generated by YouTube. The proposed solution for poor captions was utilizing manually corrected speech text using the ICS caption editor.

#### 6.1.2.1 Reasons for Errors in Captioning

As observed from the Text Accuracy Calculation, none of the lecture videos gave 100% text accuracy, as there were many errors in the captions generated by YouTube. There are various factors that could have affected the accuracy and resulted in erroneous captions. Some of them are described below:

- **YouTube Errors:** It was observed that even though the words were pronounced correctly by the instructor and the audio was clear, the tool gave incorrect captions. The creation of captions is totally tool dependent.
- **Speaker's Accent:** Sometimes the tool is not able to detect some accents. It was observed that due to the heavy accent of some instructors, YouTube was not able

to recognize certain words and gave erroneous captions. Accents result in inaccurate captions or no captions at all.

- **Interaction in Class:** For videos 180-269, the instructor often engaged in classroom interaction. YouTube was not able to recognize the mixed speech and interaction resulting in poor captions.
- **Inaudible Sound:** Sometimes instructors move away from the microphone and the volume becomes too low for YouTube to recognize. This often happens in a classroom.

### 6.1.3 Different Speech Text Content

A different speech text could be caused by issues, such as the instructor talking with the students away from the microphone, an off topic discussion like assignments or exams, a different topic discussion, etc. The speech text content of such a different discussion may not be similar to the main topic. Thus, the indexer is unable to merge such transition segments to the main topic due to its low text similarity to the main topic, resulting in an index point or the start of a new topic as expected to be detected by the algorithm. This transition point might not be considered by the instructor to be an index point in the ground truth, which causes errors.

### 6.1.4 Previous Topic Discussion

The instructor might keep discussing a previous topic as a part of a summary of the previous topic after the slide change or at the start of a new topic. This leads to unexpected results, as the text similarity could be low or high in this scenario. The text similarity is relatively low to the segment after it, but high to the segment previous to it. This causes the algorithm to detect incorrect index points when compare to the ground truth.

### 6.1.5 Speech Text Contains Less Topic Information than OCR Text

Sometimes speech text contains less topic information than the text present in the slide. The text in a slide contains definite information about a particular topic, whereas the speech texts sometimes contain more random text and fewer keywords related to the main topic. In addition, the speech tends to be improvised and less prepared, unlike the text in slides. This leads to low text similarity, when there should be more in transition frames containing topic information, and high text similarity among incorrect transition frames that do not represent a topic. However, the behavior is expected by the algorithm, as it detects the change in a topic and marks it as index point. This causes incorrect index point detection. The speech text content depends entirely on an instructor’s method of teaching a particular topic.

### 6.1.6 Summarizing Lecture in End Slide or First Slide

An instructor may choose to summarize the whole lecture verbally in an end slide, which may not be recognized as an index point by the instructor for the ground truth due to the text “end”. This may have no similarity with the previous topic, but it might have a similarity with the first slide where the instructor may choose to discuss the topics of the lecture. If this segment is marked as an index point by the algorithm (which is expected), this results in an error.

### 6.1.7 Citing Examples

Citing an example is a very common method observed in various lectures. The examples might not be present in the slide itself, but the instructor uses the example in order to explain the topic or to support certain subtopics within the lecture. This extra speech text added to the data causes low text similarity with the neighboring segments and is not able to merge with them causing a separate index point, which was not recognized by the ground truth.

## 6.2 Analysis of Errors: When OCR Text Results in Indexing Errors

The analysis of errors as discussed in Section 5.3.2 investigates and describes various reasons causing errors in indexing for OCR text. The output of each of the videos in Figure 5.6: Indexing accuracy for video lectures where uncorrected speech text performed better than OCR text containing index points was analyzed and the reasons causing the errors were investigated. Also, the number of occurrences of the errors was analyzed. Analysis of errors includes the pattern when the ground truth and uncorrected speech text mark a point as an index point, but the OCR-based slide text is not able to identify it as an index point and marks an inaccurate transition point as an index point, causing poor accuracy and failing to perform better than the uncorrected speech text. All the reasons mentioned here is where the OCR text resulted in errors in index points whereas the speech text was able to recognize the correct index points. Figure 6.2 summarizes the reasons of error in indexing and the percentage of occurrences.

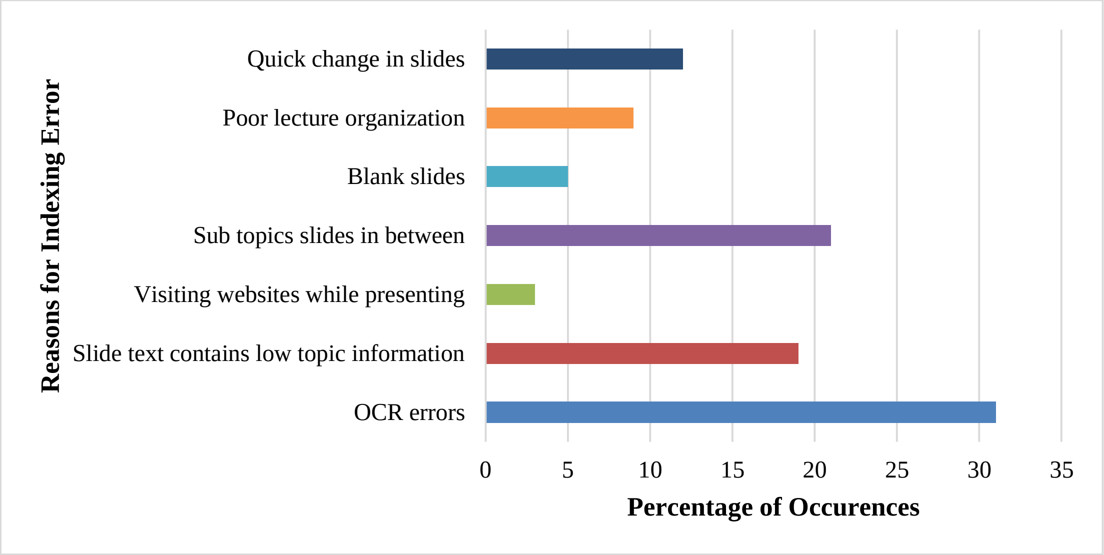


**Figure 6.2: Issues causing indexing errors with OCR text.**

### 6.2.1 OCR Errors

The OCR errors result in different or incorrect text data due to the failure to recognize the characters accurately. There could be various reasons for this, such as the size of the characters, presence of formula, or handwritten texts in a slide, etc. The OCR errors are a major source of errors causing factors in OCR text-based indexing and the results could be highly unpredictable. Speech text indexing could potentially provide significantly better indexing in this case.

### 6.2.2 Slide Text Contains Low Topic Information and More Images

In some of the slides, the text content may be very low and may not represent the actual topic information. There could be more images in the view graphs than in the textual information. The instructor may choose to explain the images verbally in addition to what is written on the slide. In this case, text-based indexing fails to decide whether to merge toward the left or right frame. A hybrid approach of combining the text, image, and audio data could be a possible solution to solve this problem.

### 6.2.3 Visiting Websites While Presenting

As observed while investigating the errors, the instructors in some of the lecture videos may choose to represent a topic or example by visiting websites (e.g., the university login website, google web page, or images page, etc.). Such scenarios cause a break in the lecture organization or a break in the topic being presented. Due to the low text similarity of the texts extracted from these web pages, they are not merged with the actual topic and are considered a new topic or are marked as an index point by the algorithm although not marked as an index point for ground truth. This causes errors and one possible solution of this could be to not consider web page text for indexing or possibly to not merge audio with these kinds of texts. Speech-based text indexing could potentially produce better results in videos having such cases.

### 6.2.4 Sub Topic Slides in Between

Occasionally, the subtopic or the outline slide in between a lecture causes a break in the linear lecture organization and leads to marking such outline slides as index points. Conversely, the outline slide contains sub-topics in the form of points for overview of the actual topic. The instructor may mark an outline slide as a new topic or index point in the ground truth, but the algorithm may merge it with the actual topic due to the presence of text data in that slide that causes errors. Speech-based indexing, however, proved to provide better indexing results in such scenarios.

### 6.2.5 Blank Transition Frames/ Slides

Occasionally, there were many blank transition frames or slides in some of the lecture videos. It was observed that these blank slides were the results of the paint screen being open for a long period of time, or dragging a document in class that contains blank first and last pages. This leads to no text content in the OCR text, and the algorithm fails to decide whether to mark it as an index point or merge it with other slides. Audio-text-based indexing gives surprisingly better results in these cases since the instructor continues to explain the actual topic, even if the slide on the screen is blank.

### 6.2.6 Poor Lecture Organization

The text-based indexing algorithm follows a linear lecture organization. Many lecture videos were found to break this rule. Some of the lecture videos followed very poor organization, such as dragging a doc or pdf file in class, jumping to web pages in between, playing a short video in between, etc. Due to all these factors, there is a break in the lecture sequence or topic sequence. The text similarity in this case is generally low with the previous segment; therefore, the algorithm marks such frames as index points causing errors.

### 6.2.7 Quick Change in Slides

The OCR is not able to detect text or transition frames when there is a quick change in slides during a lecture. The text related to topic information is missed or not extracted completely. The incomplete text does not represent the actual topic as a whole and results in inaccurate index points since it greatly affects the text similarity.

## 6.3 Analysis of Errors: When Corrected Speech Resulted in Better Results than Uncorrected Speech

The main reasons corrected speech text proves to be better option are as follows:

- Incorrect spellings,
- Unrecognized/inaudible words,
- Out of vocabulary important keywords/words or domain specific terms,
- Complete sentences making more sense,
- Better text quality than poor OCR and original speech text, and
- Words with multiple usages or meanings.

The bullet points above are described in detail below.

The captions generated by the speech recognition software produce errors at the word level. Sometimes, the words are out of their vocabulary. The correction of uncorrected or original speech text performs better in terms of indexing accuracy and generating accurate index points. The analysis was done on the corrected speech text in order to determine the probable causes for better performance. The most common reason found was that most of the poorly captioned words were captioned correctly and helped the indexing algorithm determine text similarity between segments. This leads the transition frames to merge if it contains certain topic information within it. The corrected captions made more sense than the uncorrected version, where sometimes most of the sentences in a transition frame as a whole did not make any sense due to poor captioning, unavailability of captions for reasons, such as distance of the instructor from the audio recording tool, or discussion with students, etc., and missing out on important words in a sentence. The corrected text helped the algorithm determine text similarity among transition frames and mark a transition point as an index point. However, corrected speech text obtained better results for some selected lecture videos, yet it is still difficult to prove that it is the best text type for indexing all the classroom video lectures since the caption correction experiment was performed on eleven selected video lectures. Moreover, the correction of captions is a tedious, manual, and time-consuming task and not feasible for a large number of videos. There are also words used by instructors in lecture videos that have multiple meanings in relevance to various topics and sometimes this leads to errors in corrected speech text indexing. The most significant error similarity among these selected corrected video lectures was that the OCR text was very poor and the original speech was also of bad quality with a scale rating from one to three. This gave an added advantage to the corrected speech text type to perform better in terms of indexing accuracy. For such lecture videos where the OCR text is not available or completely erroneous, corrected speech-based text indexing could be utilized.

## 6.4 Overall Summary

Different cases were evaluated in order to analyze the indexing errors. The text-based indexing algorithm favors linear organization of lectures. When lecture organization is abrupt or includes an outline or examples in between, it tends to lead to indexing errors. Also, hybrid-based text indexing was evaluated where the text from OCR and text from uncorrected speech were combined and evaluated against the indexing algorithm.

Typically, poor captions and OCR errors were found to be the most common reasons for indexing errors in both the cases where one text type was better than the other text type. Various other reasons were analyzed and recognized as causing indexing errors in both the text types. Speech text indexing performs fairly well when lecture organization is not linear since this is a big challenge for text-based algorithms based on OCR text. Speech is improvised, and there is no control over the content the instructor discusses in class as homework, off topic discussions, exams, interaction with students, an inaccurate speech recognition tool, and words out of vocabulary, etc. lead to the occurrences of errors in speech text-based video indexing. For this, correction of captions makes a huge impact on increasing the indexing accuracy. Corrected captions make more sense in terms of words, forming complete sentences, etc. Organization of lecture and properly captioned speech text are also important factors for the performance of indexing accuracy. The text similarity between segments does not necessarily indicate a topic, but the algorithm is based purely on that. Hence, there is a need for a more robust and better algorithm that utilizes the relative strengths of both OCR and speech text types.

# Chapter 7. Conclusions

## 7.1 Conclusions

The indexing of video lectures has a significant impact on the student learning and development process. The content of a video lecture holds equal importance for content mentioned in the slides as well as the speech of the instructor. The first conclusion is that the speech text and slide texts are not same. Speech text is more abundant and can be more or less descriptive depending on the instructor's method of teaching. Speech is improvised, but the text in the slides is well-prepared and rarely improvised. The indexing algorithm used to perform indexing was based on OCR text from the slides of the lecture video. The data in the slide text are not enough to present a topic. Therefore, the speech-based method was proposed, developed, and evaluated against the OCR text-based methods. The overall average accuracy seems to be marginally higher for OCR-based text indexing than the uncorrected version. Enhancement to the text data was achieved by further proposing a hybrid approach and combining the OCR and uncorrected speech text data for evaluation on few of the lecture videos. It was found that there is a minimal difference in performance, and OCR text still performs better. The output of each video lecture for the OCR text type and the uncorrected speech text type was analyzed in order to determine the causes of error when one text type performs better than other and vice versa. It was found that there are various scenarios when uncorrected speech text performs better than OCR text and vice versa. Uncorrected speech-based text indexing detects topic changes better when the OCR text is erroneous, the presentation is not organized linearly, the slide changes are really quick, etc. On the other hand, the OCR text performs better when the uncorrected speech text is erroneous and contains different random data than in the topic. The speech text content completely depends on the way an instructor talks or explains a topic. There can be concise and accurate lectures or others with a lot of grammatical errors. A set of twenty-five video lectures is not enough to come to a decisive conclusion. The scale rate of the quality of each lecture video was designed and given based on how accurate the quality of the caption text is in order to determine how quality impacts indexing accuracy for uncorrected speech-based text. It was found that the quality of speech text affects indexing accuracy. Therefore, some of the lectures were selected at random for manual speech correction, and their results were evaluated to determine the impact of the corrected speech text version on indexing accuracy. It was observed that corrected speech text performs fairly well over OCR text and uncorrected speech text in detecting the topic changes accurately. This proves that the word level errors in the speech-based text directly impact performance. The accuracy of the indexing algorithm is limited by the text similarity metric, ground truth, and other factors, such as lecture organization, text quality, etc. The results indicate that different media other than just slide text contains important keywords and relevant information about a topic and can be utilized  for achieving better indexing results.

## 7.2 Future Directions

The text data from the speech or from the slide of a video lecture alone may not be enough to represent complete topic information. There could be a possibility of getting better results if hybrid tests are conducted on a larger set of video lectures in order to determine the topic organization. Tests could be performed using various text similarities metrics in order to determine the effect on indexing accuracy. The OCR text data could also be corrected and combined with corrected speech text for further analysis. Another area of consideration would be the effect of unsupervised machine learning techniques to determine the topic boundaries and the effect on indexing accuracy. The development of a better model to represent a lecture by an instructor would also prove to be useful in enhancing accuracy results. There could also be the possibility of a change in accuracy results with the change in the method of collecting ground truths by additional observation and evaluation by a team of experts in a particular field.  There is also the possibility of finding methods to detect the text even if slide changes are quick and the lecture is not organized linearly. An automated tool development for correcting and processing audio content of video lectures could also be one major enhancement in the field of corrected audio-text-based indexing. In short, by utilizing the speech text and slide text we can design an indexing strategy that uses the relative strengths of both text types.

# References

[1] J. Subhlok, O. Johnson, V. Subramaniam, R. Vilalta and C. Yun, "Tablet PC video based hybrid coursework in computer science: report from a pilot project," *ACM SIGCSE Bulletin,* vol. 39, pp. 74-78, 2007.

[2] J. V. Miro, R. N. Spencer and A. Pérez González de Martos, "Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures," *Open Learning: The Journal of Open, Distance and e-Learning,* vol. 29, no. 1, pp. 1-2, 2014.

[3] B. Coüasnon and E. . K. Ringger, "A machine learning based lecture video segmentation and indexing algorithm," in *Document Recognition and Retrieval XXI,* San Francisco, 2014.

[4] H. Yang, F. Grünewald, M. Bauer and C. Meinel, "Lecture Video Browsing Using Multimodal Information Resources," *Advances in Web-Based Learning-ICWL, Springer,* vol. 8167, pp. 204-207, Berlin Heidelberg, 2013.

[5] T. Tuna, J. Subhlok, L. Barker, V. Varghese, O. Johnson and S. Shah, "Development and evaluation of indexed captioned searchable videos for STEM coursework," in *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education,* pp. 129-134, 2012.

[6] H. Yang and C. Meinel, "Content Based Lecture Video Retrieval Using Speech and Video Text Information," *IEEE,* vol. 7, no. 2, pp. 142-145, 2014.

[7] T. Tuna, " Search in Classroom Videos with Optical Character Recognition for Virtual Learning," M.S. thesis, University of Houston, Houston, 2012.

[8] "ICS Videos," University of Houston, [Online]. Available: www.icsvideos.uh. [Accessed 28 October 2014].

[9] R. Borgaonkar, "Captioning for Classroom Lecture Videos," M.S. thesis, University of Houston, Houston, 2013.

[10] "YouTube," [Online]. Available: www.youtube.com. [Accessed 3 September 2014].

[11] S. K. Ch and S. Popuri, "Impact of Online education," in *Innovation and Technology in Education (MITE), 2013 IEEE International Conference in MOOC* , Jaipur, 2013.

[12] H. J. Jeong, T.-E. Kim, . H. G. Kim and M. H. Kim, "Automatic detection of slide transitions in lecture videos," *Multimedia Tools Application, Springer,* vol. 73, no. 3, pp. 1-18, 2014.

[13] W. Hu, N. Xie, L. Li, X. Zeng and S. Maybank, "A Survey on Visual Content-Based Video Indexing and Retrieval," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—,* vol. 41, no. 6, pp. 797-819, 2011.

[14] H. D. Wactlar, T. Kanade, A. Smith and S. M. Stevens, "Intelligent access to digital video: Informedia project," *Computer*, vol. 29, no. 5, pp. 46-52, 1996.

[15] R. Lienhart and W. Effelsberg, "Automatic text segmentation and text recognition for video indexing," *Multimedia System*, vol. 8, no. 1, pp. 69-81, 2000.

[16] H. Mo, H. Yamagishi, I. Ide, N. Katayama, S. Satoh and M. Sakauchi, "Key image extraction from a news video archive for visualizing its semantic structure," *Advances in Multimedia Information Processing-PCM, Springer,* vol. 3331, pp. 650-657, 2004.

[17] J. Nandzik, B. Litz, N. Flores-Herr, A. L¨ohden, I. Konya, D. Baum, A. Bergholz, D. Sch¨onfu, C. Fey, J. Osterhoff, J. Waitelonis, H. Sack, R. K¨ohler and P. Ndjiki-, "Contentus—technologies for next generation multimedia libraries," *Multimedia Tools and Applications,* vol. 63, no. 287-291, 2012.

[18] H. Yang, H. Sack and C. Meinel, "Lecture video indexing and analysis using video ocr technology," *International Journal of Multimedia Processing and Technologies,* vol. 2, no. 4, pp. 176-196, 2012.

[19] H. Yang, C. Oehlke and C. Meinel, "An automated analysis and indexing framework for lecture video portal," in *International Conference on Web Based Learning*, Sinaia, Romania, 2012.

[20] D. Zhang and J. F. Nunamaker, "A Natural Language Approach to Content-Based Video Indexing and Retrieval for Interactive E-Learning," *IEEE,* vol. 6, no. 3, pp. 450-457, 2004.

[21] S. Repp and . C. Meinel, "Semantic indexing for recorded educational lecture videos," in *4th IEEE Conference on Pervasive Computing and Communications Workshop,* Pisa, Italy, 2006.

[22] A. Gandhe, L. Qin, F. Metze, A. Rudnicky and I. Lane, "Using web text to improve keyword spotting in speech," in *Proceedings of Automatic Speech Recognition and Understanding, IEEE,* Olomouc, 2013.

[23] M. Cooper, "Presentation Video Retrieval using Automatically Recovered Slide and Spoken Text," in *Proceedings of SPIE-IS&T Electronic Imaging,International Society for Optics and Photonics*, 2013.

[24] T. Kawahara, Y. Nemoto and Y. Akita, "Automatic lecture transcription by exploiting presentation slide information for language model adaptation," in *Proceedings of Acoustics, Speech and Signal Processing, IEEE,* Las Vegas, 2008.

[25] R. Swaminathan, M. . E. Thompson, S. Fong, A. Efrat, A. Amir and K. Barnard, "Improving and aligning speech with presentation slides," in *International Conference on Pattern Recognition*, Istanbul, 2010.

[26] G. Szaszak, M. Cernak, P. N. Garner, P. Motlicek, A. Nanchen and F. Tarsetti, "Automatic speech indexing system of bilingual video parliament interventions," Idiap-RR, 2013.

[27] "Coursera," Coursera, [Online]. Available: https://www.coursera.org/. [Accessed 24 September 2014].

[28] "TomatoSoft," [Online]. Available: http://www.tomatosoft.biz/blog/2012/10/03/free-video-cutter/. [Accessed 3 December 2013].

[29] V. K. Varghese, "Development and Evaluation of Text-based Indexing for Lecture Videos," M.S. thesis, University of Houston, Houston, 2014.