Captions Versus Transcripts for Online Video Content

Raja S. Kushalnagar, Walter S. Lasecki[†], Jeffrey P. Bigham[†]

Rochester Institute of Technology
96 Lomb Memorial Dr.
Rochester, NY 14623 USA
rskics@rit.edu

[†]University of Rochester
160 Trustee Rd.
Rochester, NY 14627 USA
{wlasecki, jbigham}@cs.rochester.edu

ABSTRACT

Captions provide deaf and hard of hearing (DHH) users access to the audio component of web videos and television. While hearing consumers can watch and listen simultaneously, the transformation of audio to text requires deaf viewers to watch two simultaneous visual streams: the video and the textual representation of the audio. This can be a problem when the video has a lot of text or the content is dense, e.g., in Massively Open Online Courses. We explore the effect of providing caption history on users' ability to follow captions and be more engaged. We compare traditional on-video captions that display a few words at a time to off-video transcripts that can display many more words at once, and investigate the trade off of requiring more effort to switch between the transcript and visuals versus being able to review more content history. We find significant difference in users' preferences for viewing video with on-screen captions over off-screen transcripts in terms of readability, but no significant difference in users' preferences in following and understanding the video and narration content. We attribute this to viewers' perceived understanding significantly improving when using transcripts over captions, even if they were less easy to track. We then discuss the implications of these results for on-line education, and conclude with an overview of potential methods for combining the benefits of both on-screen captions and transcripts.

Categories and Subject Descriptors

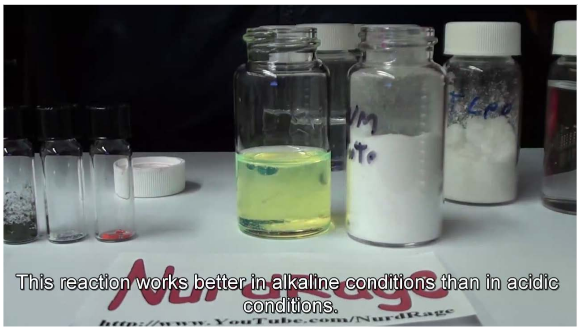
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; K.4.2 [**Social Issues**]: Assistive Technologies for Persons with Disabilities

Keywords

Captions, transcripts, deaf education, online education

1. INTRODUCTION

Captions and transcripts transform speech to text for deaf and hard of hearing (DHH) consumers. People often falsely assume that traditional captions enable full access to online video for DHH people. This assumption is detrimental because it minimizes other information accessibility issues, such as simultaneous visual streams or content complexity.



This reaction works better in alkaline conditions than in acidic conditions.

Figure 1: Video with wrap-around-line on-screen captions

Captions show one or two lines of text, which represent approximately 1-2 seconds of audio, and is overlaid on the video screen, which can sometimes obscure video visuals as in Figure 1. Transcripts show many lines of text representing several seconds of audio, but cannot be overlaid on the video screen, as the text would obscure too much information. So they are shown in a separate window, next to or under the video screen, but this can also be hard to read as the text is further away from the video.

Many variables influence users' ability to read and follow captions or transcripts, including content complexity and degree of visual stream simultaneity. For example, children's shows such as cartoons have less complexity than advanced chemistry lectures. We explore how these factors impact consumers' preferences in following video content with captions or transcripts. This is becoming important due to the meteoric rise in the popularity and availability of Massively Open Online Courses (MOOCs). DHH consumers' accessibility may improve with caption interfaces that match the content complexity and degree of simultaneity.

We compare traditional on-screen captions which show one or two lines long to off-screen transcripts that do not have the same length limitations, and explore if the tradeoff between requiring more effort to switch between captions and visuals is offset by the benefits of content history.

2. BACKGROUND

Prior research has shown that the cognitive process of reading an transcript or caption that constantly changes is very different from the cognitive process of reading print that does not change during the course of reading [11]. For

W4A2013 - Communication, May 13-15, 2013, Rio de Janeiro, Brazil. Co-Located with the 22nd International World Wide Web Conference.

static text, the average college student reads between 280-300 wpm [10, 1]. By contrast the average caption rate for TV programs is 141 wpm and the most comfortable caption reading rate for deaf and hearing is around 145 wpm [3]. Unlike print, captions force readers to read the text at a variable pace; and the readers cannot control or predict that pace. Viewers need time to read the captions, integrate the schema conveyed by the captions and picture to form a single coherent schema and narrative.

Captions can be hard to read when overlaid over a continuously changing video, especially when video text is under the caption text as shown in Figure 1. Captions may also be hard to read by viewers with physical or situational visual impairments, for example in cloudy or dim environments [11], and a 'caption history' may help in these situations.

2.1 Cognitive Overload

Accessible multimedia that includes visual representation of the audio stream (i.e. sign language interpreters or captions) may result cognitive overload, and is a major reason why deaf and hard of hearing students get less out of classroom lectures than their hearing peers [7]. Therefore, accessible multimedia that includes visual representation of the audio stream must be presented in a way that reduces the effects of visual dispersion and cognitive overload [5]. Previous research shows that hearing students benefit from combined visual and auditory materials [9] and multi-modal classrooms are now becoming the norm. Furthermore, while hearing students can simultaneously view a region of interest and listen to the audio using separate modal senses and working memory, this processing strategy is not available to deaf and hard of hearing students receiving accessible presentations. Instead, deaf students multiplex their single visual focus between the visual representation of the auditory channel and the instructor's current visual focus, usually the slides or white board [8].

For captions, readers tend to spend more time on the captions and view the video using their peripheral vision. An eye-tracking study focused on captions found that subjects looked at the captions about 84% of the time [4]. By contrast, subjects do not need to spend as much time on transcripts while viewing video, and they look at the transcript about 68% of the time [2]. It may be that viewers are able to spend less time reading the captions because the video is more likely to be out of their peripheral vision, but also because it is easier for the viewers to alternate between the video and transcript since there is more 'history' to refer to.

2.2 Caption User Interface

There is no single standard for displaying visual transcription (captions or transcripts) on the web, unlike TV captioning. Most visual transcription interfaces of web videos are displayed through browser plugins (e.g., QuickTime) or through built-in browser video functionality (e.g., HTML5). Furthermore, most current web captions continue to use interfaces that hew to TV caption standards that were limited technical constraints in 1970's era technology, such as low bandwidth capacity or lack of options for font type, size or colors. For backward compatibility and consistency, many captioned videos use default features and do not even use common TV captioning best practices, such as multi-line text. As a result, many users of captioned videos can find the videos hard to follow in comparison with TV captioning. All major online video sites (YouTube, Netflix, Amazon, etc) have visual transcription interface options that largely hew to these standards dating from the 1970s despite the technical constraints that have long since disappeared.

3. CAPTIONS

The addition of visual translation of audio essentially transforms the simultaneous viewing and listening experience of watching a video into a sequential reading and viewing experience. In other words, the viewer's cognitive task could be regarded as reading an video enhanced narrative presented through automatic scrolling of text, that is supported with both video and audio supplementary material.

Historically captions have been designed for television entertainment programs, and not for other kinds of programs including education. The number of spoken words, their length, frequency and other factors can influence reading intelligibility and presentation style. Educational videos tend to be more 'heavy' and textual. Typically the presenter uses slides, text or other structured visual materials in their presentations along with their narration. Presenters also often include a great deal of non-verbal contents, e.g., software demonstrations or experiment manipulations to illustrate a lecture. As a result it becomes very difficult for viewers to watch both the video and the captions at the same time [6]. With the advent and popularity of online education, especially MOOCs, it becomes even more imperative to offer adaptable and optimized captioning displays to give maximum benefit to viewers. We compare two presentation styles, on-screen captions and off-screen transcripts below.

3.1 On-Video Captions

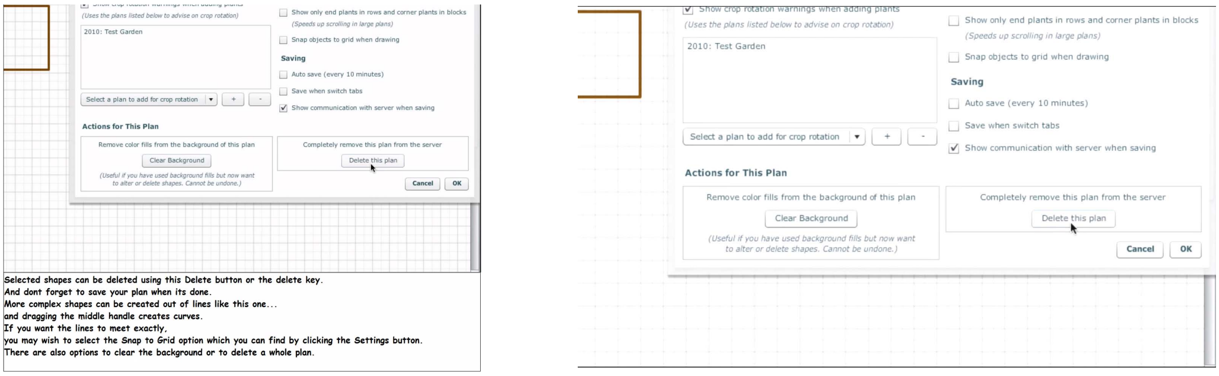
On-screen captions continue to be the most popular visual representation of speech on television and the web due to its ubiquity and simplicity. We present users with single-line captions appearing at the bottom of the screen. While some captioned videos have more than one line, we analyze videos that use captions with only one line because it is the most common case, and because it more clearly characterizes the setup of on-screen captions: small amount of text that are more easy to switch between viewing the video and the captions. The most significant limitation of captions is the amount of information that can be presented without taking up a significant amount of the viewable area of the screen. Also, captions have variable and unpredictable length, and do not visually convey the length of pauses between utterances well, especially when the speaker is not visible.

3.2 Off-Video Transcripts

Inspired by real-time captioners who often display multi-line transcripts on a laptop, we present a similar transcript view separately from the main video. In contrast with limited display time and obstructive nature of on-screen captions, a transcript allows viewers to view content and a brief history without blocking the video. However, viewers expend additional effort when switching between the video and captions as the transcript and video are more separated than video and captions.

4. EMPIRICAL STUDY

Four short, captioned video clips from YouTube were selected. To avoid confounding factors such as pre-existing knowledge, we picked clips with content unlikely to have



☑ Show crop rotation warnings when adding plants
(Uses the plans listed below to advise on crop rotation)
2010: Test Garden
□ Show only end plants in rows and corner plants in blocks
(Speeds up scrolling in large plans)
□ Snap objects to grid when drawing
Saving
□ Auto save (every 10 minutes)
□ Save when switch tabs
☑ Show communication with server when saving
Actions for This Plan
Remove color fills from the background of this plan
(Useful if you have used background fills but now want to alter or delete shapes. Cannot be undone.)
Completely remove this plan from the server

Selected shapes can be deleted using this Delete button or the delete key.
And dont forget to save your plan when its done.
More complex shapes can be created out of lines like this one...
and dragging the middle handle creates curves.
If you want the lines to meet exactly,
you may wish to select the Snap to Grid option which you can find by clicking the Settings button.
There are also options to clear the background or to delete a whole plan.

Figure 2: Examples of a video with transcripts that show several seconds of text captions (left) and with captions that show only one to two seconds of text captions (left), where the captions have disappeared during a pause in speaking.

been seen by college-age participants. The first clip (Chemistry Lecture) explained chemical properties of glow-in-the-dark fluids; the second clip (Optical Illusion) explained an unusual optical illusion; the third clip (Software Demonstration) demonstrated a software program; and the fourth clip (Bike Race) showed a video narration of a rapid action bike race. We selected these videos to compare participant viewing preferences between transcripts and captions.

4.1 Procedure

Of the four selected YouTube videos, three were educational videos that contained a significant amount of content complexity and simultaneous transcribed audio and video visuals such as formulas or optical illusions. The fourth was a bike race event that also had a significant amount of simultaneous speech and visual demonstration, but not content complexity. We divided each into two equal segments, the first with captions and the second with transcripts. We randomly selected each video to display and then administered a questionnaire, in a balanced and repeated measures design. The total time for the study was about 10 minutes. The questionnaire had three Likert questions and an open-ended question. Q1 asked 'How easy were the captions to read?', with a Likert scale that ranged from 1 through 5, with 1 being 'Very hard' to 5 being 'very easy'. Q2 asked 'How easy was it to follow the content in the video?', with the same Likert scale response as in question 1. Q3 asked 'How well did you understand what the video was explaining?', with the same Likert scale response as in questions 1 and 2. Finally an open-ended question was presented to let participants provide their thoughts on the accommodations.

We recruited 17 deaf and hard of hearing participants for the study, of which 6 were female. All were students, ranging from 18-24 years old. All had requested sign language interpreters or captioners for classes on campus, and were familiar with both transcripts and captions. All had used real-time transcripts typed by a captioner in their classes. All had watched captions on television and online videos, such as YouTube. After responding to a short questionnaire to determine eligibility for the test, the participants sat in front of a computer and watched the study videos.

4.2 Participant Comments

Several common themes emerged during the open feedback period. The first theme was that for many lecture videos, it was hard for participants to read the caption text that was overlaid on a background that may have the same color and shapes (letters) as the captions themselves: *Software Demonstration: "make the color of the font yellow so it easier to read"*; *Bike Ride: "The captions were hard to see sometimes."*; *Chemistry Lecture: "Make sure that we can see the caption clearly not confuse with background color."*

A second theme was that participants felt that captions were too fast when information was dense: *Chemistry lecture: "I cannot understand captions when they are too fast. Show more captions."*; *Software Demonstration: "I liked looking at transcript previous lines when I am confused. Captions should have more lines like transcript."*

A third theme was that participants felt that the transcript text was not as big or noticeable as closed caption text: *Bike Race: "The captions are easier to find, but were sometimes hard to read. I want to see the transcript text be bigger and easier to find."*

4.3 Results

We analyzed the responses using the Wilcoxon Signed-Rank test. Over all videos, the responses to Q1 (Whether the video with captions or transcripts were easy to read) were on average slightly higher for captions than for transcripts over all four videos, there was a statistically significant difference in favor of captions ($Z = 144.5$, $p < 0.001$, $r = 12.39$). However, the responses to Q2 (whether the video with captions or transcript was easy to follow) was not statistical significant ($Z = 342.0$, $p = 0.114$, $r = 29.33$). Similarly, the responses to Q3 (whether the video with captions or transcript was easy to understand) also was not statistically significant ($Z = 290.0$, $p = 0.243$, $r = 24.87$).

In summary, when students were asked whether they are able to read the audio transcription while watching the video, students preferred captions over transcripts. This is attributable to transcript text being physically farther from the video as opposed to captions that overlay the video. On the other hand, when students were asked whether they were could follow both the video and audio transcription, students reported no significant difference between following the video with captions or with transcript. Similarly, when the students were asked whether they were able to understand the video and audio transcription, the students reported no significant difference between understanding the video with captions or with transcript. Educational videos usually require viewers to attend to video details such as slide or demonstrations, while reading the captions simultaneously. In fact, there was a slight preference for transcripts in terms of following and understanding as shown in

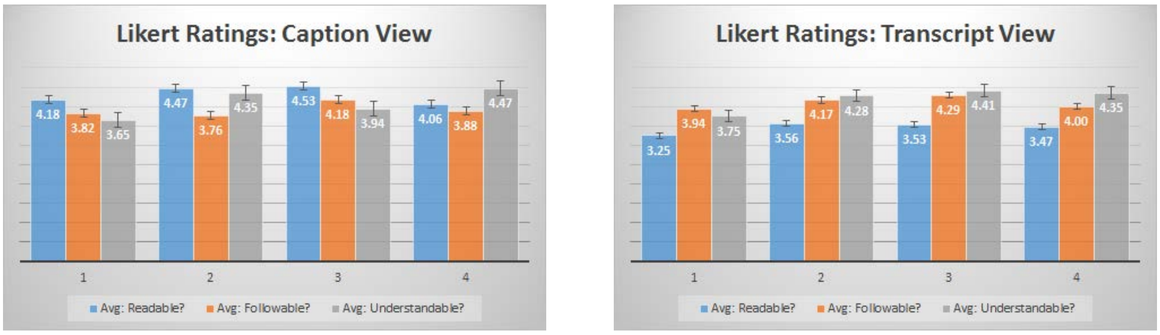


Figure 3: Participant feedback for video with single-line on-video captions (*left*) and multi-line off-video transcript (*right*).

Figure 3, though it was not statistically significant. This suggests that viewers who are following and understanding captioned videos with complex or simultaneous visual and aural content may benefit from showing transcripts. The advantage of viewing eight lines of text off-video appears to cancel out the disadvantage of greater visual dispersion.

Overall, the questionnaire responses suggest that viewers should have a choice of caption views so that they can adapt to varying amounts of simultaneous visual narration text and video visuals, or to the varying amount of vocabulary complexity, or to use alternate views (transcript) when there is a lot of visual text or the background .

5. CONCLUSION

We have presented a comparison of real-time video captions to real-time transcripts. Our results show that captions, which are the most readily used method of displaying speech-to-text content, are preferred by users in typical use cases, but transcripts, with their longer content history, are preferred for more technical content. This finding indicates that online education sources (such as MOOCs) may benefit students by providing real-time transcripts, in place or or in addition to their typical on-screen captions.

Our results show that providing additional captioning history may be worth the attention switching overhead added by having to look farther away from the screen. The additional history enables viewers to review words that stay on screen up to eight times as long on average. Previous eye-tracking studies that tracked percentage of time on the audio transcription indicate that readers devote up to twice as much time on the video while viewing transcripts than captions [4, 2]. This fact supports an inference that readers are able to look back and re-read the words to re-intepret what they have read and viewed. In other words, a longer caption history makes it easier for viewers to integrate and reinforce their learning from multiple, complex video sources. This means that the overwhelmingly popular practice of using on-screen captions likely needs to be rethought in situations where synchronous visual and verbal information is presented, such as in online education, e.g., MOOCs.

6. FUTURE WORK

Our results suggest that personalized caption interfaces could help viewers to adapt and follow better the wide variety of speeds and complexity of different content categories.

There is great potential for captioning interface enhancements that combine the best features of on-video captions with off-video transcripts. Our results will inform the design of captioning user interfaces that best balance the disruption of looking away from the screen, the obstruction of content by on-screen captions, and the benefits of multi-line transcripts. In response to the difficulty in reading captions against changing video background, we plan to study automatic adjustment of the video background or caption text so as to improve its readability. We also will study usability of moving the caption text a minimum distance so as not to obscure the video background or text.

7. ACKNOWLEDGEMENTS

This work is supported by grants from the National Science Foundation IIS-121-70356 and IIS-1218056.

8. REFERENCES

[1] R. Carver. Silent reading rates in grade equivalents. *Journal of Literacy Research*, 21(2):155–166, 1989.

[2] A. C. Cavender, J. P. Bigham, and R. E. Ladner. ClassInFocus. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '09*, pages 67–74, New York, New York, USA, 2009.

[3] C. Jensema. Closed-captioned television presentation speed and vocabulary. *American Annals of the Deaf*, 141(4):284–292, 1996.

[4] C. J. Jensema, R. S. Danturthi, and R. Burch. Time spent viewing captions on television programs. *American annals of the deaf*, 145(5):464–8, Dec. 2000.

[5] R. S. Kushalnagar, A. C. Cavender, and J.-F. Pâris. Multiple view perspectives: improving inclusiveness and video compression in mainstream classroom recordings. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '10*, pages 123–130, New York, New York, USA, 2010.

[6] R. S. Kushalnagar and P. Kushalnagar. Deaf and Hearing Students' Eye Gaze Collaboration. In K. Miesenberger, A. Karshmer, P. Penaz, and W. Zagler, editors, *The International Conference on Computers Helping People*, volume 7382 of *Lecture Notes in Computer Science*, pages 92–99, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[7] H. G. Lang. Higher education for deaf students: Research priorities in the new millennium. *Journal of Deaf Studies and Deaf Education*, 7(4):267–280, 2002.

[8] M. Marschark, P. Sapere, C. Convertino, and R. Seewagen. Access to postsecondary education through sign language interpreting. *Journal of Deaf Studies and Deaf Education*, 10(1):38–50, Jan. 2005.

[9] R. E. Mayer and R. Moreno. A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, 90(2):312–320, 1998.

[10] S. E. Taylor. Eye Movements in Reading: Facts and Fallacies. *American Educational Research Journal*, 2(4):187, Nov. 1965.

[11] F. Thorn and S. Thorn. Television captions for hearing-impaired people: a study of key factors that affect reading performance. *Human factors*, 38(3):452–63, 1996.