

CAPTIONING 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

Rucha Borgaonkar

December 2013

CAPTIONING FOR CLASSROOM LECTURE

VIDEOS

Rucha Borgaonkar

APPROVED:

Dr. Jaspal Subhlok, Advisor

Dr. Olin Johnson

Dr. Nouhad Rizk

Dr. Holly Hutchins

Dr. Lecia Barker

Dean, College of Natural Sciences and Mathematics

Acknowledgements

I am thankful to my advisor, Dr. Jaspal Subhlok, for his encouragement and guidance throughout the course of this thesis. Our discussions always helped me to look through a new perspective and helped me in accomplishing the goal of this work. I want to thank Dr. Olin Johnson, Dr. Shishir Shah, Ms. Leigh Hollyer, and Dr. Rizk for their co-operation during the various evaluation phases. I want to thank Dr. Lecia Barker for conducting the focus groups with us and for her expert comments and feedback for conducting the surveys. I also want to thank Dr. Holly Hutchins for her expert advice on the survey questions and guiding me to improve the questionnaire.

This work has been partially supported by National Science Foundation's Division of Undergraduate Education under the Course, Curriculum and Laboratory (CCLI) program with Award No. DUE-0817558. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

I will always be thankful to my parents and my in-laws for believing in me. I also want to thank my brother, Rohan, and sister-in-law Dipali for always keeping me in a positive spirit. I also want to thank Ketki and Aayush for their love and support.

Last but not the least, I want to thank my husband Gaurav Deshpande. This thesis would not have been possible without his love and support.

CAPTIONING 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

Rucha Borgaonkar

December 2013

Abstract

Tablet PC-based lecture videos are widely used by students at the University of Houston. The goal of the ICS (Indexed, Captioned, and Searchable) Videos Project is to enhance user experience and improve usability of classroom lecture videos. The goal of this thesis is to assess the value as perceived by students, of having captions for classroom lectures and to find methods of generating captions as automatically and efficiently as possible.

In this thesis, we established through survey and focus groups, that students highly valued having captions for their classroom lectures. The accuracy of currently available speech recognition tools was assessed with the assistance of three faculty members. This revealed that the output of these tools for spontaneous speech cannot be used directly as captions due to its low accuracy level. This text needs to be manually corrected. To edit these automatically generated captions, we designed, implemented, and evaluated a web-based crowd-sourcing caption editing tool. This tool can be used by a group of people to correct the captions simultaneously. This reduces the time and effort required to correct the captions, which otherwise is a monotonous task. We evaluated the caption editor through a study and a survey with 28 students from two classes. The students in the class were able to effectively combine their efforts such that a full class lecture was captioned with a very modest effort from each student.

In this thesis we established a semi-automatic process to generate corrected captions for classroom lectures efficiently. The captioning module is integrated with the ICS Videos player, thereby allowing navigating the videos from text captions. Focus

groups and other research show that captions and transcripts are highly valued by students, especially those with hearing disability and whose first language is not English.

Table of Contents

Abstract	v
Chapter 1 : Introduction	1
1.1 Motivation	1
1.2 Background.....	2
1.2.1 Video Indexes	3
1.2.2 Search Inside the Videos	4
1.2.3 Captions	5
1.3 Thesis Outline.....	6
1.4 Related Work.....	7
1.4.1 Assessment of Current Speech Recognition Tools.....	7
1.4.2 Speech Recognition Technology	8
1.4.3 Caption Editors	9
Chapter 2 : Assessment of Current Speech Recognition Tools	12
2.1 Dictation	13
2.2 Lecture Transcription	14
2.3 Parroting	16
2.4 Analysis of Errors.....	17
2.5 Conclusions	21
Chapter 3 : ICS Caption Editor	23
3.1 Usability Requirements	23
3.2 Design.....	24
3.3 Implementation.....	30
Chapter 4 : Evaluation of the Caption Editor	36
4.1 Study and Results	36
4.1.1 Conclusions	43
4.2 Survey Results	43
4.2.1 Conclusions	57
Chapter 5 : Assessment of Value of Captions	59
5.1 Results of Focus Group Conducted in Fall 2012	59
5.2 Survey Conducted in Spring 2013	60
5.2.1 Survey Results	61
5.2.2 Conclusions	73
Chapter 6: Summary	74
6.1 Conclusions	74
6.2 Future Work.....	76

List of Figures

Figure 1-1: ICS Video Player	3
Figure 1-2: ICS Video Player with index image highlighted	4
Figure 1-3: ICS Video Player with search keyword and corresponding index point highlighted.	5
Figure 1-4: ICS Video Player with captions and transcript box highlighted	6
Figure 3-1: ICS Caption Editor login screen	25
Figure 3-2: ICS Caption Editor interface	26
Figure 3-3: ICS Caption Editor interface when no text from ASR is available.....	28
Figure 3-4: ICS Caption Editor interface with audio converted as video	29
Figure 3-5: Communication between Java program and YouTube to get the caption file	31
Figure 3-6: Communication between the Caption Editor, Server, and database	35

List of Tables

Table 2-1: Accuracy with WSR and DNS for Dictation	14
Table 2-2: Accuracy of DNS and YouTube for lecture transcription of Dr. Shah	15
Table 2-3: Accuracy of DNS and YouTube for lecture transcription of Dr. Johnson	15
Table 2-4: Accuracy of DNS and YouTube for lecture transcription of Ms. Leigh	16
Table 2-5: Results of parroting lecture audio	17
Table 2-6: Analysis of errors of recognition by DNS.....	18
Table 4-1: COSC1300: Lecture 23: Performance of individual participant	37
Table 4-2: COSC1300: Lecture 23: Progress of work.....	38
Table 4-3: COSC2410: Lecture 11: Performance of individual participant	39
Table 4-4: COSC2410: Lecture 11: Progress of work.....	40
Table 4-5: COSC2410: Lecture 13: Performance of individual participant	41
Table 4-6: COSC2410: Lecture 13: Progress of work.....	42

List of Charts

Chart 4-1: COSC1300: Lecture 23: Progress of work	38
Chart 4-2: COSC2410: Lecture 11: Progress of work	40
Chart 4-3: COSC2410: Lecture 13: Progress of work	42
Chart 4-4: Academic Year	44
Chart 4-5: Gender	44
Chart 4-6: Ethnicity	45
Chart 4-7: Age	45
Chart 4-8: Major	46
Chart 4-9: Fluency with English Language	46
Chart 4-10: Course of the lecture captioned	47
Chart 4-11: Ease of use	47
Chart 4-12: Use of slide images	48
Chart 4-13: Technical quality of the audio	49
Chart 4-14: Technical quality of the audio for COSC1300 course	49
Chart 4-15: Technical quality of the audio for COSC2410 course	50
Chart 4-16: Clarity of professor's speech	50
Chart 4-17: Clarity of professor's speech for course COSC1300	51

Chart 4-18: Clarity of professor’s speech for course COSC2410	51
Chart 4-19: Use of PlaySpeed tool	52
Chart 4-20: Use of “Needs a review” feature	53
Chart 4-21: Indication of captions that are complete and those that needed more work by the interface	54
Chart 4-22: Use of How-to-use video	54
Chart 4-23: Use of the HELP link	55
Chart 4-24: Position of elements on the interface	56
Chart 4-25: Students interest to edit captions	57
Chart 5-1: Academic year	61
Chart 5-2: Gender	61
Chart 5-3: Identification of disability	62
Chart 5-4: Ethnicity	62
Chart 5-5: Age	63
Chart 5-6: Fluency with English language	63
Chart 5-7: Major	64
Chart 5-8: Course for which captions were viewed	65
Chart 5-9: Usage of videos with captions for course COSC1300	65
Chart 5-10: Usage of videos with captions for course COSC2410	66
Chart 5-11: Usage of videos with captions for course COSC1300	66
Chart 5-12: Usage of videos with captions for course COSC2410	67
Chart 5-13: Use of captions to understand the professor	67
Chart 5-14: Accuracy of captions	68
Chart 5-15: Effect of captions on learning experience	69
Chart 5-16: Use of transcript to understand video contents quickly	70
Chart 5-17: Use of transcript to go directly to a point in the video	70
Chart 5-18: Usefulness of the highlighting feature of the transcript	71
Chart 5-19: Preference to video with/without captions	72
Chart 5-20: Usefulness of captions in another language	72

Chapter 1 : Introduction

1.1 Motivation

Lecture videos are now widely used by many academic institutions. The overwhelming response at institutions such as the University of Houston, Khan Academy, Stanford University, MIT, etc. for these videos proves that the lecture videos are a very powerful resource. With recent developments in education such as MOOCs (Massive Open Online Course), there is a need to make these videos available as widely as possible to make the mission of “providing high quality education to anyone, anywhere” [38]. Captioning is important to understand the content of these videos for the deaf or hearing impaired. Captioning helps students, especially whose first language is not English, to use and understand the content of the videos of lectures delivered by professors with heavy accents. Moreover, most of the lecture videos have specialized vocabulary. In such cases, “seeing” the spoken words can be very helpful. If the transcription is available in one language, the captions can be made available in any language desired to cater to the target audience.

There is a lot of ongoing research to ease the navigation and search within these videos. Since captioning provides synchronization between the audio and the text, it can make the video content searchable. This is very useful, especially for lectures where students remember a particular reference made by the professor during the lecture. If the

transcript is searchable, students can pin-point the specific time in the video. This can avoid having to watch the whole video to find the topic of interest.

According to the Twenty-first Century Communications and Video Accessibility Act (CVAA) [39], the law updates federal communications law to increase the access of persons with disabilities to modern communications. It requires that the video programming that is closed captioned on TV to be closed captioned when distributed on the Internet. This shows that laws are being updated to make as much content available to a wider audience.

To make captioning available for these videos, we need to generate the captions as automatically and as efficiently as possible. Currently available speech recognition technology is not up to the mark for such videos. The output given by these tools needs to be corrected manually, as it cannot be used directly as captions. Therefore, we are motivated to develop a system that combines the automatic generation of captions with a step of manual correction. As manual correction of the captions is a monotonous task, we need to build a tool that makes this task easier. Moreover, to reduce time, effort, and the heavy workload on the person correcting the captions, we need to divide the task among a group of people.

1.2 Background

At the University of Houston, lectures are recorded and made available for students through the web. The research conducted by the ICS (Indexed, Captioned, and Searchable) Videos Project is focused on providing ease of

navigation, and efforts are being made to make these videos available as widely as possible [40, 46, 47, 48]. A customized Video Player – the ICS Video Player (Figure 1.1) is built that supports all the advance functionalities needed by ICS Videos. It displays the indexes, search box, and captions. Indexes, search feature, and captions are discussed in the following section.



Figure 1-1: ICS Video Player

1.2.1 Video Indexes

The Video Indexes are topics identified in the video so that students can easily navigate to their point of interest in the video [47, 48]. Changes in the video frames called as ‘transition points’ are identified. Text-based analysis is used on these transition points to filter transition points and identify indexes. An index point is shown in Figure 1-2.



Figure 1-2: ICS Video Player with index image highlighted

1.2.2 Search Inside the Videos

This module enables the ‘Search’ functionality inside the video. Optical Character Recognition (OCR) technology is used to identify the text on video frames, which are basically images. This text is stored in the database in the form of ‘keywords’. Users can use these keywords to search inside the video. Digital image processing techniques have been used to enhance the accuracy of the OCR tools [40]. In Figure 1-3, the search box and the search result is highlighted.



Figure 1-3: ICS Video Player with search keyword and corresponding index point highlighted.

1.2.3 Captions

Closed captions are created to enhance the accessibility to the lecture videos. Clickable captions provide a way to quickly reach the point when a particular sentence was spoken or a related topic was discussed. Captions are displayed on the video as well as a complete transcript is displayed in the ICS Video Player. The captions and transcript can be turned on/off by the user. The main goal of this thesis is to find ways of delivering captions efficiently. In Figure 1-4, the captions and transcript are highlighted.



Figure 1-4: ICS Video Player with captions and transcript box highlighted

1.3 Thesis Outline

This thesis is organized as follows: Chapter 2 describes the assessment of the accuracy of currently available speech recognition tools for dictation, lecture transcription, and parroting. It explains why we cannot use the output given by these tools directly as captions. We also present the analysis of errors produced. In Chapter 3, we discuss the design and the implementation of the web-based crowd-sourcing caption editor. The study and the evaluation results of the caption editor are explained in Chapter 4. In Chapter 5, we look at the results of the assessment of value of captions for classroom lectures. A summary of the thesis, the final conclusion, and an overview of future work are discussed in Chapter 6.

1.4 Related Work

This section surveys some of previous efforts in the field of assessment of speech recognition tools, speech recognition technology, and caption editing.

1.4.1 Assessment of Current Speech Recognition Tools

The accuracy of commercial automated speech recognition (ASR) systems in conversational speech was assessed by Broughton [10]. Two commercial tools were assessed, Dragon Naturally Speaking 5.0 and IBM ViaVoice 8.0. It is concluded in the paper that there is a significant degradation in the accuracy of commercial ASR tools when conversational or spontaneous speech is used.

Kate Hone in [28] reports a questionnaire measure for the Subjective Assessment of Speech System Interface (SASSI). This research program intends to produce a valid, reliable, and sensitive measure of users' subjective experiences with speech recognition systems. The research suggested factors like system response, accuracy, likeability, cognitive demand, annoyance, habitability, and speed in gauging users' perception of speech systems. Casali, Williges, and Dryden [29] determined adjective pairs such as accurate/inaccurate, consistent/inconsistent, simple/complicated, pleasing/irritating, facilitating/distracting, etc. for acceptance of speech systems.

In this thesis, we evaluate the accuracy of commercial ASRs in terms of percentage of words accurately identified by Dragon Naturally Speaking Preferred 10 and Windows Speech Recognition (Windows 7) for dictation and parroting. We also evaluate

the accuracy of Dragon Naturally Speaking Preferred 10 and Google Speech Recognition for lecture transcription.

1.4.2 Speech Recognition Technology

There have been extensive efforts in the field of speech recognition technology. There is a lot of room for improvement in this technology, as the tools currently available provide very low accuracy. The most widely used approaches in speech recognition are Hidden Markov Model (HMM)[31] and Dynamic Time Warping (DTW)[31].

Ongoing research at Google [30] is focusing on improving algorithms for speech recognition. YouTube provides captions for uploaded videos using Google Speech Recognition. Misra [31] proposes better segmentation methods by using alternative audio features and a discriminative classifier that could be important for web videos that have noisy backgrounds.

Many speech recognition tools use a parameter called the ‘Confidence’ for the hypothesized words. ‘Confidence’ is the probability that the recognition is correct [33]. There has been a lot of research to determine how this value could be accurately measured and used. New methods of rejecting errors and estimating confidence are presented in [33]. Authors of [34] present how the error rate can be reduced by integrating confidence scores with the language understanding and dialogue modeling components of the system.

Research and improvement in the speech recognition technology will help us to generate more accurate captions for videos. In this case, the manual correction efforts could be minimized.

1.4.3 Caption Editors

There are quite a few caption editors available in the market. Subtitle Workshop [35] is a downloadable caption editor that provides an interface to view the video and type in captions. In this caption editor, the video needs to be manually started and stopped by the person editing the captions.

Very similar to Subtitle Workshop is Express Scribe [36], another downloadable caption editing software. In this software too, the editor needs to start and stop the audio for typing the captions. It has an interesting tool to slow down the speed of the audio to match the speed of your typing. The PlaySpeed tool in the ICS Caption Editor is inspired by this.

In CaptionMaker[23], one can import an existing transcript and edit it. The software can break the text into captions based on parameters such as number of characters per line, start a new caption for new sentence, etc. It has an interesting feature to Auto Timestamp the captions according to the audio. Jubler [22] is an open source subtitling software under the GNU public license. It is required that JRE be installed on the system to use it and the MPlayer be installed to view the subtitles etc. EZTooSoft Video Subtitle Editor [21] is another such subtitle editor. This requires manual start and stop of the video to determine the duration of the caption.

YouTube has its own caption editor where the owner of the video can edit the captions. It does not loop over the caption or does not slow down the speed of the video to help a person edit the captions. There are a number of commercial transcription services available. For example, 3PlayMedia [18]. They provide transcription and captioning services to a number of academic institutions for a fee. They have developed a workflow to provide captions efficiently. Academic institutions such as Georgia Tech and Penn State use third party transcription and captioning solutions [20] to provide captions to their students.

Stanford University has developed the Stanford Captioning System [37]. They let users upload their videos and download captions from their system to efficiently streamline the captioning process. They use the services of a third party transcription and captioning service to expedite the captioning process.

Camtasia Relay 3.0 offers an integrated caption editor [16]. Since this editor is not web-based, it cannot be used by multiple people at a time, and may require a lot of time and effort before the captions can be published.

IBM CCES [15] is an exceptional effort in the field of collaborative caption editing. Its workflow decomposes audio data of video into segments and distributes (for example 1 minute audio) to its registered editors. Since this tool is web-based, it can be accessed from remote locations. It loops over the audio so that the audio need not be manually started or stopped. Their system divides the audio into pieces using a low power point such as breath. It loops over the audio for its users.

Authors of [24] have developed a web-based caption editing tool ‘Synote’. ‘Synote’ stores edits of all users and uses a matching algorithm to see whether the users are in agreement. It uses a ‘number of users necessary to make an agreement’ parameter. Multiple users are required to edit the same caption. They also suggest incentives could be motivational for students to correct the captions for their classroom lectures.

The caption editor implemented in this thesis, utilizes the captions given by Google Speech Recognition via YouTube. The video slides can be referred to while correcting the captions. The video loops for the currently chosen sentence, allowing focus on correcting the caption rather than handling video player controls. Controls on the interface indicate status information such as which caption is corrected, which needs more work, or how many captions are undone. A group of people can simultaneously edit different sections of the caption file. As the editor is web-based, it can be used from anywhere no extra setup is required to use it.

Chapter 2 : Assessment of Current Speech Recognition

Tools

To accomplish our goal of captioning video lectures, we first decided to take the help of automatic speech recognition tools. We decided to assess the accuracy of 3 tools: Dragon Naturally Speaking Preferred 10 (DNS), Windows Speech Recognition (WSR) (packaged with Windows 7), and YouTube.

This study was conducted with 3 professors at University of Houston.

1. Dr. Shishir Shah, faculty in the Computer Science Department.
2. Dr. Olin Johnson, faculty in the Computer Science Department
3. Ms. Leigh Hollyer, faculty in the Mathematics Department.

This study had 3 phases – Dictation, Lecture Transcription, and Parroting. Dictation was done with DNS and WSR only, as YouTube does not support dictation. Lecture transcription was done with DNS and YouTube, as the version of WSR that supports transcription was not available.

Each participant enrolled on DNS and WSR to create their speaker-dependent acoustic model for each ASR system. “U.S.-accented English” and “general” vocabulary options were chosen while creating the profiles. The minimum training required by both the ASR systems was carried out.

For assessment of accuracy of the text, the ASR hypothesis (ASR output) is compared with the ground truth (what was actually said) using a PHP script. The working of the script is described below with an example:

ASR Output	Tricone one	naturally	seeking	engine	nice
Ground truth	Dragon	naturally	speaking	engine	

Error 1

Error 2

Error 3

$$\text{Accuracy \%} = (100 - (\text{Number of errors} / \text{Total number of words in ground truth})) * 100$$

2.1 Dictation

In this phase, a head-worn, noise-cancelling microphone by Cyber Acoustics (Cyber Acoustics AC201 Speech Recognition Stereo headset and Boom mic) was used. The participants were asked to read a 328 word paragraph from [41] to the WSR system. Along with normal text, this paragraph has numbers, acronyms, proper nouns, and some technical words. The dictated text was stored in MS Word. The dictation was also recorded using the Sound Recorder system on Windows 7. This recorded dictation was transcribed using DNS and stored in MS Word. The ASR output of both the tools was compared with the ground truth. Table 2-1 shows the results.

Table 2-1: Accuracy with WSR and DNS for Dictation

Participant	Accuracy with WSR	Accuracy with DNS
Dr. Shishir Shah	86.58%	89.93%
Dr. Olin Johnson	81.95%	83.13%
Ms. Leigh Hollyer	85.18%	83.03%
Average	84.57 %	85.36%

The results in the above table show that the accuracy of DNS and WSR is comparable. Their average accuracy is fairly good with dictation, which is a type of prepared speech where speakers read out some text. Speakers seem better able to enunciate the words while reading out some text.

2.2 Lecture Transcription

In this phase, the lecture audios were transcribed using DNS and YouTube. Camtasia was used to capture the videos. Dr. Shishir Shah and Dr. Johnson used the Sony Wireless WCS-999 microphone system to record the audio. Ms. Leigh Hollyer used the in-built laptop microphone to record the audio. 2 lecture videos delivered by each of the 3 participants were randomly chosen. The audio from these videos were extracted using a freely available video to MP3 converter. DNS was used to transcribe these audios. To assess the accuracy of YouTube, the lecture videos were uploaded to YouTube and the captions were downloaded for comparison. It should be noted here, that DNS is speaker-

dependent, whereas YouTube is speaker-independent. The transcription of the lecture audio was compared to the ground truth. Results are shown in the tables below.

Participant: Dr. Shishir Shah

Table 2-2: Accuracy of DNS and YouTube for lecture transcription of Dr. Shah

Lecture	Accuracy with DNS	Accuracy with YouTube
Lecture 1	38.01%	63.56%
Lecture 2	32.09%	79.23%
Average	35.05%	71.395%

The accuracy degraded significantly with DNS when we used it for transcribing lecture audio. The accuracy fell by 50% compared to the accuracy of dictation of Dr. Shah. The average accuracy with YouTube was much better than DNS.

Participant: Dr. Olin Johnson

Table 2-3: Accuracy of DNS and YouTube for lecture transcription of Dr. Johnson

Lecture	Accuracy with DNS	Accuracy with YouTube
Lecture 1	55.67%	67.72%
Lecture 2	49.23%	56.56%
Average	52.45%	62.14

The accuracy given by DNS fell by about 30% when compared to dictation accuracy. Here too, YouTube did better than DNS. The difference in the accuracy levels of DNS and YouTube are not as much as we had with Dr. Shah.

Participant: Ms. Leigh Hollyer

Table 2-4: Accuracy of DNS and YouTube for lecture transcription of Ms. Leigh

Lecture	Accuracy with DNS	Accuracy with YouTube
Lecture 1	70.66%	71.45%
Lecture 2	75.87%	70.16%
Average	73.265%	70.805%

The accuracy with DNS and YouTube fell by about 10% when compared to Ms. Leigh’s dictation. The differences in the accuracy of DNS and YouTube are not very significant. In this case, DNS did slightly better than YouTube.

2.3 Parrotting

Parrotting is the technique of repeating/imitating the words of another speaker. Parrotting can be used to enunciate the words to the speech recognition engine. DNS and WSR are “speaker-dependent” tools. These tools need to create a voice profile of the speaker who needs transcription. Hence transcription of an audio by many users or with multiple speakers becomes difficult, as having the acoustic model of each speaker becomes difficult.

In this phase of the study, the lecture audio was repeated (parroted) to DNS. The lecture audios used in the lecture transcription phase were used here. 30 minutes of Dr. Shah’s and Ms. Leigh’s lecture audio and 4 hours of Dr. Johnson’s lecture audio were

parroted. The parroted audio was done by the author of this thesis. The results are shown in Table 2-5.

Table 2-5: Results of parroted lecture audio

Professor of whose lecture was parroted	Accuracy
Dr. Shishir Shah	94.76%
Dr. Olin Johnson	96.63%
Ms. Leigh Hollyer	96.1%
Average	95.83%

It can be seen from the results that the accuracy is very good when the parroted technique is used as the lecture becomes prepared speech when it is parroted. The accuracy is consistent for all three speakers, as it is one person repeating all the lectures. The output of this method is very good to be used for manual correction.

2.4 Analysis of Errors

The transcription given by DNS was analyzed to assess the nature of errors introduced in the transcription. Lecture transcriptions of two professors, Dr. Shishir Shah, and Dr. Olin Johnson were analyzed. Transcriptions of 20 minutes of Dr. Shah's lecture and 10 minutes of Dr. Johnson's lecture were analyzed. Results are shown in Table 2-6.

Table 2-6: Analysis of errors of recognition by DNS

Category	Nature of Error in DNS	Dr. Shah	Dr. Johnson	Total	%	Total % per category
Tool's weakness	Incorrect interpretation by the tool	113	39	152	51.01	51.01
Speaker's weakness	Disfluent speech	21	10	31	10.40	40.93
	Heavily accented speech	7	11	18	6.04	
	Conversational speech	18	12	30	10.07	
	Mixed words/ not enunciated well	25	17	42	14.09	
	Very low volume or moving away from microphone		1	1	0.34	
Independent	out- of- vocabulary words	1	2	3	1.01	9.73
	Similar sounding words	1	1	2	0.67	
	Ungrammatical construct (because of technical words)	19	0	19	6.38	
	Inaudible (student interaction)	5		5	1.68	
	Total	205	93	298		

The errors mentioned in the Table 2-6 are explained below:

1 .Incorrect interpretation by tool: In this, the tool interpreted the words completely wrong, though the words were pronounced correctly by the speaker and no other factors affected the audio. This is considered as the tool's weakness to identify the correct words.

2. Disfluent speech: Speech disfluencies are any of various breaks, irregularities, or non-lexical vocables that occur within the flow of otherwise fluent speech. These include false starts, i.e. words and sentences that are cut off mid-utterance, phrases that are restarted or repeated and repeated syllables, fillers i.e. grunts or non-lexical utterances such as "uh", "erm" and "well", and repaired utterances, i.e., instances of speakers correcting their own slips of the tongue or mispronunciations [44].

E.g.: objects that you have..or...gives you the..uh...uh...give you... give you...uh...everything.. All the regions...

3. Heavily accented speech: This kind of speech is when different speakers stress on different words in a different way. This kind of speech is especially influenced by the country or state of the speaker.

4. Conversational speech: This is interactive or spontaneous communication between two or more people. For example:

we will look at the textbook in a minute..okay?

otherwise your result is a zero..right?

let's say they were grayscale values right so it's an 8-bit value... sorry.. mode?
Yes.. Close.

5. Mixed words together/ Not enunciated well: In this, the words are spoken too fast and appear to be mixed up and not enunciated well enough for the tool to understand.

6. Very low volume: The speaker's volume is very low for the tool to be able to identify the spoken words.

7. Out-of-vocabulary words: Some technical words or acronyms that are not in tool's vocabulary cannot be identified by the tool. These errors may be eliminated by adding the word in tool's vocabulary.

8. Similar sounding words: These are the words in the language that sound very similar. For example, to, too, and two; see and sea. These words are identified by the tool based on context, but if there are errors before and after these words, the context itself can be identified wrongly and can result in these words being interpreted incorrectly.

9. Ungrammatical constructs: The errors observed as ungrammatical for these lectures were due to technical words. For example:

to perform AND operation

combination of NOT AND and OR

nothing but AND's NOT's and OR's

10. Inaudible (Student interaction): The audio when the students ask questions or are answering some question has very low volume as the current classroom infrastructure does not capture student speech clearly.

2.5 Conclusions

We performed a three-phase study to assess the general accuracy of the speech recognition systems.

In the dictation phase, the accuracy of the tools was observed to be fairly good. The accuracy of DNS and WSR is very much comparable. Accuracy rates are better with isolated speech pattern such as dictation (reading out a document), as the words are clearly enunciated to the speech recognition tool.

For the lecture transcription, though the audio quality of the lectures is good (free from static or very minimal noise), various factors mentioned in section 2.4 degrade the accuracy provided by ASR tools significantly. Spontaneous speech makes the text given by ASR tools erroneous due to factors inherent in a lecture audio, like repetitions, false starts, corrections, or filler words such as “Um” , “Ah” ,”hmm”. The output of the speech recognition tool depends on the “context” – the words appearing before and after. For example a sentence like

I eat. I scream

Since the tool would not know there is a period after “I eat” in the conversation, the output text can be

I eat ice cream.

Since there is no explicit information on punctuations, or sentence start/end in spontaneous speech which is important to understand the context, errors may be introduced in the text.

In the case of out-of-vocabulary words or acronyms, which are possibly used in a technical lecture, the ASR tool will still make a guess based on the words it has in the vocabulary. That affects the context and can result in the whole string being misrecognized.

The analysis in section 2.4 shows 51% of the errors are due to the tool's weakness to identify the correct words, even though there are no other factors affecting the audio. About 41% of the errors are speaker dependent, and these types of errors can be reduced if the speech pattern is modified. Some words were observed to be repeatedly misrecognized even though they were in the tool's vocabulary. For example: Boolean algebra, blob, etc. These errors may be reduced by more training by individual speakers. Some of the errors are independent of the tool or the speaker such as homonyms and ungrammatical constructs introduced by technical terms.

The accuracy level given by YouTube is observed to be better than DNS for two speakers. Lectures audios of Leigh Hollyer were a special case. They were "prepared" lectures. These lectures were not delivered in a class. She had prepared a transcript of the lecture and read it to the Camtasia recorder. Hence, the accuracy level given by DNS for her lecture is better than that is given for other two professors. However, the average accuracy level given by YouTube is less than given by DNS for her lecture audio.

Based on the results, we concluded that transcription given for lecture audio by ASR cannot be directly used, as the accuracy level is too low. In this scenario, a technique such as parrotting can be used, as the accuracy levels by parrotting are good. However, this would require one person training the ASR tool and repeating the entire

lecture. This would be somewhat impractical when the number of lectures and courses increase that need captioning.

Chapter 3 : ICS Caption Editor

The conclusion of Chapter 2 states that the output given by the ASR cannot be used as is. Manual correction of the captions is required to be able to deliver correct captions. To assist in the task of editing the captions given by the speech recognition tool, we designed, implemented, and evaluated a caption editing tool. The ICS Caption Editor is a custom built web-based crowd-sourcing tool to help in editing the captions efficiently. This chapter discusses the design and implementation of the ICS Caption Editor.

3.1 Usability Requirements

The usability requirements we considered when we designed the editor interface are listed below. Usage studies were conducted in Fall 2012 and Spring 2013, and users were asked to tell us about their experience. User feedback was also considered to make design decisions.

Simple Navigation and Ease of Use: The tool should be easy to use. The interface must be simple to navigate. The controls on the interface must be intuitive.

Enable editing captions from anywhere: The interface needs to be web-based to make it convenient to use the tool from anywhere.

Security: The tool should provide user management (use of usernames and passwords).

Avoid high workload: The tool must enable crowd-sourcing to avoid high workload on users.

Help and tutorial: Appropriate HELP section and tutorials must be available for users for usage instructions.

Hover help (tooltips): Interface must have tooltips to provide hints for the interface elements.

3.2 Design

In this section we discuss the design of the ICS Caption Editor. In the design, we have considered the usability requirements discussed in the previous section. Users will be able to access the editor interface through the web. With this editor the users will be able to see the caption text given by the speech recognition tool. They will be able to listen to the audio and correct the captions given by the speech recognition tool. Users will be able to refer to the video slides while editing. After the caption text is edited, the users will be able to save the caption text. The caption file is organized into sections of 5 sentences each. Each section is an individually editable part. Different users will be able to edit different sections simultaneously. Users will also be able to indicate and view the status of the caption text, whether it is complete or needs some more work. Details of the design are discussed further.

There were multiple iterations of development for the ICS Caption Editor. Figure 3-2 shows the design of the final version of the editor. The ICS Caption Editor (including a previous version of the editor) has been in use for editing captions since Fall 2012.

To begin using the ICS Caption Editor, users should click on the link that has been provided to them. Users will be redirected to a login screen shown in Figure 3-1.

New users can register their username and password using this screen. Once registered, users can login using the login screen. Users can also change their password using the “Change Password” button. After the user is logged in, the user will see the Editor Interface.

ICS Caption Editor [Help](#)

Username *

Password *

[Change Password](#) [Log In](#)

[Not a Registered User? Sign up](#)

Figure 3-1: ICS Caption Editor login screen

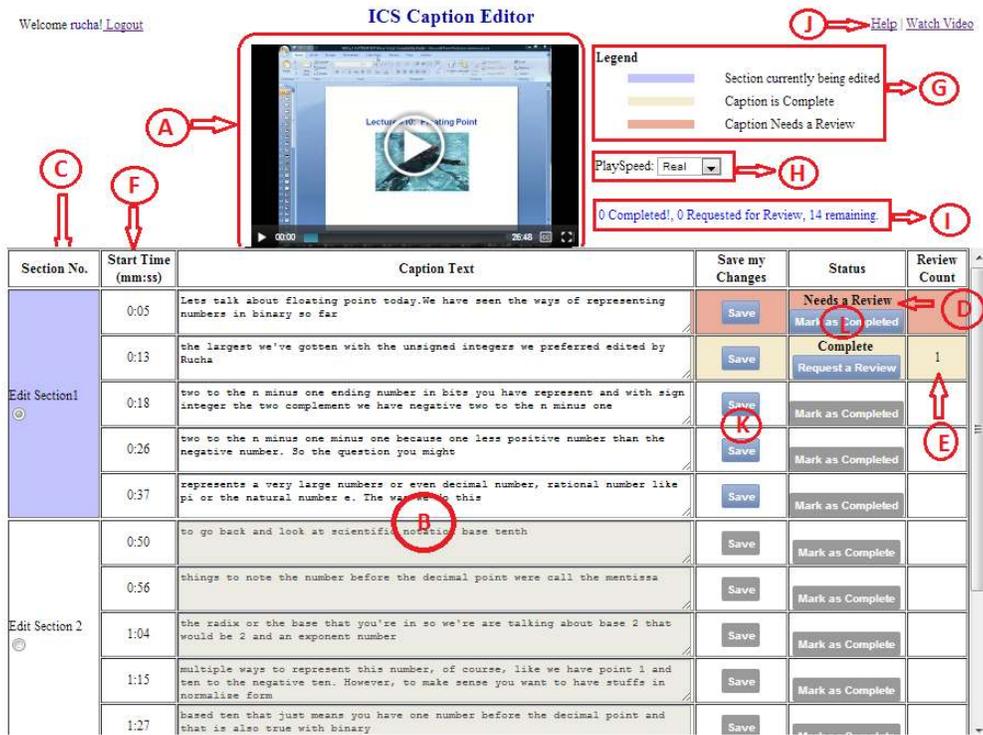


Figure 3-2: ICS Caption Editor interface

The ICS Caption Editor design is explained below:

A : This is the video that is to be captioned. It has an option to display captions by clicking the “CC” button. This video has the Play and Pause control. The video can also be viewed in full screen using the Fullscreen control.

B: This is the caption text given by the speech recognition tool.

C: Caption text is divided into sections of 5 sentences each. The last section of the caption file may have less than 5 sentences. A section is an individually editable part of the caption file. These way different sections can be edited by different users simultaneously.

D: Status of the caption text. Explanation of the caption statuses is given below:

(a) Needs a Review: This caption text is checked for correction, but either the editor is unable to hear the audio/speech clearly or the editor is not sure of the accuracy of the corrected caption. In this case, the editor would like another editor to take a look at the caption text to correct it.

(b) Complete: The caption text is checked for correction and is approved by the editor.

E: This displays the review count of the caption text. It is the number of times this caption text has been approved by editors.

F: This displays the start time when the sentence is spoken in the video.

G: Legend displays the color coding used in the editor for the different statuses of the caption text and the section.

H: PlaySpeed tool can be used to adjust the speed of the audio.

I: This section displays how many captions have status as Complete, Needs a Review, and how many captions still needs to be worked upon.

J: This link opens up a HELP window that has instructions on how to use the editor.

K: The Save button saves the edited text to a file on the server and changes the status of the caption to 'Needs a Review'

L: Mark as Completed button changes the status of the caption text to 'Complete' and increments the review count.

The users can logout using the Logout link in the upper left corner. The users are automatically logged out after 10 minutes of inactivity or when the browser is closed. The process of logging out will unlock the user's locked sections.

Welcome rucha! [Logout](#)

ICS Caption Editor

[Help](#) | [Watch Video](#)



Legend

- Section currently being edited
- Caption is Complete
- Caption Needs a Review

PlaySpeed: Real

0 Completed!, 0 Requested for Review, 394 remaining.

Section No.	Start Time (mm:ss)	Caption Text	Save my Changes	Status	Review Count
Edit Section 1	0:00		<input type="button" value="Save"/>	<input type="button" value="Mark as Complete"/>	
	0:10		<input type="button" value="Save"/>	<input type="button" value="Mark as Complete"/>	
	0:20		<input type="button" value="Save"/>	<input type="button" value="Mark as Complete"/>	
	0:30		<input type="button" value="Save"/>	<input type="button" value="Mark as Complete"/>	
	0:40		<input type="button" value="Save"/>	<input type="button" value="Mark as Complete"/>	

Figure 3-3: ICS Caption Editor interface when no text from ASR is available

As shown in Figure 3-3, the ICS Caption Editor can also be used if the caption text from automatic speech recognition tool is not available. The captions text boxes can be split at a distance of for example 10 seconds. The editor can listen to the audio and type the text inside each text box.

ICS Caption Editor

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 3-4: ICS Caption Editor interface with audio converted as video

The Caption Editor can also be used if only audio is available. The audio can be converted to video using programs like Windows Movie Maker and then the process for a regular video is followed.

How to Use:

To edit a section, click on the “Edit Section” button. The text boxes with the caption text will be enabled. To edit a sentence, double click inside the textbox. The video will start playing that portion of the video in a loop. Edit the caption inside the textbox, if it needs to be corrected and save the caption text by clicking on the “Save” button.

After the caption text is saved, the status of the caption changes to “Needs a Review”. If you are unable to hear the speech clearly or are unsure of the accuracy of your changes to the caption, you may leave the status of the caption as “Needs a Review” and move on to another sentence. If you approve the changes to the caption, click on “Mark as Complete”. This will change the status to Complete.

To edit another sentence, select a sentence that has a blank status or a sentence that has the status “Needs a Review”. After you have finished editing a section, you may move on to another section. Try to find the sections that are marked in white. A section that is marked blue indicates that the section is currently unavailable for editing.

You can use the PlaySpeed tool to adjust the audio speed. The Legend displayed next to the video, indicates the color coding used for different statuses of the caption text and the section. You can also refer to these instructions by clicking on “HELP” on the upper right corner. After you have finished editing, you may logout from the system.

These captions and an interactive transcript are made available via the ICS VIDEO PLAYER.

3.3 Implementation

In this section, we will discuss the implementation details of the ICS Caption Editor. We will describe how the system works from uploading a video to making the captions available.

Eclipse Java EE IDE for Web Developers: Version: Indigo Service Release 2 is used to develop Java program. This Java program interacts with YouTube using Google API Java clients 1.10.3. Tokens are used to establish authenticity and sessions.

The process is as follows:

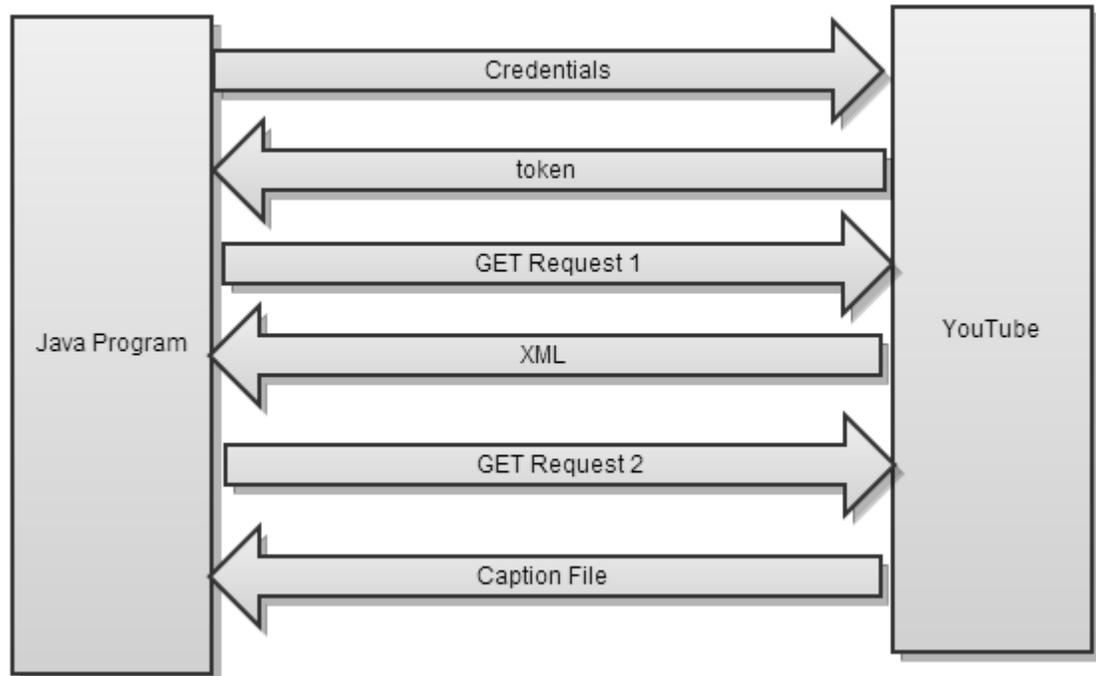


Figure 3-5: Communication between Java program and YouTube to get the caption file

The Java Program authenticates itself using the credentials and retrieves an authorization token from YouTube. The Java program then sends a HTTP GET request to YouTube. The token and the video ID whose captions we want to retrieve are passed.

URL for GET request:

```
"https://gdata.youtube.com/feeds/api/videos/"+videoID+"/captions";
```

YouTube returns an XML in response. An example of the part of the XML that is useful to retrieve the captions is given below: The important tags are boldfaced.

<entry gd:etag="W/"CUMDRnk5eyp7I2A9WhBQFEw.""/>
<id>tag:youtube.com,2008:captions:Ch8LEO3ZhwUaFgjT3Zbptd3V3tJlbhoAIgNhc3IM</id>
<published>2013-03-15T22:57:57.723-07:00</published>
<updated>2013-03-15T22:57:57.723-07:00</updated>
<app:edited>2013-03-15T22:57:57.723-07:00</app:edited>
<category scheme="http://schemas.google.com/g/2005#kind" term="http://gdata.youtube.com/schemas/2007#captionTrack"/>
<title/>
<content type="application/vnd.youtube.timedtext"
src="https://gdata.youtube.com/feeds/api/videos/371W610lrtM/captiondata/Ch8LEO3ZhwUaFgjT3Zbptd3V3t8BEgJlbhoAIgNhc3IM" xml:lang="en"/>
<link rel="self" type="application/atom+xml"href="https://gdata.youtube.com/feeds/api/videos/371W610lrtM/captions/Ch8LEO3ZhwUaFgjT3Zbptd3V3t8BEgJlbhoAIgNhc3IM"/>
<link rel="edit" type="application/atom+xml"href="https://gdata.youtube.com/feeds/api/videos/371W610lrtM/captions/Ch8LEO3ZhwUaFgjT3Zbptd3V3t8BEgJlbhoAIgNhc3IM"/>
<link rel="edit-media" type="application/vnd.youtube.timedtext"href="https://gdata.youtube.com/feeds/api/videos/371W610lrtM/captiondata/Ch8LEO3ZhwUaFgjT3Zbptd3V3t8BEgJlbhoAIgNhc3IM"/>

```
<yt:derived>speechRecognition</yt:derived>
</entry>
```

The Java program parses the XML and retrieves the URL given by the “src” attribute of the <content> tag. The <yt:derived> tag in the parent <entry> tag must have the value as “**speechRecognition**”. This would mean that the captions available at URL given by the “src” attribute of <content> tag are automatically generated using speech recognition technology.

The Java program then sends another HTTP GET request to the URL given by the “src” attribute retrieved in the previous step.

Example URL retrieved : We call it captionTrackSrc.

captionTrackSrc:

<https://gdata.youtube.com/feeds/api/videos/371W610IrtM/captiondata/Ch8LEO3ZhwUaFgjT3Zbptd3V3t8BEgJlbhoAIgNhc3IM>

URL used for GET request:

```
captionstrack = captionTrackSrc+"?fmt=srt";
```

Here the fmt parameters denotes the format in which we wish to retrieve the captions. The value “srt” means SubRip tex format. The format of this caption file is given below:

```
Subtitle number
Start time --> End time
Text of subtitle (one or more lines)
Blank line
```

The start and the end time format is: hh:mm:ss,msec.

Example .srt file:

1
00:00:04,640 --> 00:00:11,169
and uh so the quiz will start again at 6:15.

2
00:00:11,169 --> 00:00:17,730
and uh so it'll be a 30 minute quiz and it

3
00:00:17,730 --> 00:00:25,059
is very much like the other one except that

After the caption file is retrieved, the ICS Caption Editor program can read the caption file and display the captions in the interface. The ICS Caption Editor also saves and retrieves some user login/logout information, the statuses of the captions, caption text information, logging information, etc.

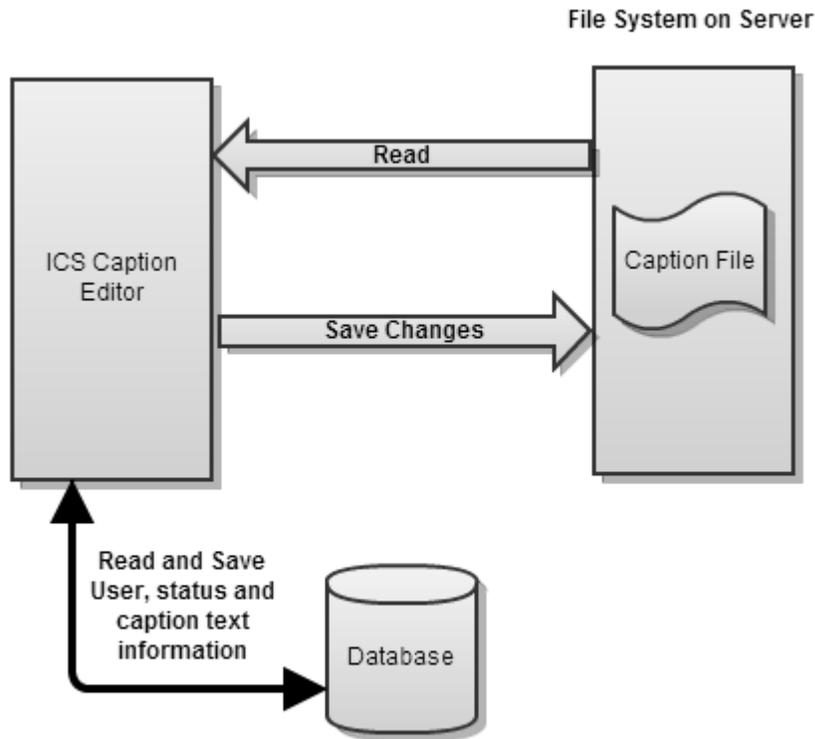


Figure 3-6: Communication between the ICS Caption Editor, Server, and database

The Caption Editor is built using PHP Version 5.2.4. The Apache Tomcat 2.2 web server is used. The video is displayed using HTML5 elements and the mediaelementplayer.js [43]. MySQL 5.2 is used in the backend for user management, to store editor related data such as statuses of captions, locked sections etc. Database updates are done in the background using AJAX. JQuery (version: jquery-1.4.1) JavaScript library is used to traverse the HTML document and to make AJAX requests.

Chapter 4 : Evaluation of the Caption Editor

The evaluation phase helped us in knowing how the caption editor works with the classroom lecture videos. It also helped us to know which features worked well and which ones require some modifications.

4.1 Study and Results

A study was conducted with videos of classroom lectures being edited by students enrolled in that class. We involved a total of 3 videos from the following 2 classes:

1. COSC1300: Introduction to Computing
Professor: Dr. Olin Johnson
2. COSC2410: Computer Organization and Programming
Professor: Dr. Nouhad Rizk

For this study, students from both the classes were invited to participate. ICS Caption Editor links for the videos were provided to participating students. A how-to-use video of the ICS Caption Editor was provided to them to understand the steps of editing. The students created their own user ID and password to login to the caption editor. Students were divided into 3 groups and given 3 separate links of the 3 videos which were to be captioned. Every group was given 5 days to complete the editing. Following are the results of captioning individual lecture:

COSC1300 – Lecture 23

Study period: Wednesday, March 20, 2013 to Sunday, March 24, 2013

Duration of the lecture = 1:10:45 (hh:mm:ss)

Number of Sentences to caption= 352

Number of Sections = 71 sections

Number of participants = 11

Performance of each participant:

Table 4-1: COSC1300: Lecture 23: Performance of individual participant

Users	Time Taken in minutes	Captions Saved	Captions Completed
User1	10.06	10	9
User2	16.38	20	25
User3	25.35	20	33
User4	38.5	35	42
User5	41.41	35	35
User6	45.21	42	50
User7	56.12	32	45
User8	60.66	57	54
User9	64.13	50	55
User10	69.16	54	52
User11	76.11	31	0

Median value: 45.21 minutes

Accuracy of corrections:

8 words - incorrect

3 words - missing

Total words: 3536 words

Progress of the work:

Table 4-2: COSC1300: Lecture 23: Progress of work

Day	Day of week	Sentences Completed	% Work Done
1	Wednesday	0	0
2	Thursday	26	7.38
3	Friday	155	44.03
4	Saturday	134	38.06
5	Sunday	37	10.51

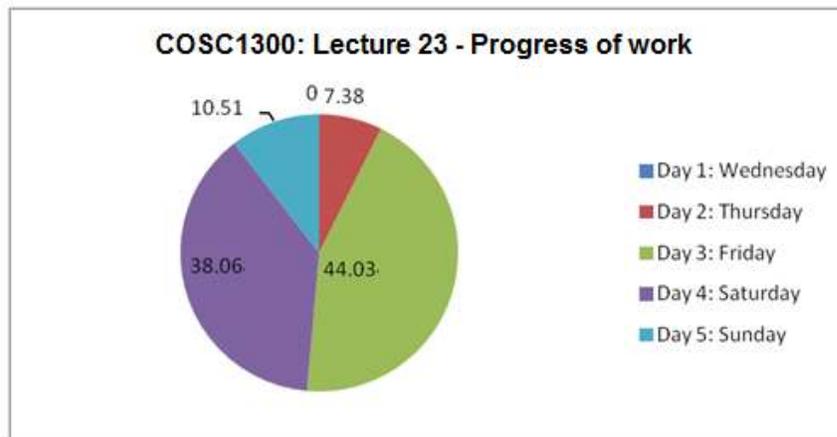


Chart 4-1: COSC1300: Lecture 23: Progress of work

As shown in Table 4-1, 11 participants worked on the Lecture 23 of COSC1300 class. Table 4-1 shows the time and the performance of individual participant. We consider the median value as a measure for the time taken to edit the complete lecture. From the results, it can be seen that a lecture of approximately 1 hour and 10 minutes was corrected in about 46 minutes which is about 0.6 times the time of the original lecture. The corrected lecture was analyzed for errors. Eight words were found to be incorrectly edited and 3 words were missing in the final correction. Table 4-2 and Chart 4-1 show the

progress of work for this lecture. They show that most of the work was done on Day 3 and Day 4 of the study.

COSC2410 – Lecture 11

Study period: Friday, March 22, 2013 to Tuesday, March 26, 2013.

Duration of the lecture = 0:31:00 (hh:mm:ss)

Number of Sentences to caption = 314

Number of Sections = 63 sections

Number of participants=10

Performance of each participant:

Table 4-3: COSC2410: Lecture 11: Performance of individual participant

Users	Time Taken in minutes	Captions Saved	Captions Completed
User1	3.48	1	1
User2	10.56	10	10
User3	15.22	12	11
User4	21.06	15	15
User5	34.5	33	33
User6	35.2	30	36
User7	47.94	45	50
User8	51.24	47	52
User9	62.38	60	60
User10	68.76	65	63

Median value = 34.85

Accuracy of corrections:

6 words: incorrect
1 word: typo
Total words: 3576 words

Progress of the work:

Table 4-4: COSC2410: Lecture 11: Progress of work

Day	Day of week	Sentences Completed	% Work Done
1	Friday	0	0
2	Saturday	5	1.59
3	Sunday	146	46.49
4	Monday	106	33.75
5	Tuesday	57	18.15

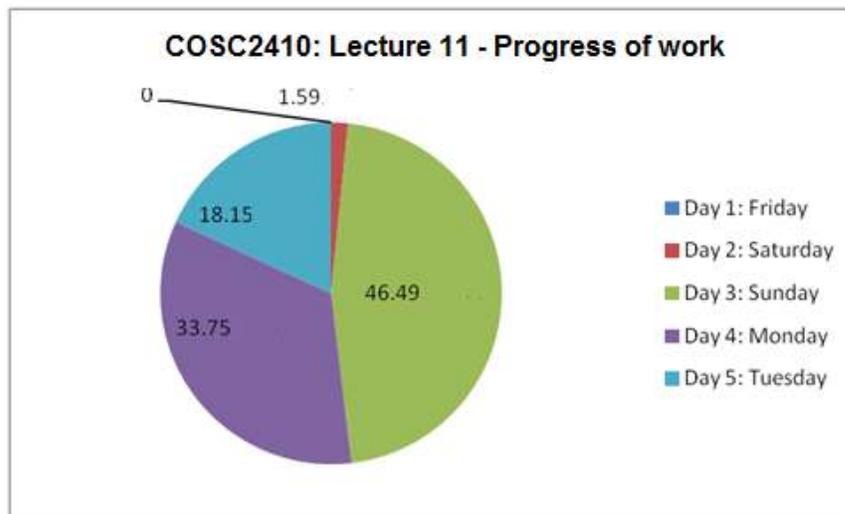


Chart 4-2: COSC2410: Lecture 11: Progress of work

As shown in Table 4-3, 10 participants worked on the Lecture 11 of COSC2410 class. Table 4-3 shows the time and the performance of individual participant. From the results, it can be seen that a lecture of approximately 31 minutes was corrected in about 35 minutes which is about 1.12 times the times the time of the original lecture. The corrected lecture was analyzed for errors. Six words were observed to be incorrect and 1 spelling mistake was observed. Table 4-4 and Chart 4-2 show the progress of work for this lecture. They show that most of the work was done on Day 3 and Day 4 of the study.

COSC2410 – Lecture 13

Study period: Friday, March 22, 2013 to Tuesday, March 26, 2013.

Duration of the lecture: 0:15:13 (hh:mm:ss)

Number of Sentences = 163

Number of Sections = 33 sections

Number of participants = 7

Performance of each participant:

Table 4-5: COSC2410: Lecture 13: Performance of individual participant

Users	Time Taken in minutes	Captions Saved	Captions Completed
User1	1.2	1	0
User2	16.32	15	15
User3	22.3	20	16
User4	26.45	20	19
User5	39.34	30	33
User6	45.15	35	40
User7	46.06	45	52

Median value=26.45

Accuracy of corrections:

2 words: incorrect
1 word: typo
1 word: missing
Total words: 1899 words

Progress of the work:

Table 4-6: COSC2410: Lecture 13: Progress of work

Day	Day of week	Sentences Completed	% Work Done
1	Friday	0	0
2	Saturday	0	0
3	Sunday	75	46.01
4	Monday	53	32.51
5	Tuesday	35	21.47

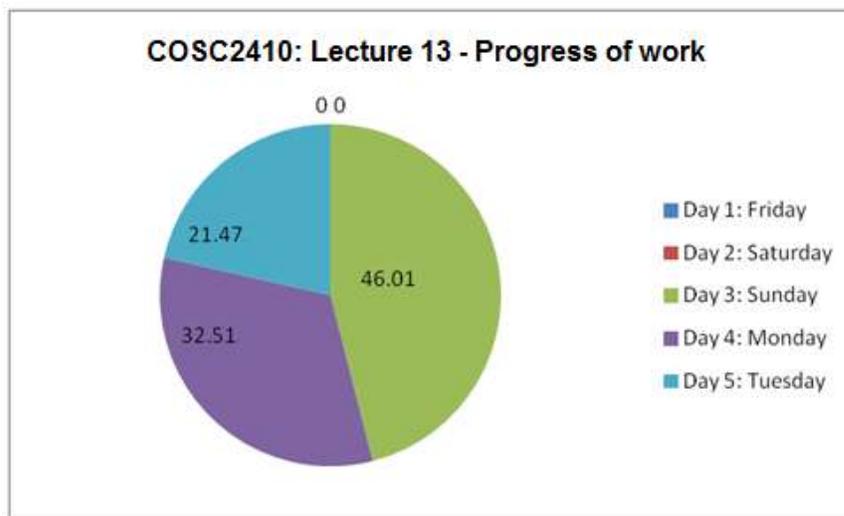


Chart 4-3: COSC2410: Lecture 13: Progress of work

As shown in Table 4-5, 7 participants worked on the Lecture 13 of COSC2410 class. Table 4-5 shows the time and the performance of individual participant. From the

results, it can be seen that a lecture of approximately 15 minutes was corrected in about 27 minutes which is about 1.73 times the time of the original lecture. The corrected lecture was analyzed for errors. It was observed that 2 words were incorrect, 1 word was misspelled, and 1 word was missing from the final correction. Table 4-6 and Chart 4-3 show the progress of work for this lecture. They show that work was distributed across Day 3, 4, and 5 of the study.

4.1.1 Conclusions

Following the usability requirements mentioned in Section 3.1, we were able to design and implement the Caption Editor. We conducted the study with 28 students from 2 classes and the evaluated the results illustrated in section 4.1. From the study and the results, we can see that students were able to use the editor successfully to edit the captions. From the participants some did more work than the others. A small number of errors were observed in the final corrected version. For all the 3 lectures, most of the work was done on Day 3 and Day 4 of the study.

4.2 Survey Results

At the end of the study, an online survey was conducted to know about the usability of the caption editor. The survey had 20 questions. The analysis of the results of the survey is presented below:

Demographic information:

Question: Specify your academic year:

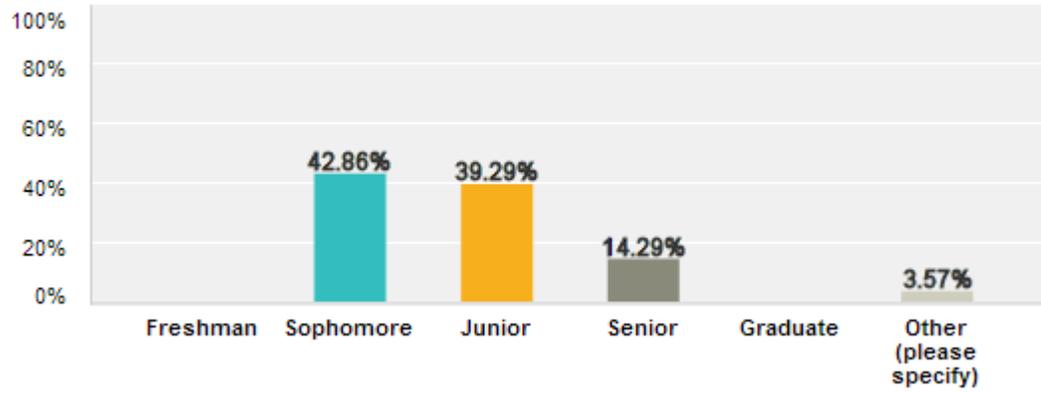


Chart 4-4: Academic Year

Question: Specify Gender:

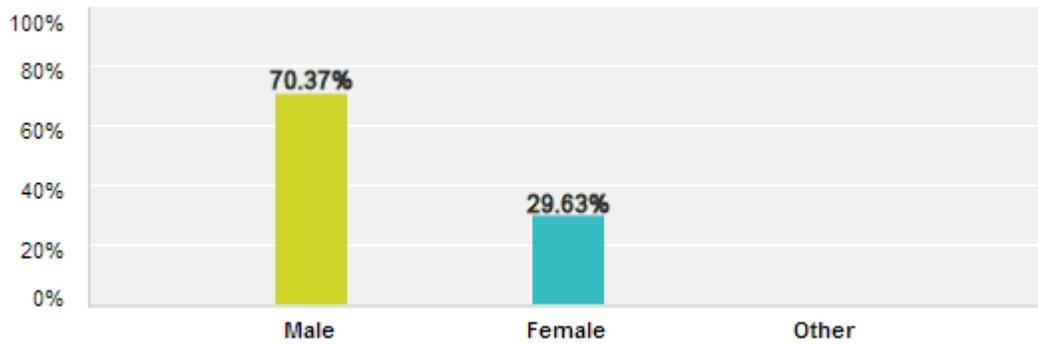


Chart 4-5: Gender

Question: Please indicate your Ethnicity: Check all that apply.

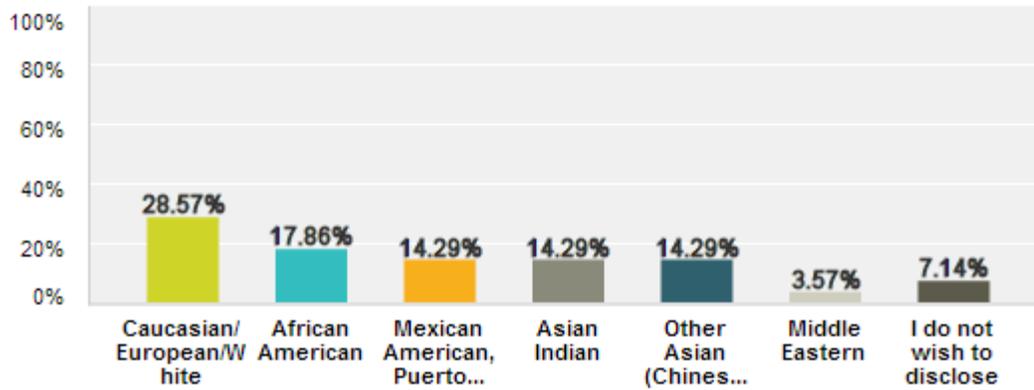


Chart 4-6: Ethnicity

Question: What is your age?

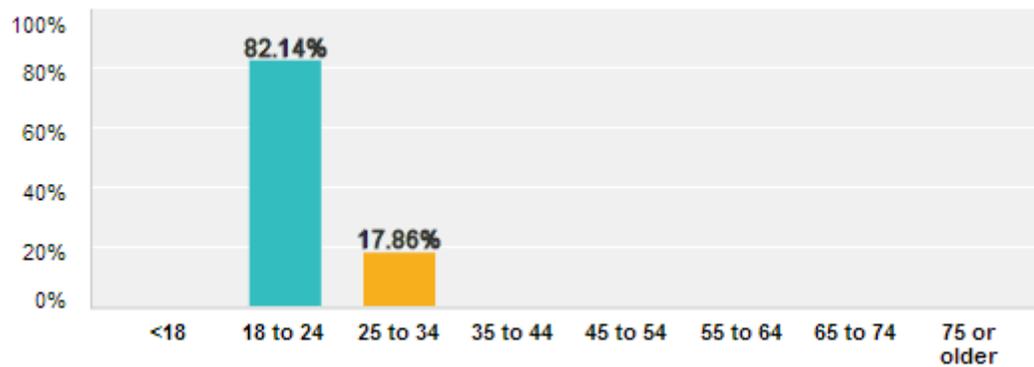


Chart 4-7: Age

Question: What is your major?

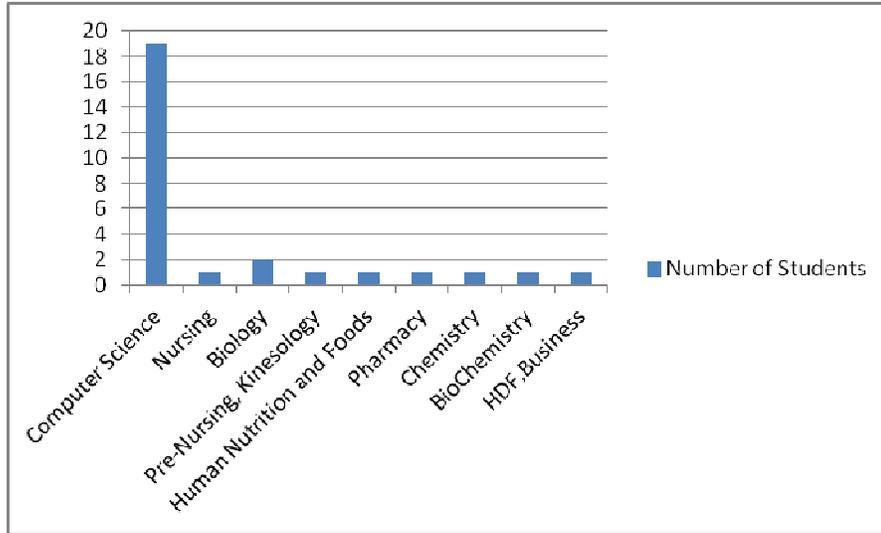


Chart 4-8: Major

Question: How would you describe your fluency with the English language?

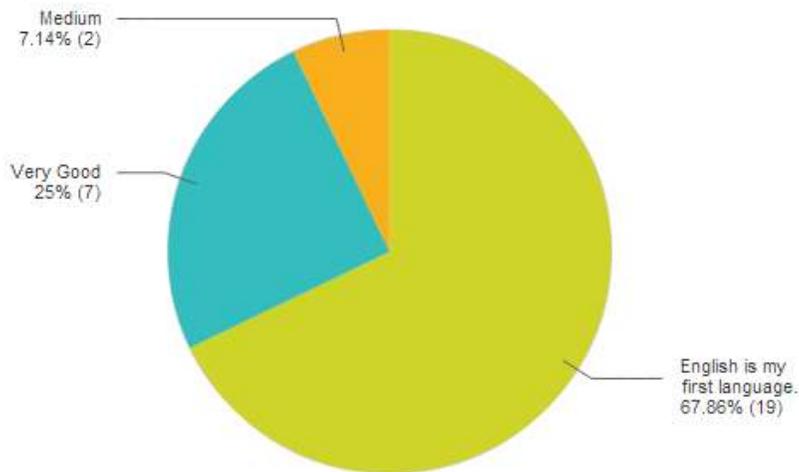


Chart 4-9: Fluency with English Language

ICS Caption Editor Feedback:

Question: Specify the course of the lecture you corrected captions for:

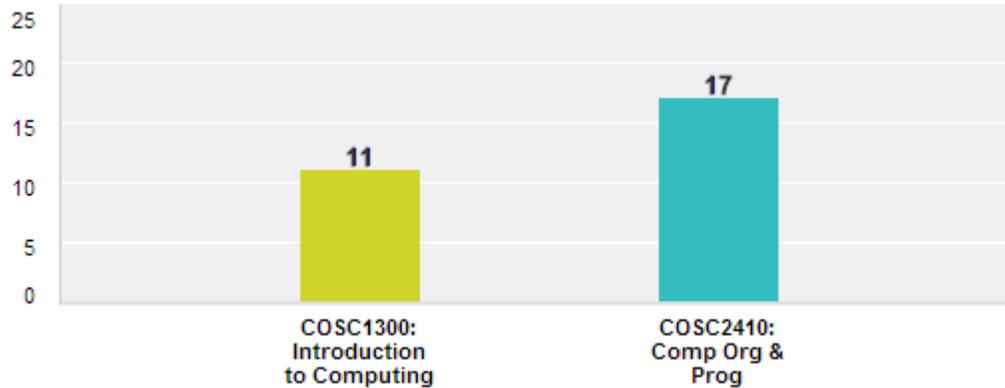


Chart 4-10: Course of the lecture captioned

Question: The ICS Caption Editor is easy to use. Please express the strength of your agreement.

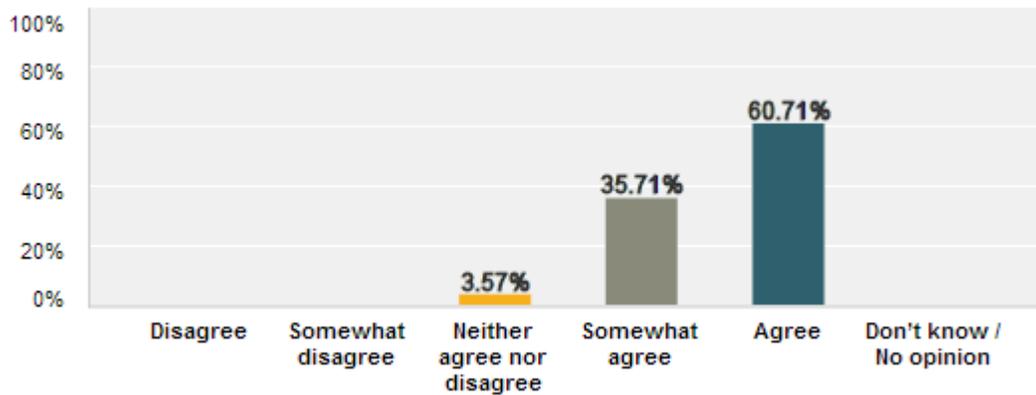


Chart 4-11: Ease of use

It can be observed that students found the editor easy to use. Some students provided comments that they found the interface simple to use and liked the video looping concept.

Question: The slide images in the video (as opposed to having ONLY audio) were helpful in editing the captions. Please express the strength of your agreement.

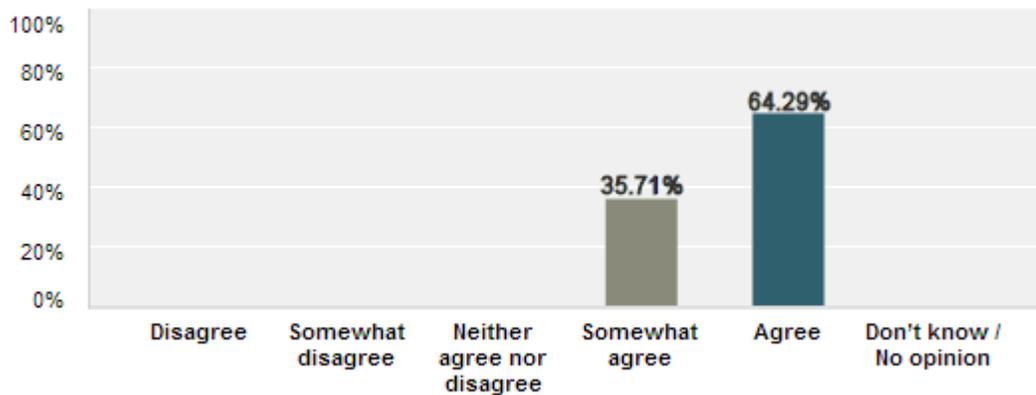


Chart 4-12: Use of slide images

Students referred to the slide images while editing and claim that it was useful to understand what was being said, when the audio was difficult to understand.

Question: The technical quality of the audio was good. Please express the strength of your agreement. (Some things that negatively affect the technical quality of audio are static, echo, electrical noise, distortions etc.)

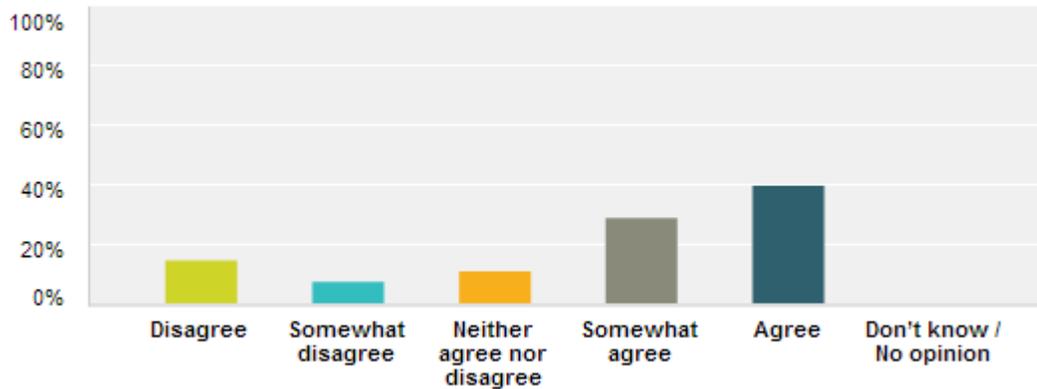


Chart 4-13: Technical quality of the audio

FILTER: COSC1300:

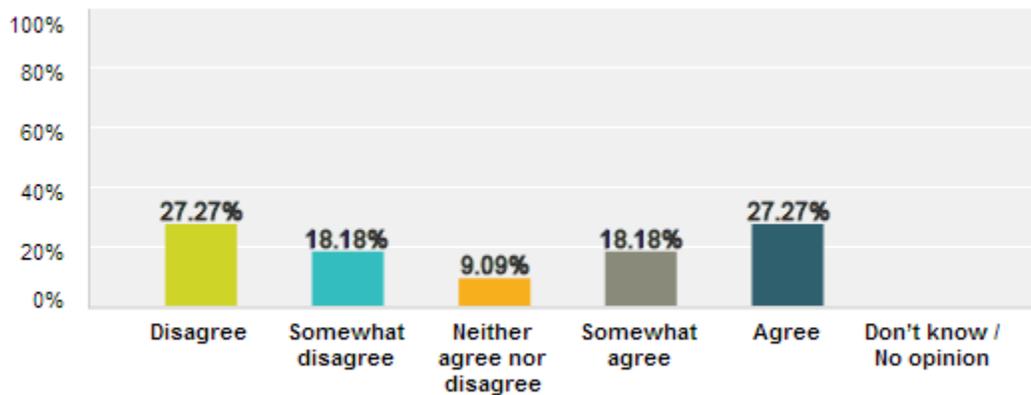


Chart 4-14: Technical quality of the audio for COSC1300 course

FILTER: COSC2410:

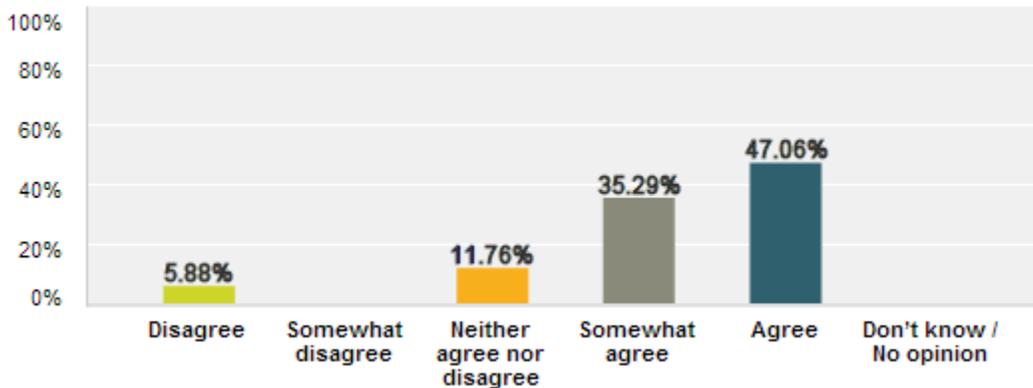


Chart 4-15: Technical quality of the audio for COSC2410 course

Question: You could hear the professor clearly. Please express the strength of your agreement. (Some things that could affect the clarity in hearing are the volume of his speech, heavy accent etc)

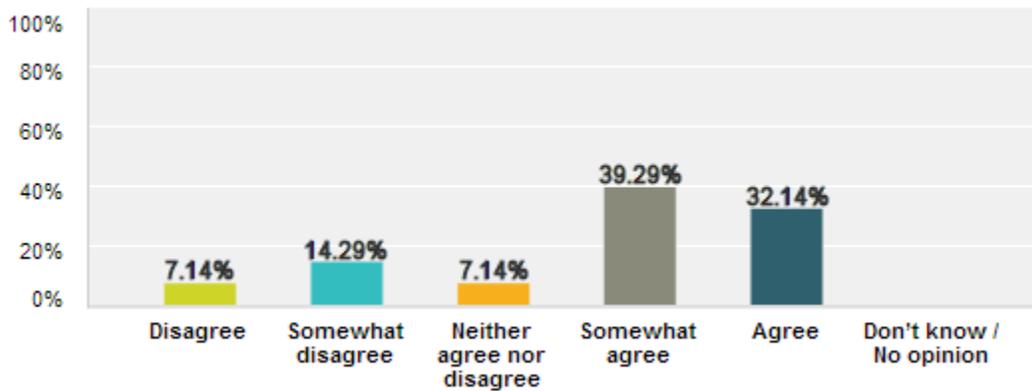


Chart 4-16: Clarity of professor's speech

FILTER: COSC1300:

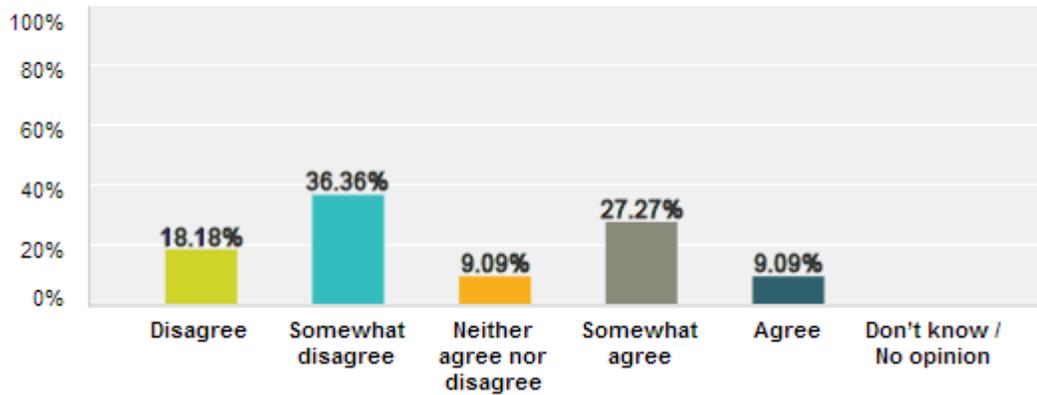


Chart 4-17: Clarity of professor's speech for course COSC1300

FILTER: COSC2410:

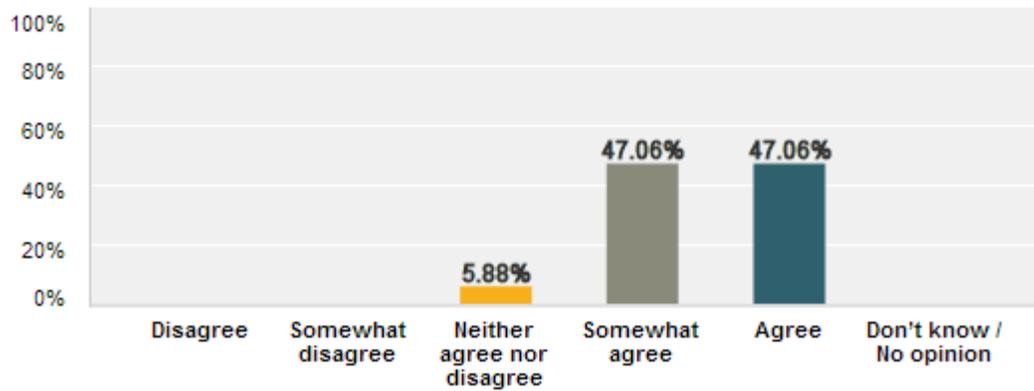


Chart 4-18: Clarity of professor's speech for course COSC2410

The results also show that the technical quality of the audio was better for COSC2410 lectures than COSC1300 lecture. Also, they could hear the lectures of COSC2410 more clearly than COSC1300.

Question: The PlaySpeed tool is useful. Please express the strength of your agreement. (The PlaySpeed tool can be used to adjust the speed of the audio)

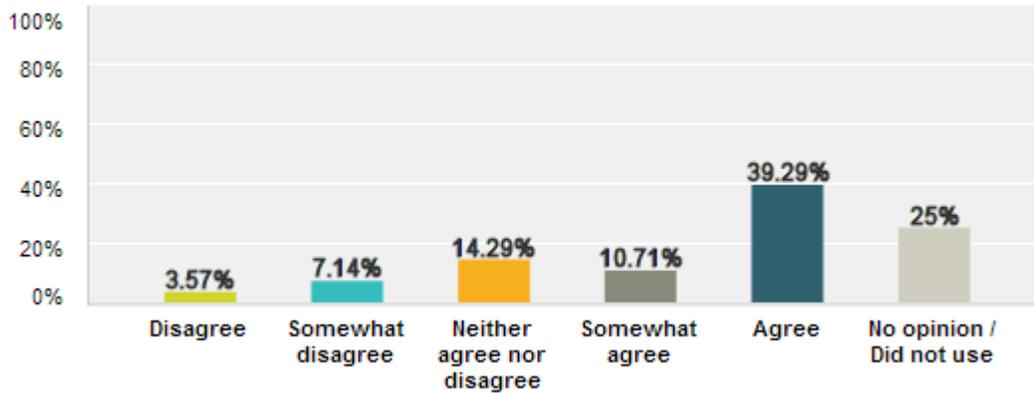


Chart 4-19: Use of PlaySpeed tool

There was a distributed response for the usability of PlaySpeed tool. Some students did not use it and some reported technical issues while using it. Some students also reported it being useful to slow down the speed to be able to hear the speaker.

Question: The feature of having the status of the caption as "Needs a Review" is useful. Please express the strength of your agreement. (If you are unable to hear the audio clearly or are unsure of the accuracy of your correction, there is an option to have the status of the caption as "Needs a Review".)

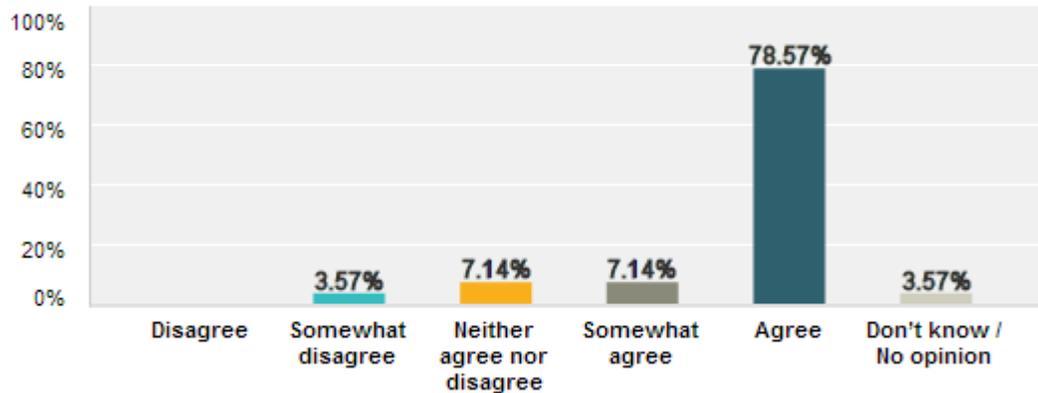


Chart 4-20: Use of "Needs a Review" feature

Most of the students agreed that being able to mark the captions for 'Review' by another editor was useful when they were unable to understand the speech or were unsure of the corrections. Some students commented 'It was nice to be able to go directly to the problem' when they wanted to review the captions that were marked for review by other students. A few students commented that they did not understand the purpose of this feature.

Question: The interface indicated clearly the captions that were complete and those that needed work. Please express the strength of your agreement. (Different color for the row and status was used to display this information)

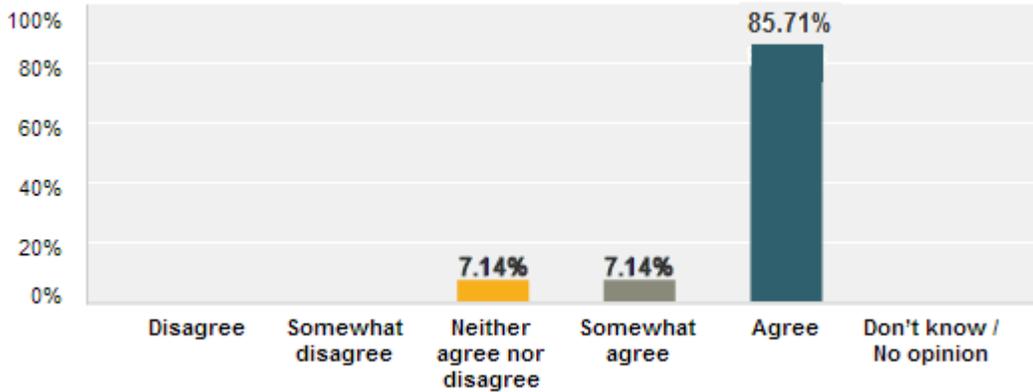


Chart 4-21: Indication of captions that were complete and those that needed more work by the interface

Most of the students agreed that they could distinguish between the captions that were complete and those that needed work by the status information and color coding.

Question: The ICS Caption Editor: How-to-use video was helpful in understanding the instructions. Please express the strength of your agreement.

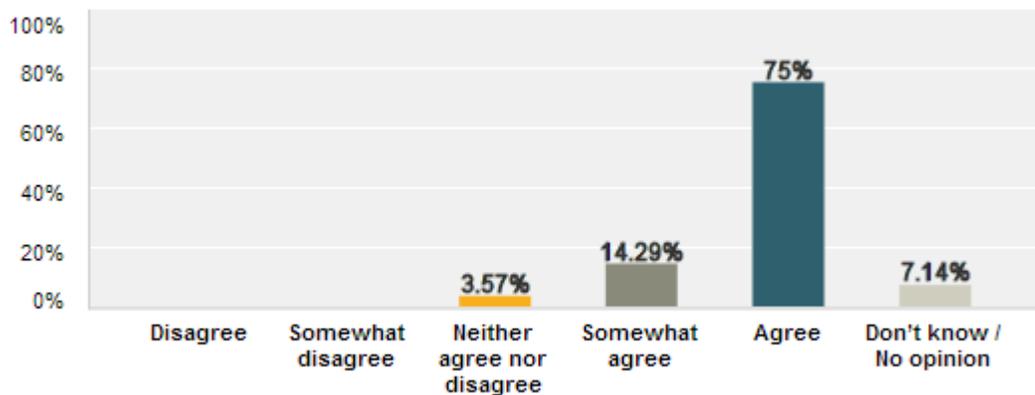


Chart 4-22: Use of how-to-use video

Most of the students could understand the instructions explained in the video and were able to use the editor.

Question: The HELP link provided on the screen was useful in understanding the instructions. Please express the strength of your agreement.

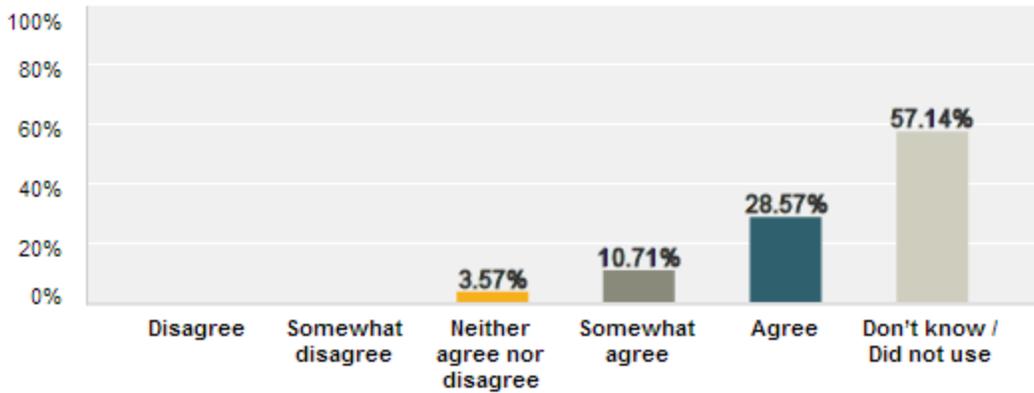


Chart 4-23: Use of the HELP link

Many students did not use the HELP link as they had already viewed the video for instructions.

Question: The placement (position) of the following elements and controls on the Caption Editor interface was appropriate. Please express the strength of your agreement.

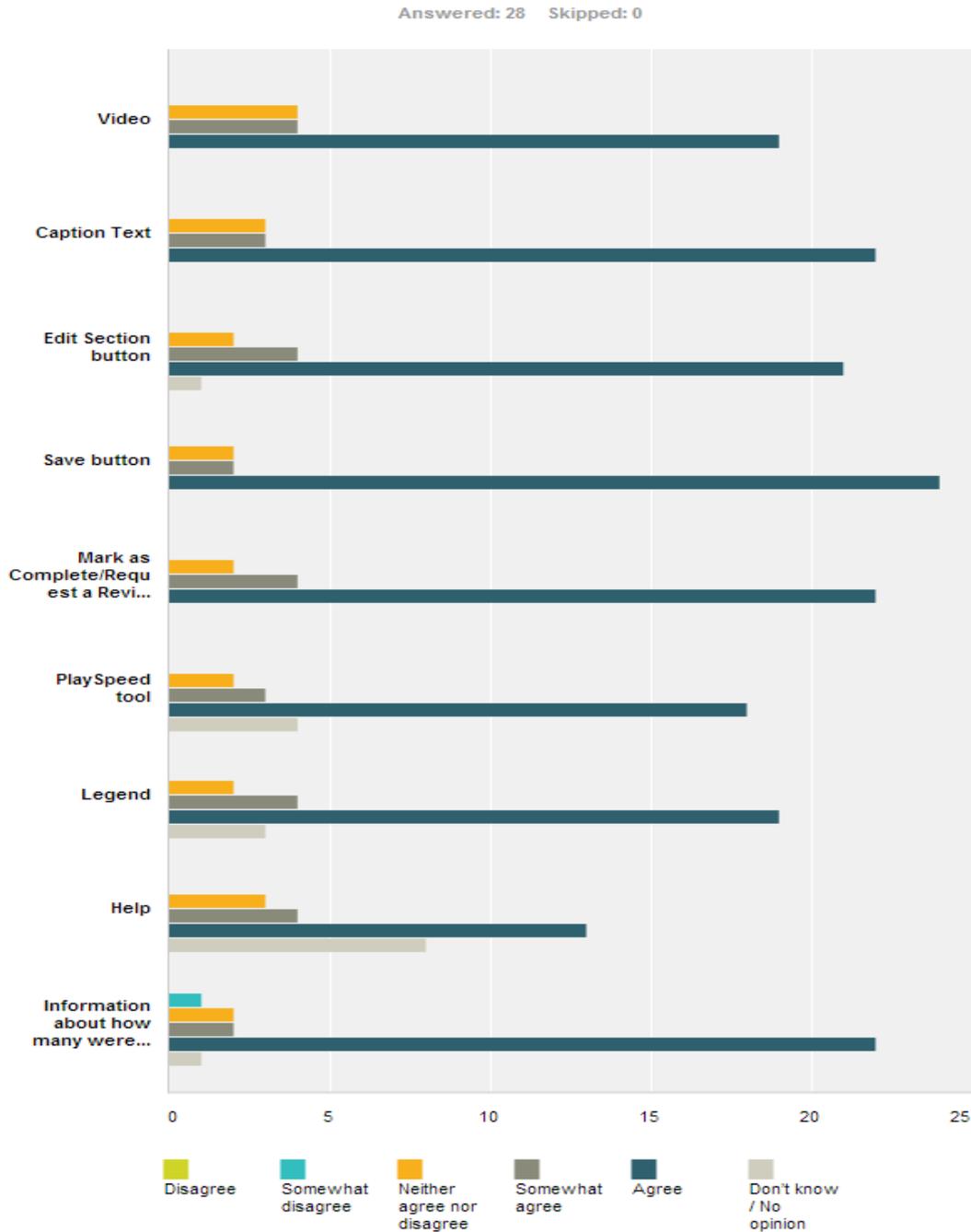


Chart 4-24: Position of elements on the interface

Most of the students agreed on the placement of the controls on the editor's interface. No comments were received by the participants who somewhat disagreed with the placement. Therefore, we are unable to take decisions on improving the placement of controls at this time.

Question: Would you be interested in working with other students to correct captions for your class lectures using this caption editor if you receive some incentive (for example academic credit)?

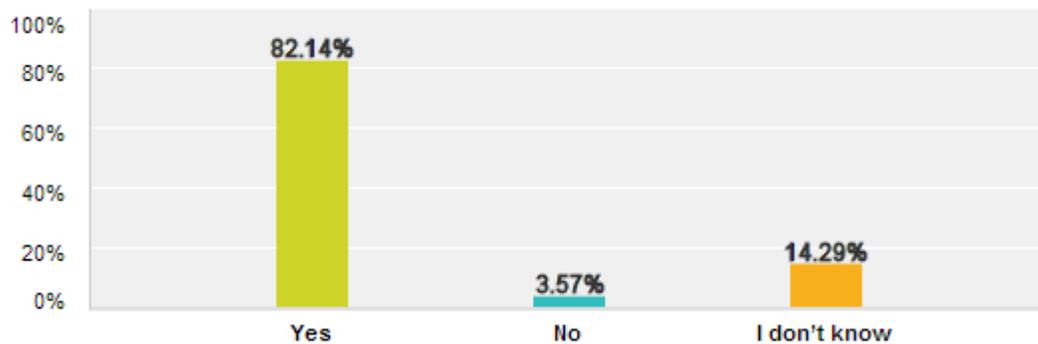


Chart 4-25: Students interest to edit captions

A mostly positive feedback was received for the above question. Student comments showed that they would like to edit the captions in return of some incentive.

4.2.1 Conclusions

We were able to gather information from the participants through the survey which is analyzed in Section 4.2. Demographic information from the survey shows about 70% were male and 30% were female participants aged between 18 and 34 years. The

participant group was highly diverse as demonstrated by the ethnicity graph. Though maximum number of participants has their major as Computer Science, there were also participants from other majors such as Nursing, Chemistry, Biology, etc. as demonstrated in the Chart 4-8.

From the survey results, it was evident that the interface was found to be easy to use. Participants from not only computer-related majors but also other majors were able to use the editor successfully. The video proved to be very useful as a tutorial to use the editor. They like the idea of being able to mark the captions for review. In the survey, we also received some information about problems faced by the participants and some suggestions.

Based on those, following improvements were made:

1. Double click to loop the video: When participants clicked inside the textbox, the video starts looping. But when they want to jump a few words ahead, they need to use the keyboard arrow keys which are time consuming. To avoid looping the video every time mouse is clicked inside the text box, the functionality is changed to 'double click'. The video loops only when the editor double clicks inside the text box.
2. Logout the user after inactivity or browser window is closed: If the editor does not logout, the sections locked by that user remain locked and no other editor is able to edit that section. Functionality to auto-logout the

user after 10 minutes of inactivity or when the browser window is closed was added.

Chapter 5 : Assessment of Value of Captions

5.1 Results of Focus Group Conducted in Fall 2012

Focus groups were conducted in Fall 2012 by Dr. Lecia Barker and the team to assess the value of captions for students who had video lectures captioned. This focus group was for COSC2410 class of Fall 2012. Excerpts from Dr. Lecia Barker's report:

Caption Function

“Students think that captioning is a useful feature. One said that, ‘It's cool... because sometimes there is some knowledge which is not in the presentation, but the professor mentions it.’ They see great value for captioning information which is not included in slides. For example, there may just be formulas on the slides, but the professor will spend a lot of time explaining the slides. For international students, they may be more likely to not drop a class if they know they could get captions of what the professor says. One student requested an option to select the font size of the captions. Students may be working on a variety of systems which have different screen sizes.”

Transcript

“Students who had used the transcript found it very useful. One student said, ‘I have bad hearing, so the transcript was almost everything for me.’ Another stated that ‘The transcripts helped me to search through the video to find just what I wanted.’ The

transcript also helped when the students had problems with the quiet audio. Students suggested that the transcript be made easier to view. One student said, ‘The transcript was just a big wall of text to me.’ The format that the transcript is presented in does not lend itself to ease of use. The students would like for it to be separated and not be presented as long paragraphs. They would like for the timestamps to be separated by the text, possibly with each sentence using differently colored text. Students said it would be useful if the transcript was separated by index or by slide. They suggested that a different background color be put behind the text when a slide changes. Finally, students would also like for the transcript to be searchable.”

5.2 Survey Conducted in Spring 2013

In Spring 2013, we conducted a survey to assess the value of captions as perceived by the students. Two classes were involved for this survey:

1.COSC1300: Introduction to Computing

Professor: Dr. Olin Johnson

2.COSC2410: Computer Organization and Programming

Professor: Dr. Nouhad Rizk

COSC1300 had 9 lectures captioned and COSC2410 had 12 lectures captioned. In this chapter, we present the analysis of the results of the survey. Twenty-four students participated in the survey.

5.2.1 Survey Results

Demographic Information:

Question: Specify your academic year.

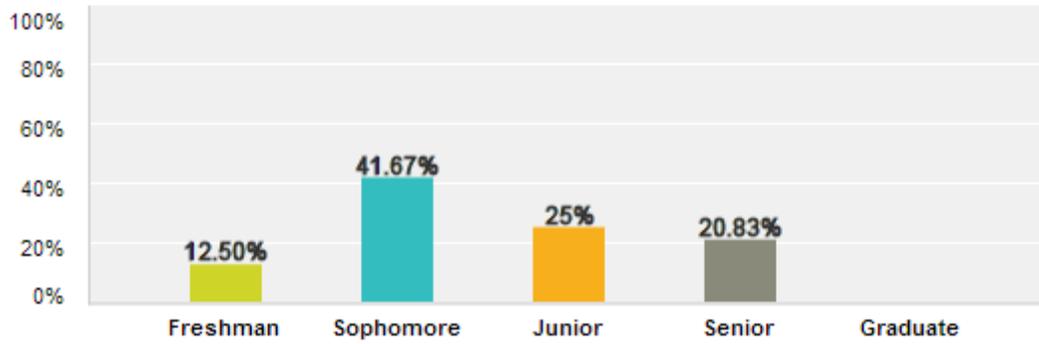


Chart 5-1: Academic year

Question: Specify Gender:

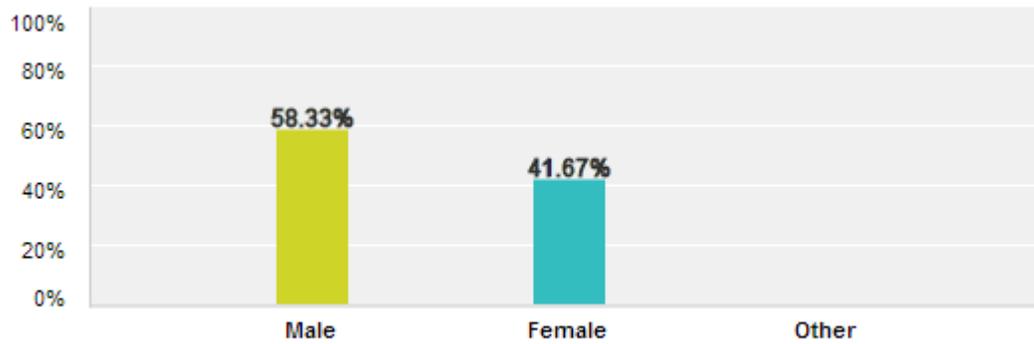


Chart 5-2: Gender

Question: Would you identify yourself as having any of the following: Choose all that apply.



Chart 5-3: Identification of disability

Question: Please indicate your Ethnicity: Check all that apply

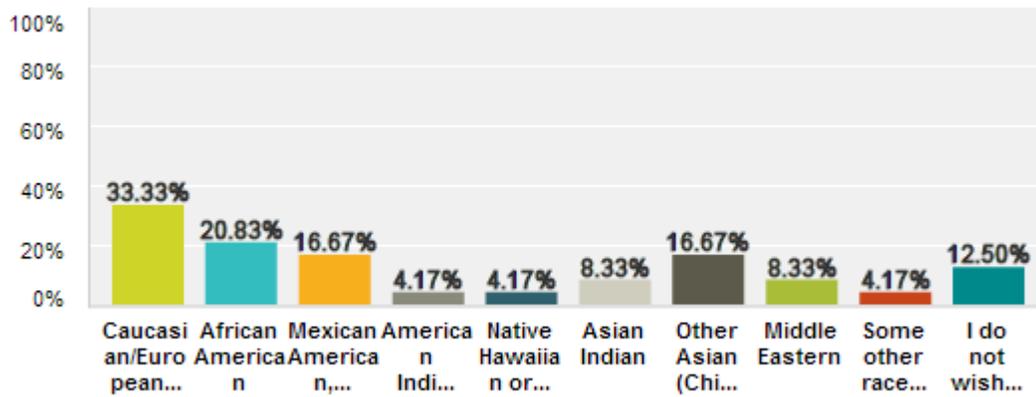


Chart 5-4: Ethnicity

Question: What is your age?

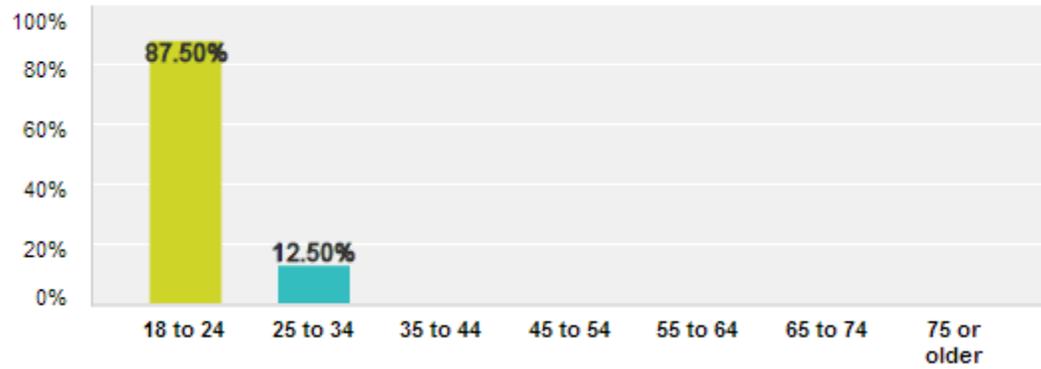


Chart 5-5: Age

Question: How would you describe your fluency with the English language?

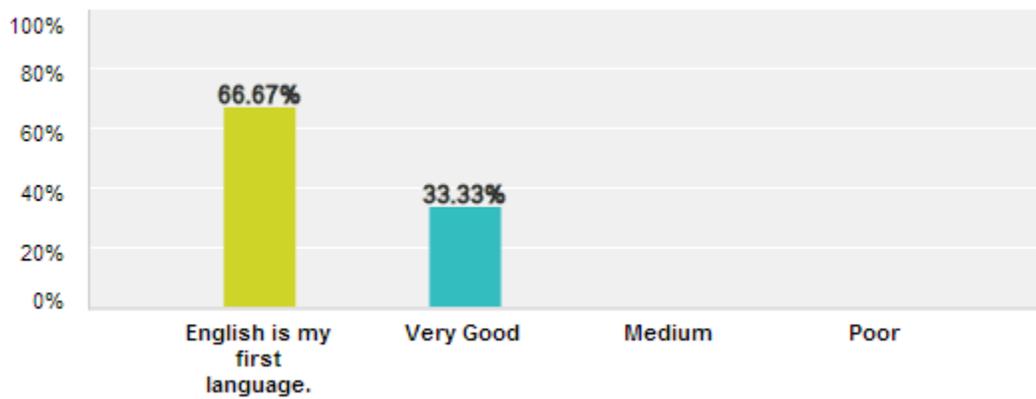


Chart 5-6: Fluency with English language

Question: What is your major?

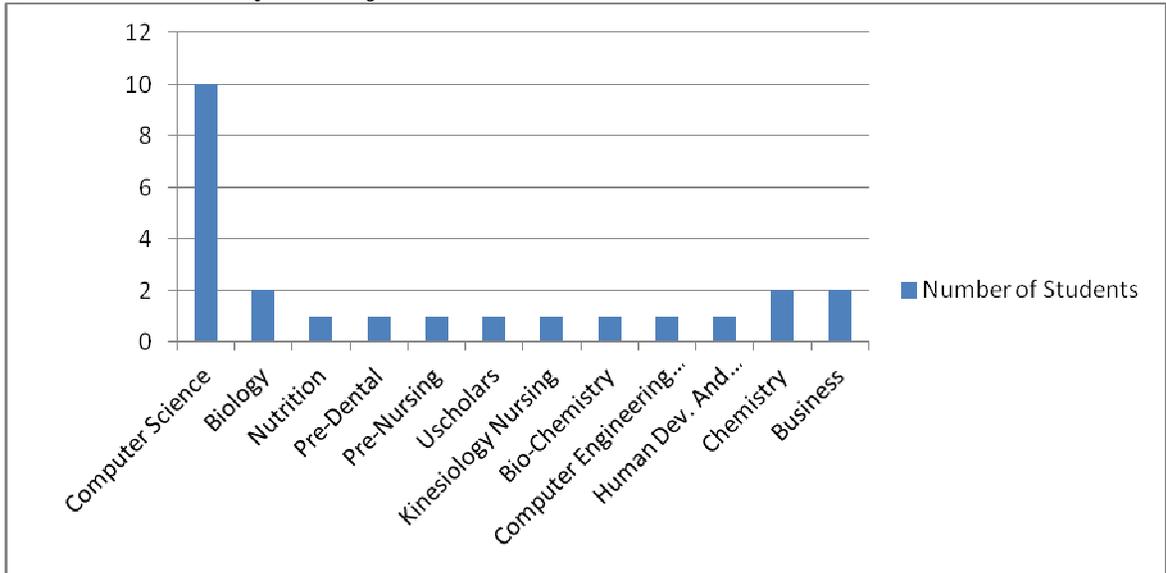


Chart 5-7: Major

Summary of Demographic Information:

The participants of this survey were 24 students from COSC1300 and COSC2410 class. 58% of the participants were male and 42% were female all aged between 18 and 34 years of age. From the ethnicity chart and from chart 5-7, it can be seen that we surveyed a very diverse group. Two of the participants identified themselves as having a visual disability, while others had none.

Question: Specify the course for which you viewed videos that had captions and a transcript:

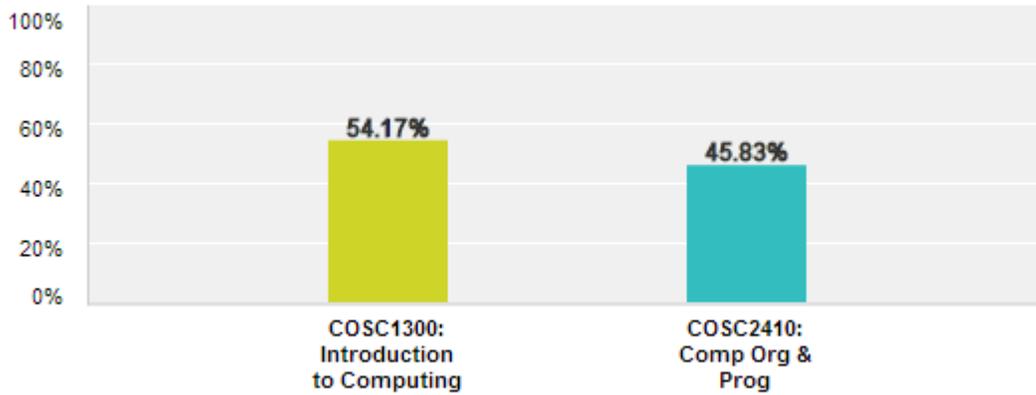


Chart 5-8: Course for which captions were viewed

Question: Nine videos of your COSC1300: Introduction to Computing class included captions/transcript of Dr. Johnson's speech. Did you use any of the videos with captions?

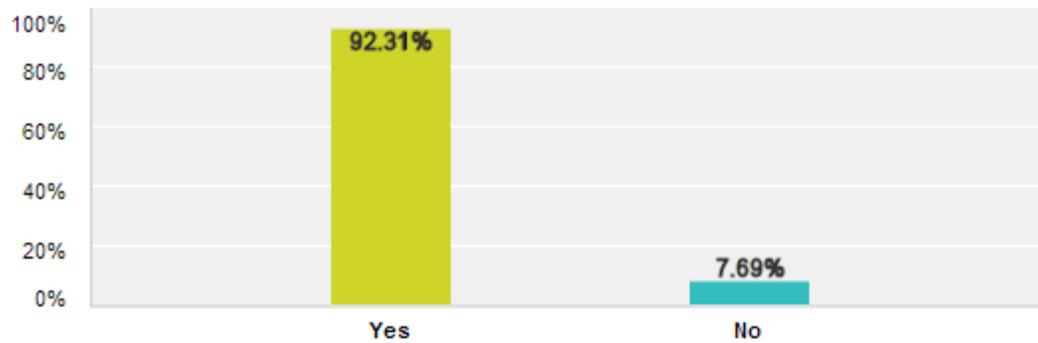


Chart 5-9: Usage of videos with captions for course COSC1300

Question: Twelve videos of your COSC2410: Computer Organization & Programming class included captions/transcript of the professor's speech. Did you use any of the videos with captions?

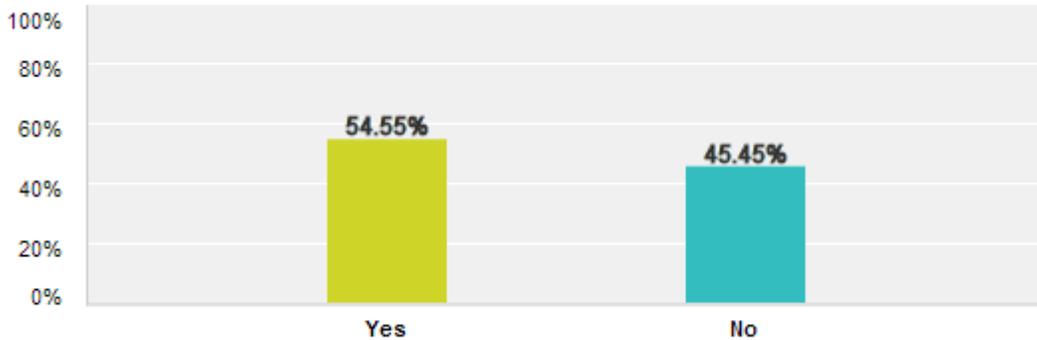


Chart 5-10: Usage of videos with captions for course COSC2410

Question: COSC1300: How many videos with captions did you use?
Average Calculated:



Chart 5-11: Usage of videos with captions for course COSC1300

On an average, six videos out of the nine captioned were viewed by the students of COSC1300 class.

Question: COSC2410 How many videos with captions did you use?
Average Calculated:



Chart 5-12: Usage of videos with captions for course COSC2410

On an average, five videos out of the twelve captioned were viewed by the students of COSC2410 class.

Question: The captions and transcript helped me understand what the professor was saying. Please express the strength of your agreement.

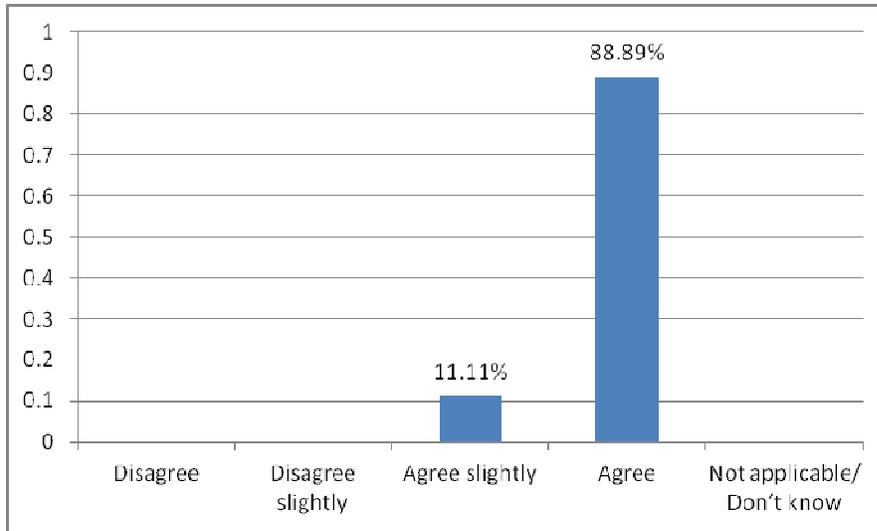


Chart 5-13: Use of captions to understand the professor

All the students either agreed or slightly agreed that the captions and transcript helped them understand what the professor was saying.

Question: The captions and the transcript represented accurately what the professor said. Please express the strength of your agreement.

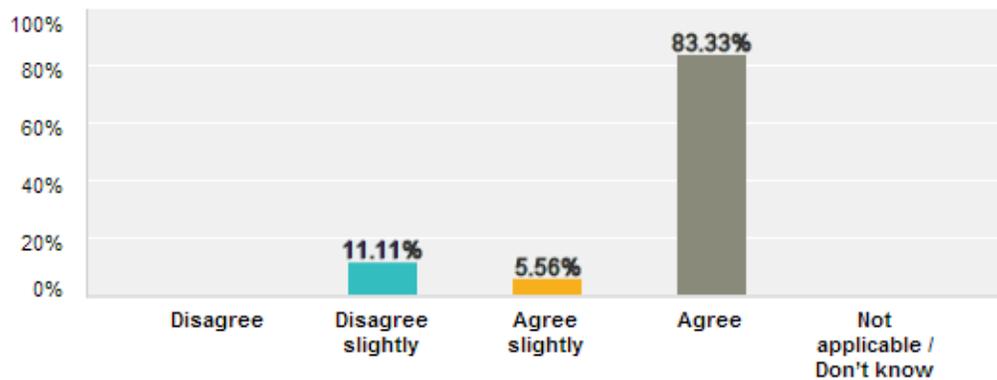


Chart 5-14: Accuracy of captions

Most of the students agreed that captions / transcript (captions and transcript is the same text presented in different format) represented accurately what the professor was saying. However, some students disagreed.

Question: Please comment on the effect captions had on your learning experience in the following aspects: Choose the option improve, no change or reduce. [45]

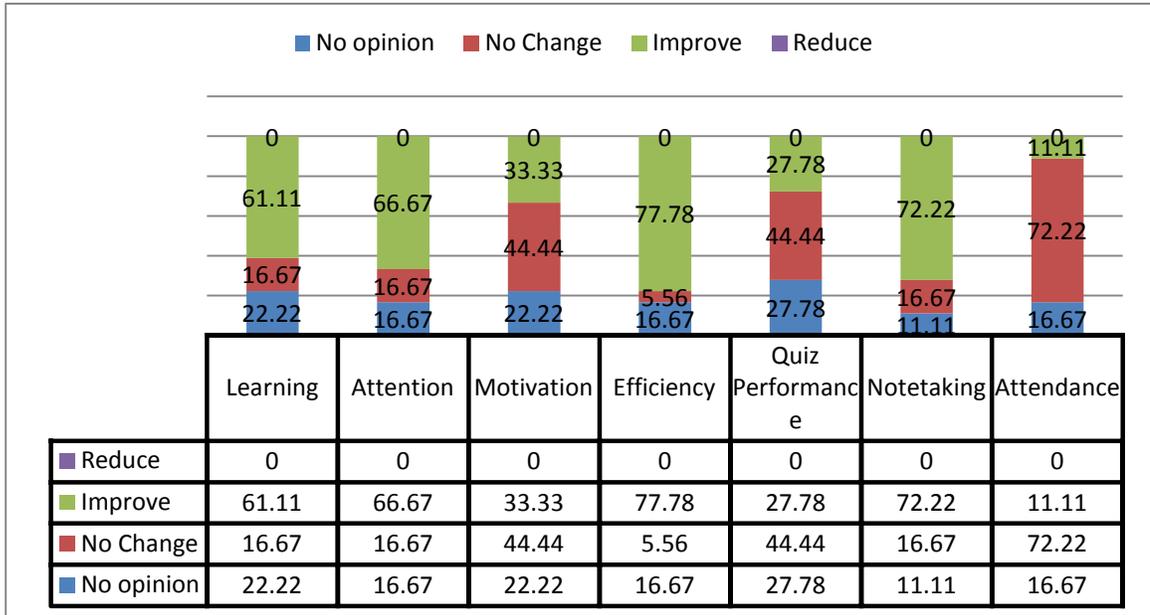


Chart 5-15: Effect of captions on learning experience

Students have claimed that captions have affected their learning experience and have had observed an improvement in their learning of the content, attention, efficiency and note taking. Not much change is observed in their attendance.

Question: Do you believe that the transcript is useful to quickly read through what is discussed in the video without having to watch the entire video. Please express the strength of your agreement.

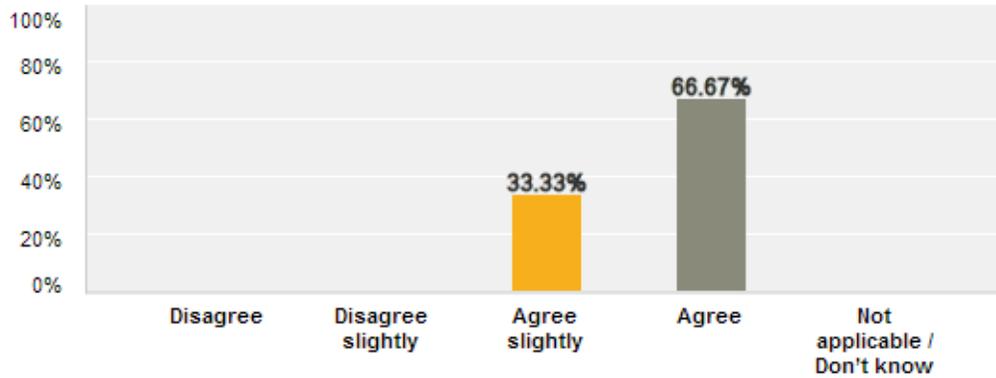


Chart 5-16: Use of transcript to understand video contents quickly

All the students either agreed or slightly agreed that the transcript was useful to go through the lecture quickly to know what was discussed without having to watch the entire video. One student commented that “This was my favorite aspect because now I can get the information at my own rate.”

Question: Clicking on any of the sentences in the transcript allows you to go directly to the point in the video when the sentence is spoken. Do you believe this feature is useful? Please express the strength of your agreement.

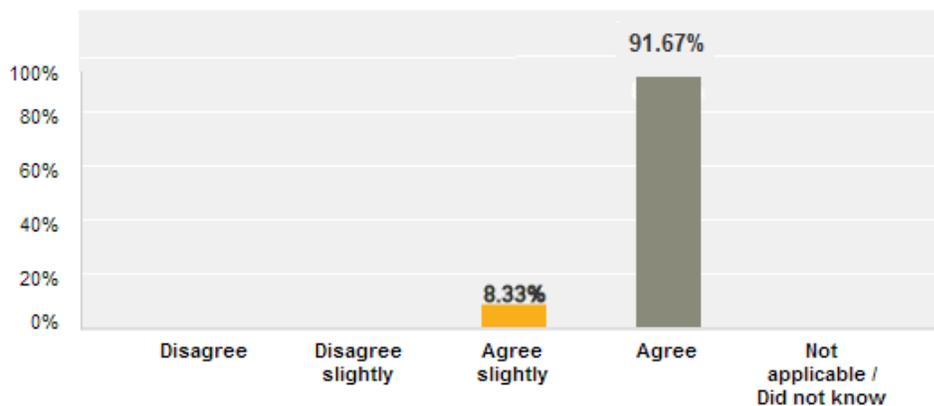


Chart 5-17: Use of transcript to go directly to a point in the video

All students either agreed or slightly agreed that they found this feature useful. One student commented ‘it can be used as a marker in case u want to review a specific problem form the lecture’. Another student commented ‘helpful when finding certain material’.

Question: When the video is played, the currently spoken sentence is highlighted in the transcript box. Highlighting the current spoken sentence is helpful to track the progress of the lecture. Please express the strength of your agreement.

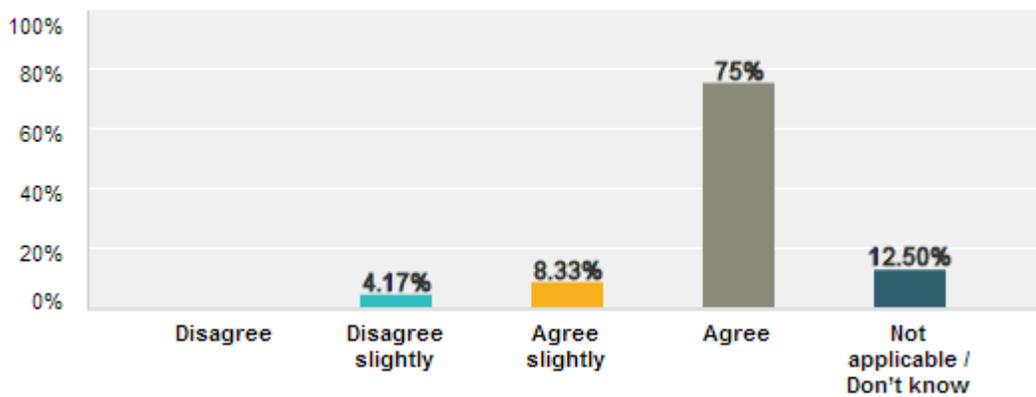


Chart 5-18: Usefulness of the highlighting feature of the transcript

Most of the students agreed or slightly agreed that this feature is useful to track the progress of the lecture. Some of the comments received were ‘Anything is good as a marker’ or ‘helpful to keep track of what teacher is saying.’ Not all the videos had this feature working. Hence, some students were not aware of this feature.

Question: The videos with captions/transcript (text given for spoken sentences) are preferable than videos without them. Please express the strength of your agreement.

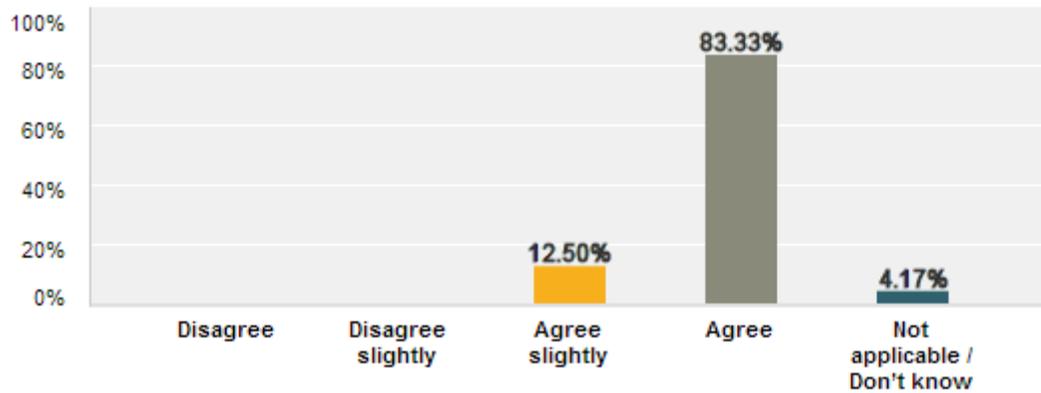


Chart 5-19: Preference to video with/without captions

Most students preferred videos with captions rather than the ones without them.

Question: Do you believe that the transcript and captions would be more helpful if they were in your native language (for example in Spanish, Chinese etc) Please express the strength of your agreement.

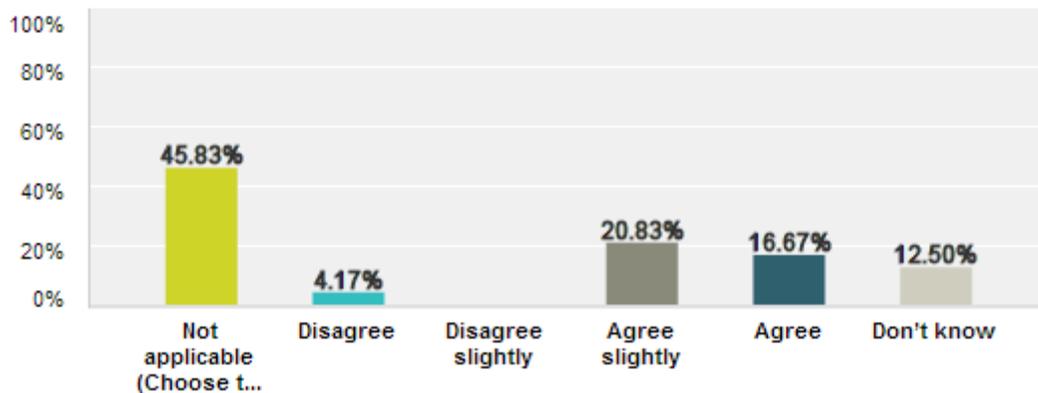


Chart 5-20: Usefulness of captions in another language

5.2.2 Conclusions

We assessed the value of captions as perceived by students to whom captioned videos were available. It can be observed from the survey results that students like having captions for their lecture videos. Students like having the transcript to be able to quickly read through the video. Their efficiency of viewing the video has improved due to interactive and clickable transcript. Focus group results show that captions and transcripts are highly valued by students with hearing disability and by international students. The question whether captions should be made available in other languages than English received a distributed response. Most of the students agreed that captions helped them understand the content of the lecture.

Chapter 6: Summary

In this chapter, we will discuss what was accomplished in this work and what future developments could be beneficial.

6.1 Conclusions

We conducted focus groups and survey to assess the value of captions for classroom lecture videos as perceived by students. Focus groups and other research show that captions and transcripts are highly valued by students, especially those with hearing disability and whose first language is not English. Text for speech helps them understand what the professor is saying. They claim that captions have improved their learning experience, especially in aspects such as efficiency, learning, and note taking. They feel that transcript helps them to quickly read through the contents discussed in the video. Clickable transcripts help them to quickly go to a point of interest in the video.

To employ the use of speech recognition technology to automate the process of generating captions, we assessed the accuracy of currently available speech recognition tools. We conducted the study with 3 faculty members. The results showed that the accuracy of these tools is low for lecture transcription. Therefore, the output given by these tools cannot be used directly as captions. The technique of parroting gives acceptable accuracy levels but the process of one person parroting the entire lecture is seen as being impractical. The process of parroting will not scale when the number of lectures that require captioning increase. It is concluded that after the text is given by the

speech recognition tool, an additional step, of correcting these captions manually is required.

For the purpose of editing the captions, we designed and implemented a caption editing tool following the usability requirements mentioned in Section 3.1. We conducted studies with 28 students from 2 classes. We also conducted a survey with the participants to get feedback about the ICS Caption Editor, to understand the issues they faced while using it and any suggestions to improve the interface. From the study and survey results we can see that the students were able to use the editor successfully and corrected the captions for the lectures. The corrected captions were generated in a modest amount of time. Combined efforts from a group of participants avoided heavy workload on individual participant. They found the interface easy to use. The tutorial provided instructions that were found to be easy to understand. Demographic information shows students not only from computer-related but also non-computer-related majors were also able to correct the captions with this editor. The feature of looping the audio was found to be useful. The feature of being able to mark and read the caption status information, whether the captions are complete or need some more work, proved to be useful.

More lectures can be captioned using this semi-automatic process if student participation is encouraged. More lectures can be provided with captions and transcripts if the videos are recorded and re-used in several semesters. More courses that re-use videos should be included in the captioning program.

6.2 Future Work

Surveys helped us to understand issues in the system and what could be improved. Question in the survey regarding having captions in their native language received a distributed response. Studies to determine which language would be most beneficial to have and how we could provide corrections for that language should be conducted.

With the ICS Caption Editor interface, we have the functionality where users could review other users' corrections. A functionality to automatically verify the corrections could be added. The corrections to a sentence by multiple users could be saved and their corrections could be matched. This would require multiple users editing the same sentence.

The corrections for the captions are manually done by editor and can be error prone. A spell-check mechanism or auto-prediction mechanism could help reduce these errors.

References

- [1] <http://www.nuance.com/dragon/index.htm>
- [2] http://en.wikipedia.org/wiki/Windows_Speech_Recognition
- [3] <http://cmusphinx.sourceforge.net/wiki/research/>
- [4] <http://www.informedia.cs.cmu.edu/dli2/index.html>
- [5] Gabriel Skantze GSLT: Speech Technology 5p. 2003-02-26: The use of speech recognition confidence scores in dialogue systems.
http://www.speech.kth.se/~rolf/gslt_papers/GabrielSkantze.pdf
- [6] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouviet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens: Speech Communication, Vol 49, Issues 10–11, October–November 2007, Pages 763-786: Automatic speech recognition and speech variability: A review
<http://www.sciencedirect.com/science/article/pii/S0167639307000404>
- [7] Schaaf T. Interactive Syst. Labs., Karlsruhe Univ. Kemp, T., Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference: Confidence measures for spontaneous speech recognition
<http://www.cs.cmu.edu/~tschaaf/MyPublications/1997-icassp-schaaf-kemp.pdf>
- [8] Hope L. Doe, Master's thesis, Virginia Polytechnic Institute and State University, July 1998: Evaluating the effects of automatic speech recognition word accuracy
<http://scholar.lib.vt.edu/theses/available/etd-7598-165040/unrestricted/thesis1.pdf>
- [9] M.A.Anusuya, S.K.Katti , (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009, Pages 181-205: Speech recognition by machine: A review
<http://arxiv.org/ftp/arxiv/papers/1001/1001.2267.pdf>
- [10] Broughton, Michael. (2002), Workshop on “Virtual Conversational Characters: Applications, Methods, and Research Challenges” 29th November, 2002, Melbourne, Australia: Measuring the accuracy of commercial automated speech recognition systems during conversational speech.

- [11] Yao, X., Bhutada, P., Georgila, K., Sagae, K., Artstein, R., & Traum, D. (2010), LREC 2010, Proceedings of the 7th International Conference on Language Resources and Evaluation. Valletta, Malta. 17-23 May, 2010: Practical evaluation of speech recognisers for virtual human dialogue systems
- [12] Ben Shneiderman: September 2000/Vol. 43, No. 9 Communications of the ACM, Pages 63-65: The limits of speech recognition
- [13] John Foliot , Sean Keegan : Stanford Captioning System
<http://captioning.stanford.edu/presentations>
- [14] Jaspal Subhlok, Tayfun Tuna, Shishir Shah, Varun Varghese, Olin Johnson , Lecia Barker: SIGCSE'12, February 29–March 3, 2012, Raleigh, North Carolina, USA: Development and evaluation of indexed captioned searchable videos for STEM coursework.
- [15] IBM Collaborative Captioning System:
http://www-03.ibm.com/able/education/downloads/Collaborative_Caption_Editing_System-CSUN-2012_accessible_IBM.pdf
- [16] Camtasia Caption Editor :
<http://www.techsmith.com/tutorial-camtasia-relay-caption-editing-prior.html>
- [17] CaptionTube : A caption Editor by Youtube
<http://captiontube.appspot.com/>
- [18] 3PlayMedia: A captioning, transcription, and translation solution
<http://www.3playmedia.com>
- [20] Georgia Tech Case Study :
<http://www.3playmedia.com/customers/case-studies/>
- [21] EZTooSoft Subtitle Editor
<http://www.eztoosoft.com/subtitle-editor.html>
- [22] Jubler Subtitle editor
<http://www.jubler.org/>
- [23] CaptionMaker: A Caption Editing Software
<http://www.cpcweb.com/nle/>

- [24] M. Wald, University of Southampton, UK: W4A2011 – 'Microsoft Challenge', March 28-29, 2011, Hyderabad, India. Co-Located with the 20th International World Wide Web Conference: Crowdsourcing correction of speech recognition captioning errors
- [25] HTML5: http://www.w3schools.com/html/html5_intro.asp
- [26] JQuery : <http://jquery.com/>
- [27] MediaElement Javascript Library: <http://mediaelementjs.com/>
- [28] Kate S. Hone: Natural Language Engineering Vol. 6, Pages 287-303: Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI)
- [29] Casali, S. P., Williges, B. H. and Dryden, R. D. (1990), Human Factors, Vol.32, No.2, Pages 183-196: Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance.
- [30] Speech recognition research at Google:
<http://research.google.com/pubs/SpeechProcessing.html>
- [31] Ananya Misra, Proceedings of InterSpeech 2012: Speech/Nonspeech segmentation in web videos
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/pubs/archive/40362.pdf
- [32] http://en.wikipedia.org/wiki/Speech_recognition
- [33] Larry Don Colton, PhD thesis, Oregon Graduate Institute of Science and Technology, October 1997: Confidence and Rejection in Automatic Speech Recognition
- [34] Timothy J. Hazen, Stephanie Seneff, and Joseph Polifroni, Computer Speech and Language (2002) Vol.16, Pages 49–67: Recognition confidence scoring and its use in speech understanding systems
http://groups.csail.mit.edu/sls/publications/2002/Hazen_CSL02.pdf
- [35] Software: Subtitle Workshop : <http://subtitle-workshop.en.softonic.com/>
- [36] Software: Express Scribe: <http://www.nch.com.au/scribe/index.html>
- [37] Stanford Captioning System:
http://captioning.stanford.edu/captionvideo_notranscript.php

[38] http://en.wikipedia.org/wiki/Khan_Academy

[39] <http://www.fcc.gov/guides/21st-century-communications-and-video-accessibility-act-2010>

[40] Tayfun Tuna, Master's thesis, University of Houston, December 2010: Search in classroom videos with optical character recognition for virtual learning

[41] http://en.wikipedia.org/wiki/Dragon_NaturallySpeaking

[42] http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects

[43] <http://mediaelementjs.com/>

[44] http://en.wikipedia.org/wiki/Speech_disfluency

[45] Mike Wald, International Journal of the Computer, the Internet and Management Vol.18, No.2, May - August, 2010, Pages 63-69: Synote: Multimedia annotation 'Designed for All'

[46] Jaspal Subhlok, Olin Johnson, Venkat Subramaniam, Ricardo Vilalta, and Chang Yun, SIGCSE '07 Proceedings of the 38th SIGCSE technical symposium on computer science education, 2007: Tablet PC video based hybrid coursework in computer science: Report from a pilot project

[47] Joanna Li, Master's thesis, University of Houston, 2008: Automatic indexing of classroom lecture videos

[48] Gautam Bhatt, Master's thesis, University of Houston, April 2010: Efficient automatic indexing for lecture videos