Contextual Captions: Improved Captioning for the Hearing Impaired

Samson B. Drews

June 9, 2017

Abstract

This paper outlines a brief history of closed captioning and addresses problems with the current system. It proposes a solution to these raised problems, especially concerning hearing impaired audiences. These solutions were then tested in a user study to determine their significance.

Contents

List of Figures

1 Introduction

Video and visual media have become an important carrier of information, whether it be for entertainment or instructional videos. As a result of recent rapid advances in video accessibility in terms of capturing, storing, and availability to the public, video is playing an increasingly significant role in people's daily lives. The explosion of video has ostracized many with hearing impairment, though. The millions that suffer from hearing impairment, whether fully or partially impaired, have more difficulty in comprehending video content due to the loss of essential auditory information. There are 124 million people with moderate to severe hearing disabilities around the world [15], and there has been little efforts in assisting this special audience since captioning was invented in the early 1970's [7]. The current method definitely offers some information for the hearing impaired, but there is significant loss of information and context that is otherwise provided by auditory means. There are four main problems with current captioning:

- Uncertainty regarding which character is presently speaking. For instance, a scene that has multiple characters in it requires a hearing impaired audience to spend the time to determine who might be speaking. This degrades entertainment value and more importantly makes it more difficult to interpret a given scene.
- Uncertainty of the pace that a given character is speaking. Currently, a portion of the script is displayed statically at the bottom of the screen with no signification of how slowly or rapidly a character is speaking. For example if a character elongates a word, it is very difficult for a hearing impaired audience to determine this. Also, if a character is speaking very rapidly, the text will have to update rapidly as well, sometimes forcing a viewer to miss a portion of a phrase.
- The lack of conveying volume variation. Different levels of volume in video can convey a vast amount of information and emotion. For instance if a character is speaking loudly or yelling, current captioning offers no way to convey this except with punctuation, which comes at the end of the sentence where it is often too late and an interpretation of a scene may have already been made.
- Displaying contextually important sounds. Sounds like a door bell ringing or a gunshot with the current method is to simply say what the sound is. For instance if a doorbell rings, current captioning will say something like "doorbell rings" instead of "ding dong" which detracts from the overall natural viewing experience.

To enhance their perception of video, this paper proposes a dynamic approach to improve current captioning to better assist and inform hearing impaired audiences when watching videos. Contextual caption-

ing will enhance the quality of viewing and interpreting videos for hearing impaired audiences. Compared to existing methods, contextual captioning will address the aforementioned problems by:

- Determining suitable script locations on screen and assigning each character a color for their caption text. Contextual captioning will place text in neutral areas around a given speaker to display the script in a manner that it is obvious who is speaking.
- Highlighting the scripts in real time to better convey speaking pace and elongation of words. For example, think of the way karaoke systems highlight scripts word by word at the pace it is meant to be sung, contextual captioning will emulate this in its own way.
- Conveying variations of volume in sounds by increasing or decreasing the font weight. For instance, if a character is whispering the text will be smaller than when speaking normally and will be larger when a character yells or raises their voice.
- Introducing as many contextually important sound words as possible.

Starting with these main concerns, the existing approach to captioning has a long way to come before it is sufficient for the hearing impaired audience. Also, in many current forms of captioning, there is no signification if the on screen script is a thought or speech and again, a hearing impaired audience must spend the time to determine if a character is speaking or not.

The contribution this proposal aims to make is an enhancement to current captioning, especially concerning hearing impaired audiences. All research aims to enhance current captioning methods to give disabled individuals more equal opportunity to enjoy visual media. Accomplishing this requires exploration of a large host of technologies. To evaluate if contextual captioning is an improvement, a user study will be conducted to compare it to current captioning systems and methods.

2 Background and Related Work

There are two forms of captioning, open and closed. Open captions can be turned on or off by the user, and closed captions cannot. Furthermore, there is a drastic difference between captions and subtitles. Subtitles are direct transcriptions of dialogue only, while captions include description of all sounds on screen, and are geared towards the hearing impaired audience [13]. However, current forms of captioning illustrate text statically and in a fixed region of the screen, which is generally at the bottom, like below in Figure 1.

Since the birth of closed captioning for television in the early 1970's, there has been little effort to make any significant improvements. The first closed captioning, created by the National Television System Com-

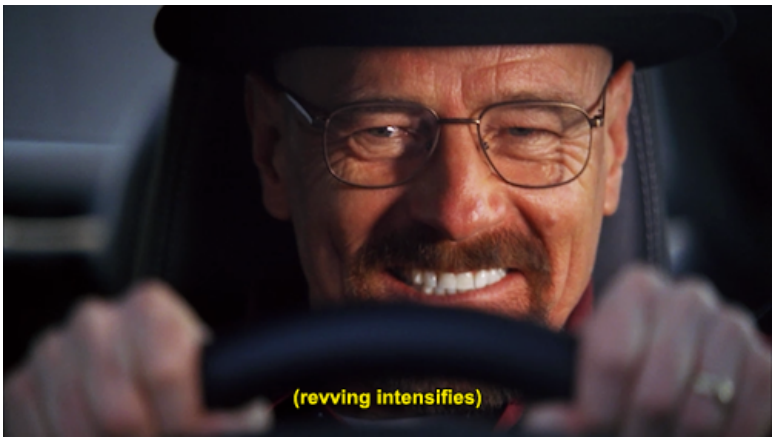


Figure 1: This image from the hit series *Breaking Bad* on Netflix shows how captioning currently conveys sound.

mittee (NTSC) at the First National Conference on Television in Nashville, Tennessee [6], is the base for television captions. For movie theatre viewing, Rear Window Captioning (RWC) is a widely accepted and used system where a discrete reflective panel can be placed in front of a viewer's seat to show captions projected at on the back wall of the theatre. Other systems allow users to add their own captions, like *Amara* [2], which a user can upload a video file and *Amara* will extract the audio and allow the user to add their own captions based off of volume changes. But, before the internet and new forms of multimedia, closed captioning was only a concern for broadcast television. The founding of the National Captioning Institute (NCI) within the Federal Communications Commission (FCC) in the early 1980's offered some relief, but only on broadcast television [12].

The introduction of new forms of media has shifted the landscape. Today, captioning is much more complex. As the internet has become the unifying medium for access to almost all information today, state and federal governments have begun to create laws and regulations surrounding closed captioning of online video. In attempt to expand accessibility standards for online video in 2010, congress passed the Twenty-First Century Communications Video Accessibility Act (CVAA). New FCC regulations that were put into place under this act mandate that all video devices that receive or display video programming transmitted simultaneously with sound, including those that can receive or display programming carried over the internet must provide closed captioning capabilities. Most recently, part of Title III of the Americans with Disabilities Act (ADA) has been expanded to include online spaces of public accommodation, claiming "no individual shall be discriminated against on the basis of disability in the full and equal enjoyment of the goods, services, facilities, privileges, advantages, or accommodations of any place of public accommodation" [9].

In spite of the the vast options for captioning today and the increasing amount of technologies available,

there is not much done in the way of analysis of these systems specifically for hearing impaired audiences. The work of Braverman and Hertzog [4] looked into analyzing language level understanding and not the rate a hearing impaired user can comprehend captions like, Jelinek [8] did. In general, much of the analysis of current captioning for the hearing impaired has concluded that it is difficult for this audience to track and interpret information from a caption in an efficient manner. Therefore, this works' proposed contextual captioning solution will aim to improve on the aforementioned identified issues in current captioning by determining suitable on screen script locations, changing text color, highlighting scripts in real time, displaying contextually important sounds, and better conveying of speaking/noise volume.

3 Technologies and Methods

3.1 Technologies

Deciding how to caption a web-based video is challenging given the vast number of options available and their inherent differences. Initial research found an embedded multimedia player to be necessary for displaying video on a web page, but all players have slightly different settings when it comes to making a caption file actually work [5]. Most web multimedia today requires captioning data to be stored in a side-car file, meaning captions are in a separate file from the video itself [11] [10]. Based off of this preliminary research, QuickTime media player seemed to be the best option because it is easily embedded in a web page and has a great platform for tailoring multimedia experiences. The side-car file for QuickTime is SMIL, which has a massive variety of styling options [14]. Later research found that a new web standard for captioning web video was created soon after the unveiling of HTML 5. The fifth version of HTML includes a <video> tag that allows for a mp4 video file to be directly referenced instead of embedding a media player. This new tag also allows for a caption track to be included in the WebVTT file format [3]. This experiment utilizes HTML's <video> and <track> tags to display the video and render captions in a web page. To deploy the page I utilized the Heroku platform, which created and updated the page with a simple PHP header. The technologies needed are straightforward and not difficult to use. All that is needed is WebVTT for the captions, HTML for the web pages, CSS for styling the pages/captions, PHP for updating the page, and little JavaScript for some nice animation.

3.2 Methods

Concerning conceptual design and styling, some challenge came in defining and using sound-related words on and off camera. Deciding which sounds are contextually important enough for understanding a given

Figure 2: These two images show how sounds are displayed in current forms of captioning (bottom) and how they are displayed in contextual captioning (top).

scene was difficult. Comics and graphic novels have been doing this for so long, it was a great place to start. For captions, as seen in Figure 1, a simple description of the sound does not suffice for the sound itself. Figure 2 shows how captions are displayed with the current industry standard versus this paper's proposed contextual captions. Captions need to include relevant descriptions of sounds and additional information that can greatly enhance a captioned video, but it is important not to congest said video with unnecessary descriptive captions. A caption viewer should not receive any more or any less information than a hearing one. Contextual captions inform the caption viewer of developments of which they would otherwise not be aware. They are to be placed at the center of the screen unless they are specific to a certain character or object.

The most challenge came in addressing uncertainty of a given character's speaking pace. Highlighting the caption text in real time was too ambitious a goal, especially considering the change in web-based video standards with HTML 5. Given more time, WebVTT does have a way to implement the karaoke-style script highlighting like contextual captioning intends. It is not difficult to implement, just very time consuming because of time stamps needing to be hard-coded within the caption text. In other words, for each word in each caption text, a start and finish time stamp must be declared for when a given word should be highlighted. Look to Future Work for more on script highlighting.

4 Experiment Design

In order to determine if contextual captioning is an improvement over the industry standard, user studies were conducted over the span of a week. The experiment was designed so that a participant was randomly assigned one of two videos and its corresponding survey using Google Forms. Each of the two videos and surveys were the exact same, the only difference was the captions. One video contained industry standard for captions and one had this paper's proposed contextual captioning. The random assignment was determined by the millisecond of the day. The participants were contacted via email and sent a link to an informed consent page, which they were then randomly assigned one of the two videos. The survey consisted of seven rating questions, one multiple choice question, and one open-ended comment question. The questions mainly focused on the four aforementioned issues with industry standard captions and how distracting the participant found the captions they viewed.

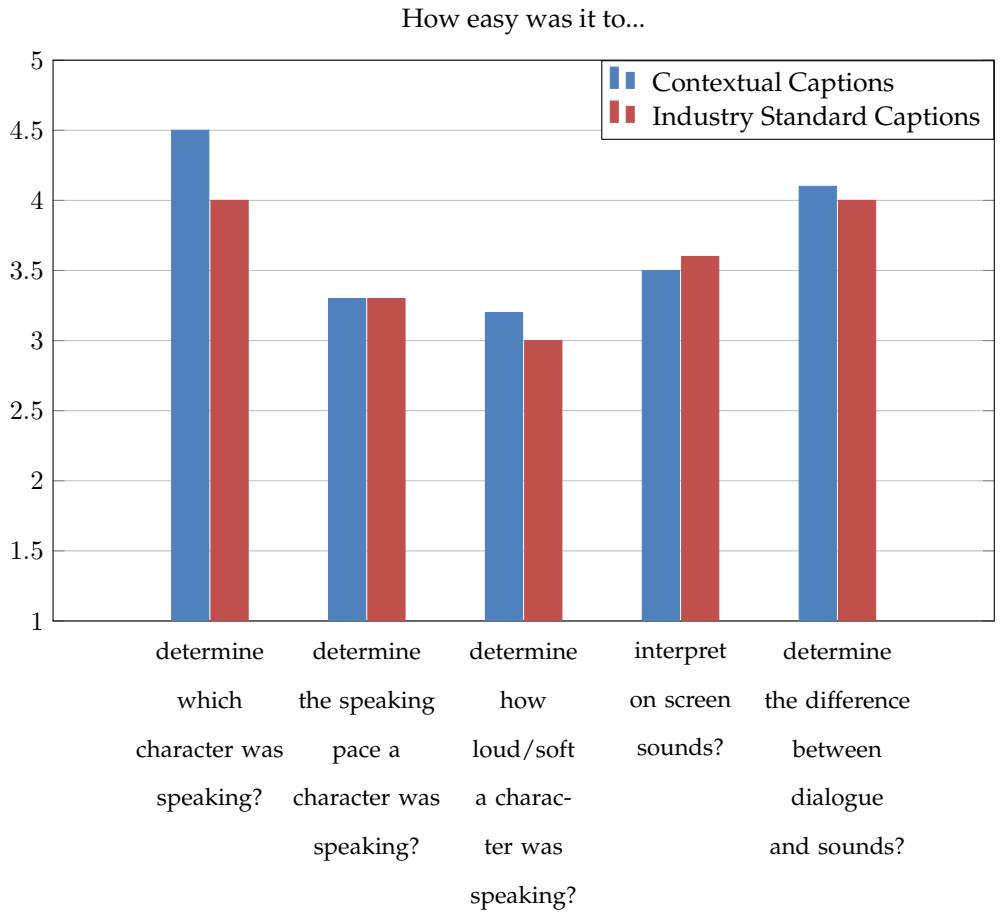
The choice of clip was very intentional. There are multiple characters on screen at a time, lots of hard-to-describe sounds, and one of the characters is masked, making it more difficult to determine his speech and facial expressions. The clip is from *Star Wars Episode 7, The Force Awakens* and mainly consists of dialogue between two characters [1]. Most importantly, the survey needed to be taken by participants that suffer from some degree of hearing impairment and frequently use captions. The familiarity with caption use was essential to gather meaningful data. The Albany chapter of the Hearing Loss Association of America graciously extended a helping hand and provided participants. The co-president of the organization sent two email blasts to their mailing list containing a link to my informed consent page and some brief instructions.

5 Results and Evaluation

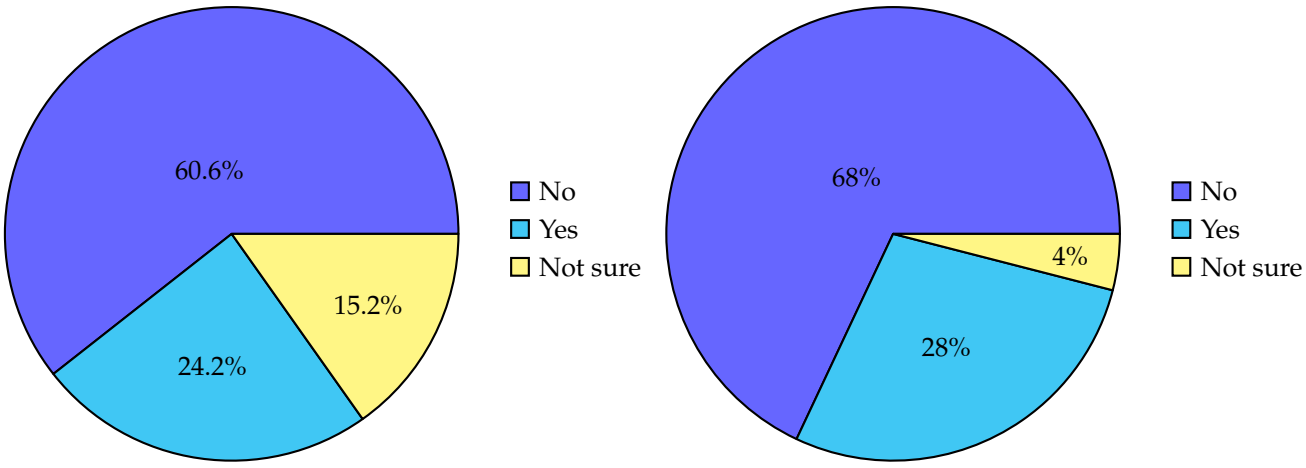
5.1 Results

In total, only fifty-eight participants watched one of the clips and took its corresponding survey. Twenty-five participants watched the industry standard clip and thirty-three watched the contextually captioned clip, and took each related survey. Since the surveys were anonymous, it is impossible to determine what portion of the participants were actually hearing impaired aside from a few of the comments that were left.

The bar graph below shows five of the seven rating questions, which asked a participant to rate each question from one to five based on their experience - one being the worst and five being the best. Look to the Evaluation section for discussion of the results.



As the bar graph above shows, contextual captions performed slightly better than the industry standard on the majority of the questions. The only rating question that industry standard captions achieved a better score in was interpreting on screen sounds. The below pie charts shows the results of the question, "Were the captions at all distracting?". The right chart contains results from the industry standard clip and the left chart contains results from the contextually captioned clip.



5.2 Evaluation

The final question of each survey asks the participant "Would you recommend this form of captioning to a friend or colleague?". The industry standard clip received an average of 6.7 whereas contextual captions received an average of 7.3. Therefore, contextual captioning received slightly higher scores on every rating question except one - determining on screen sounds. This is attributed to the way contextual captioning currently denotes sound, it is too similar to the industry standard. As shown in Figure 2, the differences are very minor between contextual and industry standard captioning. In this specific case, this occurred because too much attention was given to the word choice rather than its styling. Regarding the tied score on ease of determining speaking pace, the results would be drastically different if the karaoke-style text highlighting was implemented. This is also attributed to how contextual captioning currently displays speaking pace. As mentioned above, it is not different enough from the industry standard to give deterministic results. If karaoke-style highlighting were implemented, it may be more distracting because instead of watching what is going on on-screen, a viewer may focus on the text too much.

As the pie charts show, participants found the contextual method of captioning to be more distracting than the industry standard. This occurred because participants may have been too accustomed to the industry standard. By reaching out to a hearing loss organization and obtaining participants with caption familiarity, the vastly different placement on screen that contextual captioning implements may stray too far from the norm. Given a longer clip or entire film, participants would have had more time to adjust to contextual captioning and the methods it employs. A few participants addressed this problem directly in the comments section:

> "I think the contrasting colors of the different speakers in itself was helpful but I'm not sure I liked the placement off to the sides–I think that distracted me more from watching the action on screen. Maybe it's because I'm used to watching captions in the middle but it was a bit jarring."

> "At first the captioning was hard to figure out - usually it is at that bottom of the screen and this was in the middle near the young man. After a bit I could tell who was talking. Putting the sound of what is happening was very helpful."

> "The placement of captions across the screen disrupts the careful compositions created by cinematographers and directors, and could be seen as inappropriately altering the visual qualities of the work for a marginal gain."

This was the most negative feedback contextual captioning received. These participants quite obviously felt that contextual captions strayed too far from the industry standard to the point of disrupting the view-

ing experience. Given more time to familiarize with a clip longer than one minute, viewers would have had an easier time adjusting and may even learn to expect captions to be next to a character or sound, rather than in a fixed position at the bottom of the screen. This is true because many participants did find contextual captioning helpful and an improvement, saying:

> "This form of captioning is so much more clear than the current standard. You can easily distinguish between who is talking, the 'loudness' or emphasis of their voice, and the difference between sounds and voices"

> "It was very diverse and real to life sort to speak as the captioning was tailored to the background. It also was different colors which was great I thought. We are so used to the typical white lettering. It was a good contrast to the black background. I enjoyed participating in this study and seeing this type of captioning and would welcome this in a theater near my home. Thumbs up!"

Therefore, further testing of placement is needed to obtain deterministic conclusions about the script locations. What was very successful was participants found the color change for characters extremely helpful. In addition, there are a few confounding variables in the experiment design that may alter viewing experiences. Viewing the videos in different browsers changed how the caption text was rendered and placed. In the email sent to participants, they were encouraged to use Google Chrome for a consistent viewing experience, but it is not possible to know if they did. Also, not knowing the screen size, participants' distance from the screen, or how familiar a participant was with the video source material may have contributed.

Lastly, there may have also been some response bias for the industry standard clip. Given the blind nature of the clip assignment, a participant may have confused the industry standard clip with the contextual clip, and were therefore more inclined to rate the industry standard clip highly. Thus making it difficult to determine if my independent variable, the type/form of captioning, had any significant impact on a given participant.

6 Future Work

Moving forward, primary goals will be to add karaoke-style script highlighting and as much automation as possible. In terms of karaoke-style highlighting, initially, attempt to use WebVTT's time stamping to hard code highlighting and conduct another user study to determine its helpfulness or distraction. In terms of automation, some patterns have already been identified in the hard coding that can be made into their own functions, but this is not enough. Patterns so far mainly consist of heavily relied on styling options,

like gradually increasing text weight and size, which I can automate with some JavaScript instead of hard coding every time. Although this is minuscule in the grand scheme of automating the system, it serves as a necessary first step. Being able to deploy certain, automated styling scripts from given cues from the audio and/or video files would be the next step, but this requires a lot more research in audio/video encoding and how to parse that information, especially if the video is web-based.

In terms of making sound-related words more dynamic, just as much emphasis needs to be put into styling as the word itself. By possibly surrounding a sound word with specific brackets or fonts as signifiers, it will be easier for viewers to immediately recognize a sound, rather than realizing after already reading the caption. From there, aggregating sound styles and making them a part of automation is a logical next step. After each new caption method is implemented (like karaoke highlighting) it's important to get more participant results in how helpful or harmful the new method may be. Therefore, a more structured experiment design will be needed the next time its sent. Like mentioned above in Evaluation, the confounding variables found after obtaining survey results will be addressed if another survey was to be distributed. Having each participant view the video in the same browser is crucial for consistency in the data. Screen size and distance from the screen are much harder to accomplish because participants would need to come use the same computer rather than their own in the comfort of their home. In the grand scheme of the experiment design, screen size and distance from the screen are not nearly as important as using the same browser, for example.

In conclusion, although contextual captioning received overall higher scores, the difference was too insignificant to be considered a success. There is vast room for improvement still and as visual media starts to dominate our lives more and more, it will be all the more necessary to address this important issue.

References

[1] JJ Abrams. Star wars: The force awakens, 2016.

[2] Amara. Amara, making video globally accessible.

[3] Kyle Barnhart. Web video text tracks (webvtt).

[4] Barbara B. Braverman and Melody Hertzog. The effects of caption rate and language level on comprehension of a captioned video presentation. *American Annals of the Deaf*, 1980.

[5] Courtney Duerig. Understanding closed captioning formats: What you need for your player, 2015.

[6] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. Dynamic captioning: Video accessibility enhancement for hearing impairment. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 421–430, New York, NY, USA, 2010. ACM.

[7] National Captioning Institute. History of closed captioning. 2016.

[8] M. S. Jelinek Lewis. Television literacy: Comprehension of program content using closed captions for the deaf. *Journal of Deaf Studies and Deaf Education*, 2001.

[9] US Department of Justice. Americans with disabilities act title iii regulations. *Americans with Disabilities Act*, 1990.

[10] University of Washington. What is the difference between smil and sami?

[11] University of Washington. What types of closed caption files do video players support?

[12] Deaf News Today. Closed-captioning history.

[13] unknown. Closed captions vs. subtitles. 2009.

[14] W3C. Synchronized multimedia integration language (smil 3.0), 2008.

[15] Wikipedia. Hearing loss. 2016.