Helping Students Keep Up with Real-Time Captions by Pausing and Highlighting

Walter S. Lasecki[1], **Raja Kushalnagar**[2], **and Jeffrey P. Bigham**[3]

ROCHCI, Computer Science[1]
University of Rochester
wlasecki@cs.rochester.edu

ICS, NTID[2]
Rochester Institute of Technology
rskcs@rit.edu

HCI Institute[3]
Carnegie Mellon University
jbigham@cmu.edu

ABSTRACT

We explore methods for improving the readability of real-time captions by allowing users to more easily switch their gaze between multiple visual information sources. Real-time captioning provides deaf and hard of hearing (DHH) users with access to spoken content during live events, and the web has allowed these services to be provided via remotely-located captioning services, and for web content itself. However, despite caption benefits, spoken language reading rates often result in DHH users falling behind spoken content, especially when the audio is paired with visual references. This is particularly true in classroom settings, where multi-modal content is the norm, and captions are often poorly positioned in the room, relative to speakers. Additionally, this accommodation can benefit other students who face temporary or "situational" disabilities such as listening to unfamiliar speech accents, or if a student is in a location with poor acoustics.

In this paper, we explore *pausing* and *highlighting* as a means of helping DHH students keep up with live classroom content by helping them track their place when reading text involving visual references. Our experiments show that by providing users with a tool to more easily track their place in a transcript while viewing live video, it is possible for them to follow visual content that might otherwise have been missed. Both pausing and highlighting have a positive impact on students' scores on comprehension tests, but highlighting is preferred to pausing, and yields nearly twice as large of an improvement. We then discuss several issues with captioning that we observed during our design process and user study, and then suggest future work that builds on these insights.

ACM Classification Keywords

K.4.2 Computers and Society: Social Issues—*Assistive technologies for persons with disabilities*; H.5.m Information interfaces and presentation: Miscellaneous.

Author Keywords

Real-time captioning; caption readability; accessibility; inclusive classrooms; human factors

W4A 2014 - Technical, April 7-9, 2014, Seoul, Korea.
Co-Located with the 23nd International World Wide Web Conference.

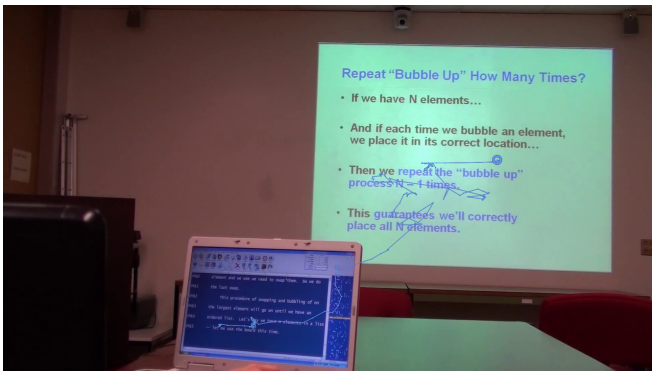

Figure 1. An eye glance trace generated by an eye tracker observing a deaf student who is viewing a classroom lecture. The gaze initially follows the captions, and then shifts to the slide displayed at the front of the room to search for the current topic that was referenced. The focus on the captions does not give the student enough time to study the slide to associate the referenced information with the lecture speech – unlike hearing students who have more time to study the slide and search for relevant information as they listen to the audio. Our goal is to reduce this difficulty by giving users more control over their captions.

INTRODUCTION

We explore methods of improving the readability of real-time captions by allowing users to more easily share their attention between multiple sources of visual information. Real-time captioning provides deaf and hard of hearing (DHH) users with access to spoken content during live events, such as classroom lectures, work meetings, and personal conversations. Moreover, these captions can even be used in situations when other forms of accommodation, such as sign language interpreters, are not available. The web has allowed these services to be provided via remotely-located captioning services, and for web content itself.

Despite the benets of captions, a subtle but important problem results when spoken content is paired with visual references to separate information i.e., on slides or in part of a demonstration (Figure 1). This is particularly common in classroom settings where multi-modal content is the norm [8], and understanding the relationship between spoken and visual information is critical. For most students, the pace of reading captions is slower than the listening pace [5]. This means that it is common for caption readers to fall behind the spoken content, especially in settings such as classrooms.

To make matters worse, captions are often positioned far away from the speaker [9], increasing the amount of time and effort required to switch between content. The reading rate

of non-hearing students is also often significantly lower than that of their hearing peers [17]. Taken together, these issues create additional barriers to learning for DHH students.

In this paper, we explore two simple but effective approaches to helping DHH students keep up with live classroom lectures with mixed visual and spoken content: *pausable* captions and last-word-read *highlighting*. Our experiments measure students' comprehension of material in a live video of a lecture by asking them to complete a comprehension test after using each of our two tools compared to the baseline case of having no control over the real-time captions.

We find that, while both approaches have a positive impact on students' scores on comprehension tests, highlighting is preferred to pausing, and yields significantly higher test scores (14.6% increase when using highlighting captions, versus 7.3% with pausing). We discuss the relevant design criteria based on the feedback we received from users during our iterative design process, and suggest future work that builds on these insights. In general, by providing users with a tool to more easily track their place in a transcript while viewing live video, it is possible for them to follow visual content that might otherwise have been missed.

The rest of this paper is organized as follows:

- We begin with a discussion of prior work in captioning technology and usage, including modern approaches that leverage the web to enable more reliable captions in settings where they may not have previously been practical.
- We then discuss our prior work that demonstrated the need for improving the way captions are presented to and controlled by students in a classroom setting.
- We then present the design of a captioning tool that allows users to *pause* captions while looking away, then easily resume them in order to catch back up to the live video content they are viewing.
- Based on the results of the pausing caption player study, we present the design of a tool that allows users to track their latest reading position in the text with a single button. Our results show that users prefer this method to pausing and perform significantly better on comprehension tests.
- We then discuss the implications of our findings, and design principles for caption presentation that we derived from the feedback of our study participants.
- We conclude with a discussion of potential future improvements to captioning that can further address the problems faced by students presented with captions alongside other sources of visual information.

BACKGROUND AND RELATED WORK

Real-time captioning and caption readability have been studied in a wide range of settings. Many of the approaches to providing usable captions have been assisted by the ability of users and workers to connect to one another via the web. However, to our knowledge, no prior work has studied the effect of tools that allow users to adjust the playback of captions for live events to help them better manage the viewing of simultaneous captions and visual content.

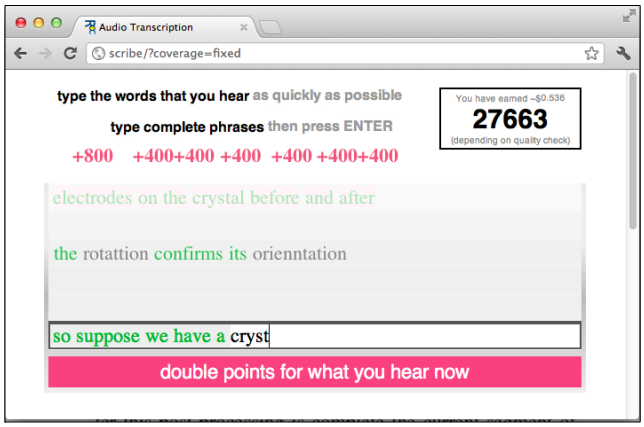

Figure 2. Legion:Scribe worker interface. Workers are prompted to type during certain segments of the audio. These clips are designed to be of a size that workers can handle, and are created by the system such that the partial captions generated by workers can be recombined into a complete caption. Workers also receive feedback about how well they are doing in the form of points.

Real-Time Captioning

Real-time captioning transforms classroom auditory information to text. The accuracy has to be high, while the delay in showing the captions should be within a few seconds.This near real-time translation of audio to text enables DHH students to participate in classroom lectures and discussions. Students usually view captions on a display on a laptop or personal device such as a phone or tablet. They prefer real-time captioning if they do not know sign language or if they find the vocabulary is more accessible through print than sign.

Even professional typists average only about 50 to 80 words a minute, which means they cannot keep up with typical speaking rates that reach approximately 170 words per minute. Therefore, special data entry methods are used to let typists keep up with normal speaking rates. Until recently, the only way to type a verbatim approach was to use a short-hand typing system. In this system, a stenographer types on a specialized shorthand keyboard that is connected to a computer with a short-hand translation program. The program then converts the shorthand to written English and displays it in real-time for the student to view.

The skill and training required to keep up with real-time speech lets professional captionists charge $100 to $300 per hour or more. Additionally, not all content can be accurately captioned by any professional captionist. Settings such as classrooms often include highly technical, domain-specific content and jargon that require that captionists have specialized knowledge [8] – making professional captionists even more expensive and harder to find.

Collaborative Captioning

Collaborative captioning has previously been applied to lecture recordings without a real-time constraint [18]. Recently, Legion:Scribe [14] introduced the idea of real-time collaborative captioning. Multiple individuals can contribute what partial captions they can, and collectively the group is able to caption the entire audio stream in real-time. This approach can leverage non-expert typists such as hearing students in a

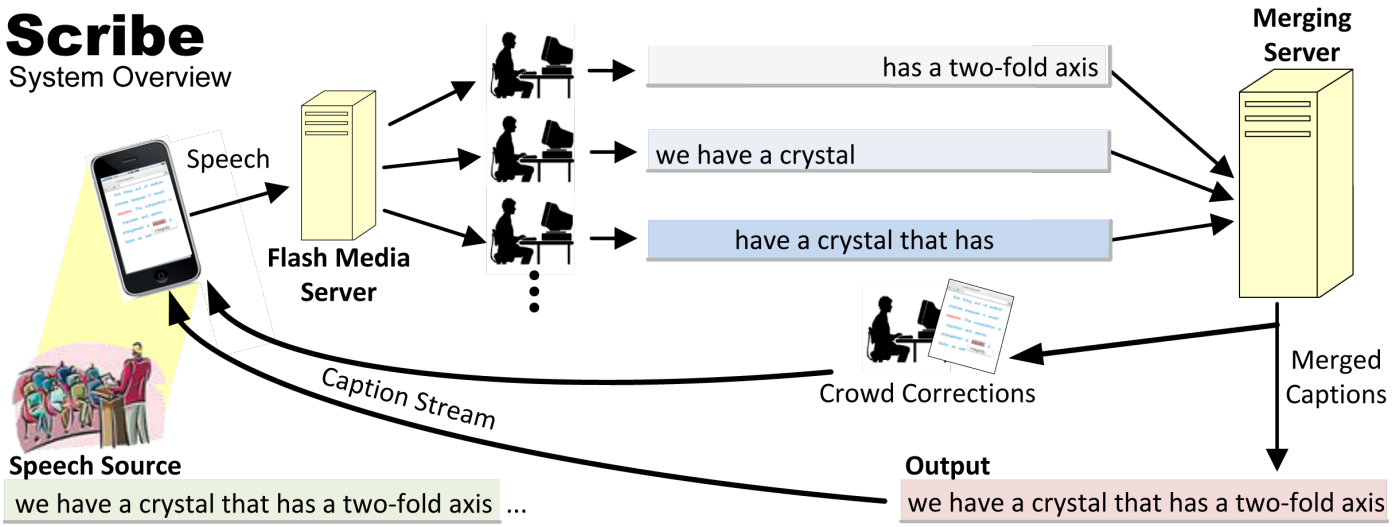

Figure 3. Legion:Scribe system. Audio is streamed from a user's mobile device to a server that divides it into pieces. Workers use the captionist interface (Figure 2) to type what they can of the content they are asked to. Multiple workers' inputs can be combined to increase coverage and decrease latency, resulting in better captions than any of the workers could produce individually.

classroom, co-workers in a meeting, online crowd workers, or any other person who can hear and type, and is willing to contribute. By removing this high barrier to entry to being a captionist, the cost of providing captions, even when multiple workers are hired, can be reduced to as little as one fifth to one half of the price of a professional captionist.

Scribe System

Scribe's worker interface (Figure 2) collects captions from workers, and displays information regarding how much they have contributed (which is converted into earning amounts for paid crowd workers) and when they should be typing the content they hear. These coordinated segments are then merged back together using a multiple sequence alignment algorithm [16]. The Scribe system diagram is shown in Figure 3. This process allows the group to produce captions far more quickly and completely than any one worker alone could. By using modifications to the streaming audio, such as algorithmically slowing down and speeding up parts of what workers hear [13], worker performance can be improved even further.

Recruiting Web Workers from the Crowd

Scribe's continuous crowdsourcing approach is derived from Legion [12], a system that allows multiple workers to control a desktop in real time. By using the web to recruit workers, such as from Amazon Mechanical Turk, we have access to a flexible source of on-demand labor. Workers can be recruited in seconds [1, 2], and can be kept active as long as they are interested in doing so [11]. Since the quality of the output from unknown web workers can be low, indicators can be used to filter out poor quality workers [10].

Automatic Speech Recognition

While Automatic Speech Recognition (ASR) works well with single speaker, in quiet environments, its performance degrades in most higher education settings. These settings usually have extensive technical vocabulary, poor acoustic quality, multiple information sources, or speaker accents. ASR also often adds processing delays of several seconds, which lengthen as the rate of speech increases. A study on untrained ASR software in lectures found that they had a 75% accuracy rate but, with training, could reach 90% under ideal single speaker conditions. However, this accuracy rate is still too low for student use [4, 6].

Unlike human powered captioners, ASR cannot currently interpret or integrate visual references to slides or demonstrations into the captions. For example, assume a teacher is displaying a slide that lists three assignments and discussing their due dates. If the teacher points to the bottom of the list, labeled "Assignment 3", and says "This will be due next Monday," a captioner would look at the teacher's visual reference and type "Assignment 3 will be due next Monday". On the other hand, ASR would not detect the visual reference and would generate "This will be due next Monday".

Classroom Captioning Readability

In the U.S., most students have grown up with classroom accessibility guaranteed by law and their expectations for full accessibility have grown as well. Despite the prior work in captioning, and the expense that is incurred to access these services, they are of no use if DHH users cannot easily read the captions that are produced. While current real-time captions provide substantial access to audio, subtle barriers remain: the presence of multiple simultaneous visuals, and the mismatch between speaking and reading rates. These factors can work against each other, so a universal design approach for captioning requires an interface that can adapt to the individual user's capabilities and preferences [8]. When captions and the lecture visuals (such as slides) are far apart, gaze switching is harder. It is a slow and effortful process that can result in considerable loss of information. Even when the information sources are brought into the student's field of view, the student has to rely on experience to decide when to switch their gaze to fully catch simultaneous visuals.

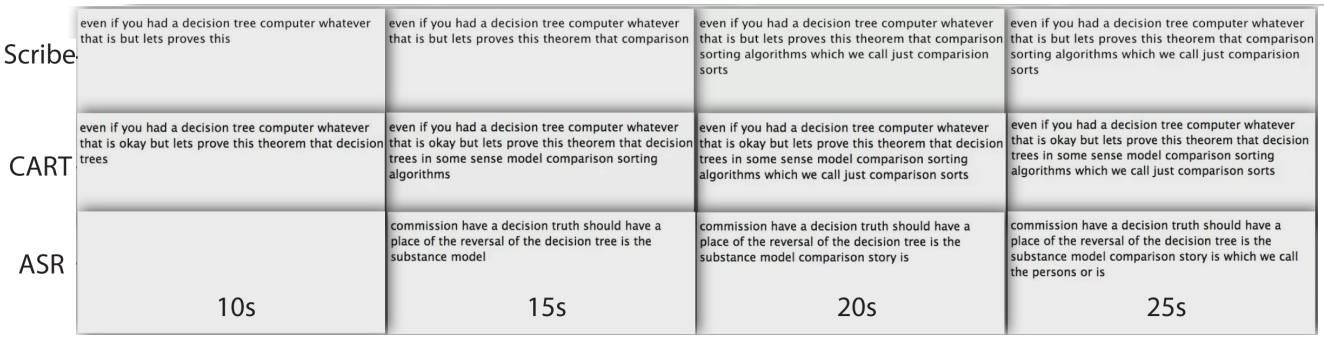

Figure 4. An example of the captions produced by professional captionists (CART), automatic speech recognition (ASR), and Scribe, over time. This shows that in addition to errors, captions can also come in "bursts" (especially with ASR), making them even more difficult for users to read.

Multiple Visuals

Just as important as the captioning accessibility on educational STEM videos is the ability to effectively manage and split attention among multiple visual information sources (e.g., simultaneous presentation of captioning and visual demonstrations), which remains an elusive goal. Although multi-modal instruction benefits hearing students, the translation of aural to visual information can result in deaf learners to miss content since deaf students spend less time watching the slides or teacher, than their hearing counterparts.

Studies have shown that deaf students who use visual accommodations look at the instructor 10% of the time and at the slides 14% of the time [7], compared to hearing students who looked at the slides for 63% of the time and at the instructor 29% of the time. Similarly, Marschark et al., and Cavender et al., found that deaf students spend 15% and 22% and 12% and 18% looking at the instructor and slides respectively [15, 3]. On average, hearing students have 3–4$\times$ more time than their DHH peers to watch the slides and process the information.

In addition to the time spent on reading the captions, deaf students spend a significant amount of time searching for the current information being discussed every time they switch gaze between the captions and the classroom view. This wasted time in searching discourages gaze shifting. Hearing viewers do not need to look at the audio source, while deaf viewers have to actively look at the visual translation of the audio in order to understand it. This is problematic for the deaf viewer who has to switch between the lecture visual and the aural-to-visual translation.

For example, in Figure 1, the teacher is explaining a procedure. In the three second time-lapse snap shot the deaf students' eye gaze paths, the eye-gaze initially is focused on the captions, i.e, visual translation of the teacher's speech. The student does not look at the slide until he or she realizes that something has changed on the screen. Only then can the deaf student switch gaze to the screen. Unfortunately because the student has not had time to visualize the screen, he or she spends extra time searching for the relevant information.

Slow Caption Reading Speeds

The problems stemming from limited student reading speeds are also often made even worse by factors such as inconsistent caption flow (Figure 4), and grammar and syntax errors common to spoken speech. These make captions even more difficult to read, and harder to associate with visual content.

In summary, most DHH students spend a majority of their time reading the captions, and thus, have little time to watch the lecture visuals[8]. This means that students often fall behind, and miss content that requires both seeing visual information as well as reading captions.

FORMATIVE STUDY

In prior work on Scribe [14, 8], we have observed that students often struggle to read the captions produced by our system – though not as much as other approaches because of the improved "flow" of collective captions. Even when using perfect captions – those that are edited offline and synchronized with non-realtime video for playback – we still observed that students struggled to read captions.

Setup

We conducted a survey of 24 DHH students from the National Technical Institute for the Deaf. Our participants all had prior experience with real-time captioning, most of which they gained in college level courses. Students were asked to view a video of a lecture while live English captions were playing, then take a survey asking them about their ability to read and understand caption content. Our example lecture contained common situations where visual information is presented concurrently with captions.

Results

As expected from prior work, students struggled with the speed of the captions, as well as the ungrammatical speech, and lack of clear signal of the speaker's emphasis and tone. We asked users' rating of how easy it was for them to follow content in the video using a 5-point Likert scale. The average rating of ease given by participants was 3.46. We then asked participants how well they felt they could follow the transcripts. The average rating of ease given by the same set of participants was just 2.75. In order to determine if this difference between the perceived difficulty of following the video versus the captions was significant, we ran a Wilcoxon Signed-Rank test on the results. We found that this difference was significant ($z = -2.81$, $p < 0.01$), suggesting that helping users better keep up with the captions would be a logical place to start.

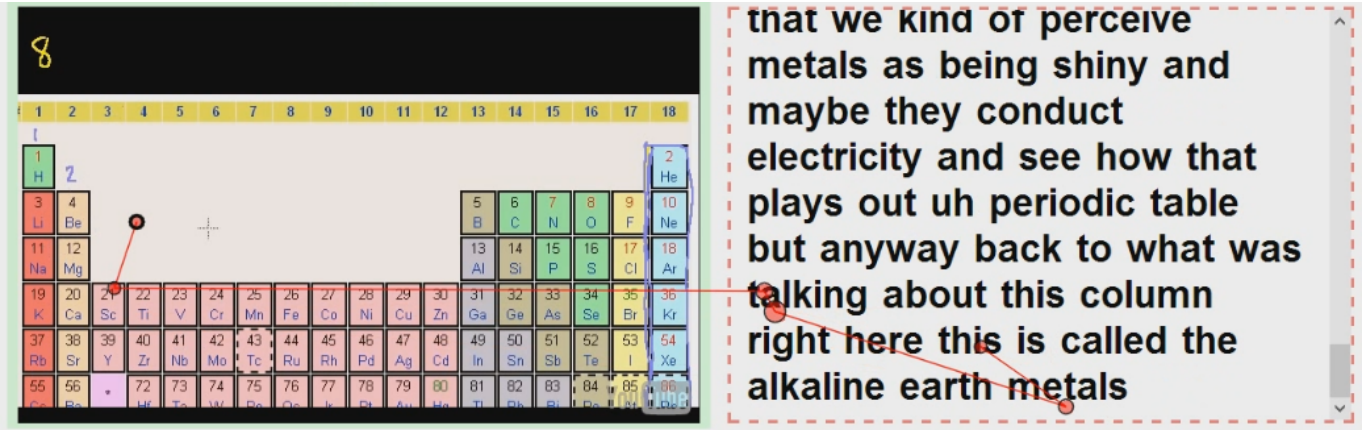

Figure 5. Eye tracking traces (red lines) from a user in our formative study. Here, the user looks away from the captions when the instructor says "right here" to look at the visual information on the periodic table that is being referred to. This figure show the user's eye path when returning from the table to the captions. First, the user under-shoots their previous position in the transcript (looking at where the content *used* to be when they looked away. They next look at the newest content in the transcript as it updates. When they realize this is not they finally scan and jump back to the correct position (the next word that they had not yet read before looking away). This multi-step search process makes it difficult for users to follow content referencing other visual information, and often results in missing even more information as the student tries to recover.

Additionally, participants commented that they struggled to keep up when the content was dividing their attention between visuals and captions. One common reason for this was that users would lose their place in the captions when they looked away to visual content (a characteristic example of which can be seen in Figure 5). For example:

> *"When they show the _x in the video, it distracts me and I kept losing when the +x go"*

> *"I get a little irritated that I would have to constantly look at the board or video to see what the speakers talking about."*

> *"I did not like it because I could never look at the periodic table and see which things the professor was referring to. The [captions were] all I could watch if I wanted to understand the teacher."*

Some students also noted the inconsistent rate of the captions. For example, one participant remarked:

> *"[The captions transition] between low speed to high speed and vice versa."*

TOOL DESIGN AND ITERATION

Our goal is to develop a tool that will allow users to better control the rate at which captions are presented in order to let students follow content that contains both captions and visual content. Our focus is on supporting live content both in classroom settings and in online streaming content.

Pausing Caption Player

There are two main issues that we observed students had during our formative study:

- Being unable to keep up with captions, because speaking rates often exceed reading rates and because captions have potentially inconsistent generation rates.
- Being unable to associate visual content with captions that reference it simultaneously.

To support both of these use cases, we developed a caption player tool that allows users to pause captions as they are presented. While we cannot pause both the captions and the video, our approach allows users to have playback control over the captions, unlike previous work. Figure 6 shows the setup of the pausing captions tool beside a streaming video.

Users are shown a transcript that is initially synchronized with the video shown and updated in real-time as new captions arrive. They are also given a set of playback controls for the captions that allow them to pause, fast-forward, and skip immediately back to the current point of the live video. Pausing can also be performed in two ways: *hold-to-pause*, which allows users to easily start and stop captions by pressing and holding a key when they need to view visual content, and *toggle pause* which allows users to press a key once to pause the captions where they are and again to continue playing them. The delay of the currently shown captions with respect to the live stream is indicated by a progress bar that is green when the user's position matches real-time, and red when the user is behind the live content. To see how far behind the live video content a user is, the bar is filled proportional to how close to real-time the current captions are.

Giving users the ability to pause and resume captions at will, as well as to catch back up to the real-time lecture content either progressively (using fast-forward) or all at once (using "go to live") lets users determine when captions are going too slow for them and they can afford to move through the spoken content faster, and when they want to go to live immediately to catch an important detail in real-time.

Initial User Feedback

To get a sense of whether or not users found the ability to pause just the captions useful during a video, we conducted a study with 10 DHH users from the National Technical Institute for the Deaf. Students were asked to view a short video stream of a class lecture, and use the pausing player tool to stop captions when they needed to.

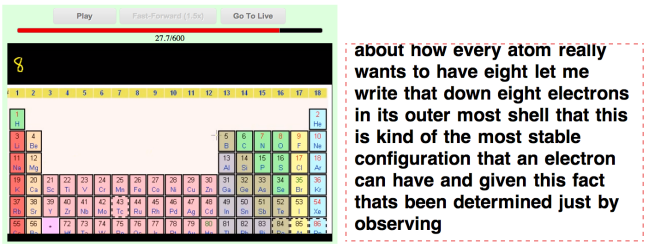

Figure 6. Pausing caption tool user interface. Users are shown a transcript that is initially synchronized with the video shown and updated in real-time as new captions arrive, and given a set of playback controls for the captions that allow them to pause (using hold-to-pause or toggle), fast-forward, and skip directly back to the current point in the video.

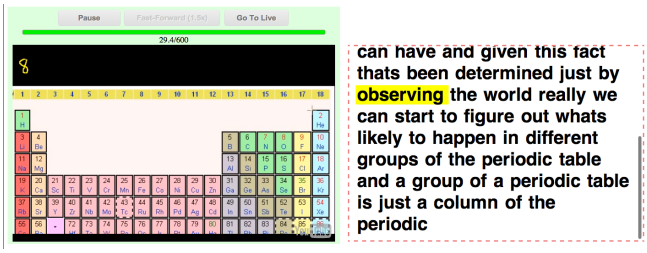

Figure 7. Highlighting caption tool user interface. Users are shown a transcript that is initially synchronized with the video shown and updated in real-time as new captions arrive, just as in the pausing player, but now are able to mark their position using the same two types of interactions as before: hold-to-mark or toggle. The last word visible when they press the hot-key will be marked by a yellow highlight, but the rest of the text will continue as normal.

As expected, users felt that they were unable to keep up with both the visual content and captions. However, several users reported that they did not like the pausing captions because it did not allow them to view what was being said currently after returning from viewing visual content. By pausing, users were unable to glance ahead and quickly assess the importance of understanding the content where they are versus catching up to real-time to be able to ask a question about the material, without losing their place in the transcript.

However, many users still liked the underlying idea of holding their place while they look away from their screens. Some users even suggested that instead of using a system to let them pause the entire caption, they would prefer something that lets them see all of the captions, while highlighting the place where they left off.

Last-Word-Read Highlighting Player

Based on the feedback from our initial users, we designed and implemented a caption highlighting player that allows users to mark the latest word displayed at a given point in time by pressing and holding, or pressing to toggle, a key. Unlike the pausing player, even when a word is highlighted, the captions continue to update as they would normally. This allows users to see and skim newer content, while keeping track of their place if they decide to look away from the captions to watch other visual content. Figure 6 shows the setup of the caption highlighting tool beside a streaming video.

Our highlighting approach does not preclude pausing. Both pausing and highlighting can be used together, providing users with more options that might be better suited to different situations (as in the example in Figure 6). For the purposes of our tests, however, we use only one of these at a time. To explore user preferences when they are able to use both tools at once, we added a session following our study that asked users to use both tools and provide feedback on their experience.

STUDY

To simulate live content while ensuring test consistency, we use pre-recorded video lectures available on YouTube, and display the video with no controls so that it cannot be paused or rewound by participants. This lets us re-create the same content and experience for all users without changing the options users would have when viewing live content.

Setup

Our experimental setup displays the video directly beside the caption (Figures 6 and 7). This setup most closely resembles the design suggested by Kushalnagar et al [9]. While this layout is not always used in classrooms (where content is often separated by much farther distances, making it harder for users to switch between content [9]) it is the view users would have for most online content, and illustrates a best-case setup in a classroom. This means that we expect any improvements to be strictly more pronounced in settings where the cost of switching between visual content is higher. Users were allowed to familiarize themselves with the tool on 3 minutes of video prior to beginning the study.

Measures

To measure the performance of our two caption players, we asked participants to take a short comprehension test consisting of 5 questions about the content they just watched. Participants were given the test after both the first and second halves of the video. The studies were run such that each pair of videos were viewed with one of the tools for one half, and no tool (baseline condition) for the other half. The order of the videos and tool usage was randomized to avoid bias.

Study Participants

Our study participants were recruited from the National Technical Institute for the Deaf at the Rochester Institute of Technology. We recruited 25 students, 10 for the pausing tool trials (7 female, 3 male), and 15 for the highlighting tool trials (17 female, 7 male), with an average age of 21.54 (median of 22) years. Users had no prior experience with our captioning tools, but have an average of 1.84 years of experience with captions and median of 2 years (only one student had no experience, and the maximum was 4 years).

RESULTS

We ran Welch's t-test to assess our pausing and highlighting tools. Welch's test was selected in lieu of a standard Student's t-test due to the unequal sample sizes and variances. We found that both the caption pausing and highlighting tools yielded an improvement in students' scores on the comprehension test, with an improvement of 7.32% and 14.56% respectively. However, of the two, only the highlighting tool resulted in a significant difference ($p < 0.001$).

Participant Feedback

Overall, participants were very positive about the tools presented, especially the ability to highlight captions when they look away to view content referenced on the board. For instance, the following quotes are from study participants:

> *"The highlighter is very helpful. When I saw "this" or "that", I immediately pressed button and looked for the "+" on the screen. It helped me see the location of the group or element being discussed."*

> *"The [highlight] helps me a lot. It helps on track [sic]. However, it is slightly difficult to keep up with."*

> *"It is very helpful to highlight and see where [the] pointer is going. For example - the drawing on video stressed fact that hydrogen is in alkaline metal group, but is not alkaline metal. I would have missed that without highlighting."*

> *"Highlighting is very helpful in reading. I don't get lost when I look at video and back at [the captions]."*

One user commented that the same type of visual would have been helpful for the pointing cursor used by the instructor:

> *"Highlighting helped me keep track of my reading whenever I looked at the video. The pointer in the video was hard to see, so as I spent much time looking for the pointer. I wish the video had highlighting too!"*

Pausing Versus Highlighting

There was a significant 98.79% increase in the improvement seen in the highlighting captions compared to the pausing captions ($p < 0.01$). This aligned with our expectations following from the initial user feedback we received about the pausing caption tool. Additionally, while there is a difference between the baseline conditions that can be seen in Figure 8, this difference was not significant ($p > 0.05$).

To better explain this effect, we also presented users of our highlighting tool with a combined version of the pausing and highlighting tools to see if using both approaches together could be more effective. However, most of these users ended up just selecting one option (usually highlighting) and sticking with it, instead of switching between using highlighting and pausing for different types of visual events. In part, this may be a result of the classroom content not presenting different enough use cases to see such behavior.

When asked to directly compare pausing and highlighting options, users seemed to find that pausing was often too heavyweight to quickly switch back and forth. We also received more feedback similar to the formative study. For example:

> *"Pausing captions meant that I could not see what was being said NOW. I did not like the fast forward in captions as it was too hard to read."*

> *"[With pause] I sometimes get confused when i finish looking at the video and look back at captions because the captions show the old information and i have to wait for it to move forward."*

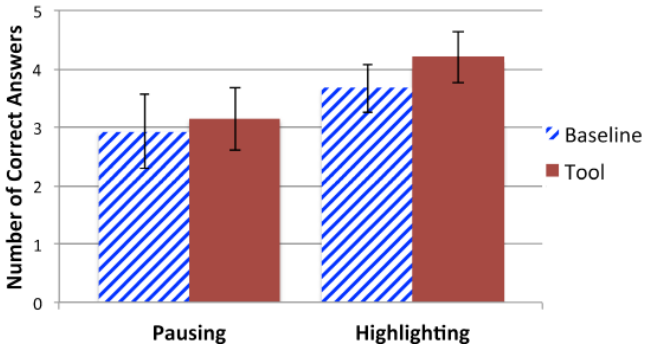

Figure 8. The results of the comprehension tests used in our study with 95% confidence intervals shown. There was a positive improvement seen in students scores using both the pausing and highlighting players (7.32% and 14.56% respectively), however, only the highlighting player was significant ($p < 0.001$). The improvement seen between the pausing and highlighting players was also significant ($p < 0.01$). While there is a difference in the score of the baseline as well, this effect was not significant ($p > 0.05$).

DISCUSSION

Users seemed to value the ability to look away and come back more than trying to make the caption rate more consistent. However, this may be because users have not previously had the ability to do this, so they are not used to the idea yet. On the other hand, using a placeholder in one source of information while viewing another is something that is common to general multi-tasking and thus appeared to be more natural for users in our study.

A vast majority of students were able to quickly use these tools. Out of all of our participants, none of the highlighting tool users, and only two of the pausing caption users actually decreased their score between the baseline and trial conditions. While we had expected some level of variation in this result (since the clips were randomized and the content is different in the two parts of the clip), this is an additional indicator that our approach is reliably beneficial (or at least not costly) to users.

FUTURE WORK

Future work will continue to explore new methods for supporting the use cases we observed throughout our studies. For example, to make a version of "pausing" work better for users, it might help to show where the current reading position is relative to the live content. Allowing users to move this player around in their view more freely than standard playback controls do will also help avoid incurring additional overhead from using the player tool itself.

We will also create a version of our tool that allows both pausing and highlighting to be used on top of existing captioning software, so that any user can modify the caption rate and keep better track of their location in the text. While none of our participants reported being distracted, future work that seeks to provide more complex feedback must keep in mind the potential pitfalls of adding too much complexity.

In general, the effectiveness of our approach suggests that allowing users to control and track streams of information presented to them on the same channel can increase their ability

to effectively comprehend it. This has implications for all users, but especially other sensory impaired users, such as blind and vision impaired people. In that case, a similar problem to that of DHH users is faced, but with multiple sources of information all being carried via audio (e.g., lecture audio as well as a screen reader or scene description). However, these problems are not identical and new issues would have to be considered in future studies.

CONCLUSION

We have presented two approaches, pausing captions and highlighting the last word read in a caption, that allow deaf and hard of hearing (DHH) users more easily deal more easily with situations in which live content and its corresponding captions must be viewed simultaneously in order to understand the content. We then created tools corresponding to each of these interactions while including DHH end-users (students) throughout the design process.

In general, our approaches allow users to keep track of their position when they are reading a real-time caption while visual information (or other visual content) in the scene around them is referenced. By keeping track of their position, users can reduce the overhead associated with switching between captions and other visual content, and thus miss fewer aspects of the content itself. The studies that we have described in this paper demonstrate that a significant improvement in students' understanding of the content of a live classroom lecture video via comprehension test results. Future work will focus on additional ways that users can help themselves better follow multiple sources of content all using the same modality. This allows users to better manage their information consumption, and reduce the time and effort needed to switch focus.

ACKNOWLEDGMENTS

We would like to thank Jason Luu for his input on this project. This work was supported by National Science Foundation under awards #IIS-1218056, #IIS-1149709, #IIS-1116051, Google, and a Microsoft Research Ph.D. Fellowship.

REFERENCES

1. M. S. Bernstein, J. R. Brandt, R. C. Miller, and D. R. Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of UIST 2011*, pages 33–42.
2. J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of UIST 2010*, pages 333–342. 2010.
3. A. C. Cavender, J. P. Bigham, and R. E. Ladner. ClassInFocus. In *Proceedings of ASSETS 2009*, pages 67–74.
4. M. Federico and M. Furini. Enhancing learning accessibility through fully automatic captioning. In *Proceedings of W4A 2012*, page 1.
5. C. Jensema. Viewer reaction to different television captioning speeds. *American annals of the deaf*, 143(4):318–24, Oct. 1998.
6. R. Kheir and T. Way. Inclusion of deaf students in computer science classes using real-time speech transcription. In *Proceedings of ITiCSE 2007*, pages 261–265.
7. R. S. Kushalnagar, P. Kushalnagar, and G. Manganelli. Collaborative Gaze Cues for Deaf Students. In *Proceedings of DuET Workshop at CSCW 2012*.
8. R. S. Kushalnagar, W. S. Lasecki, and J. P. Bigham. Accessibility Evaluation of Classroom Captions. *TACCESS*, 5(3):1–25, 2013.
9. R. S. Kushalnagar, B. P. Trager, and K. B. Beiter. Accessible Viewing Devices for Deaf and Hard of Hearing Students. In *Convention of American Instructors of the Deaf*. 2013.
10. W. S. Lasecki and J. P. Bigham. Online quality control for real-time crowd captioning. In *Proceedings of ASSETS 2012*.
11. W. S. Lasecki and J. P. Bigham. Interactive Crowds: Real-Time Crowdsourcing and Crowd Agents. Chapter In *Handbook of Human Computation*. Ed. P. Michelucci. Springer, 2013.
12. W. S. Lasecki, K. I. Murray, S. White, R. C. Miller, and J. P. Bigham. Real-time crowd control of existing interfaces. In *Proceedings UIST 2011*, pages 23–32.
13. W. S. Lasecki, C. D. Miller, and J. P. Bigham. Warping time for more effective real-time crowdsourcing. In *Proceedings of CHI 2013*, pages 2033–2036.
14. W. S. Lasecki, C. D. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. P. Bigham. Real-time captioning by groups of non-experts. In *In Proceedings of UIST 2012*. pages 23–34.
15. M. Marschark, J. B. Pelz, C. Convertino, P. Sapere, M. E. Arndt, and R. Seewagen. Classroom Interpreting and Visual Information Processing in Mainstream Education for Deaf Students: Live or Memorex(R)? In *American Educational Research Journal*, 42(4):727–761, Jan. 2005.
16. I. Naim, D. Gildea, W. Lasecki, and J. P. Bigham. Text alignment for real-time crowd captioning. In *Proceedings of NAACL-HLT 2013*, pages 201–210.
17. M. D. Tyler, C. Jones, L. Grebennikov, G. Leigh, W. Noble, and D. Burnham. Effect of caption rate on the comprehension of educational television programmes by deaf school students. *Deafness & Education International*, 11(3):152–162, 2009.
18. M. Wald. Crowdsourcing correction of speech recognition captioning errors. In *Proceedings of W4A 2011*, page 1.