Effect of Caption Width on the TV User Experience by Deaf and Hard of Hearing Viewers

Abraham Glasser
atg2036@rit.edu
Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York, USA

Joseline Garcia
joseline.garcia@gallaudet.edu
Information Technology Program
Gallaudet University
Washington, D.C., USA

Chang Hwang
chang.hwang@gallaudet.edu
Information Technology Program
Gallaudet University
Washington, D.C., USA

Christian Vogler
christian.vogler@gallaudet.edu
Technology Access Program
Gallaudet University
Washington, D.C., USA

Raja Kushalnagar
raja.kushalnagar@gallaudet.edu
Information Technology Program
Gallaudet University
Washington, D.C., USA

ABSTRACT

Deaf and hard of hearing (DHH) viewers watch multimedia with captions on devices with widely varying widths. We investigated the impact of caption width on viewers' preferences. Previous research has shown that presenting one word lines allows viewers to read much more quickly than traditional reading, while others have shown that the optimal width for captions is 6 words per line. Our study showed that DHH viewers had no preference difference between 6 and 12 word lines. Furthermore, they significantly preferred 6 and 12 word lines over single word lines due to the need to split attention between the captions and video.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**; **Empirical studies in accessibility**.

KEYWORDS

Accessibility Guidelines, Subtitles, Captions, Deaf, Hard of Hearing

ACM Reference Format:
Abraham Glasser, Joseline Garcia, Chang Hwang, Christian Vogler, and Raja Kushalnagar. 2021. Effect of Caption Width on the TV User Experience by Deaf and Hard of Hearing Viewers. In *Proceedings of the 18th International Web for All Conference (W4A '21), April 19–20, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3430263.3452435

1 INTRODUCTION

Captioned television is essential visual access to auditory information for the 36 million Americans who are Deaf or Hard of Hearing (DHH). Captions can be defined as a text representation of speech in television, in the shape of one or more lines of written text presented on screen, that is in sync with the speech. Access to captioned television has a direct impact on participation in society, as it is everywhere: entertainment, news, political engagement, government, schools, post-secondary education, at-home learning, social engagement, and much more.

However, captioning has not kept up with the shift from broadcast TV to video that can be produced by anyone. The technology and processes for creating captions are fundamentally the same as in the 1980s and 1990s, and do not serve the needs of consumers today. Today, our personal devices have high-quality screens and can support customized captions. The differences between resolutions and viewing size between large screen television displays and small screen phone displays, can influence viewer preferences depending upon the program speaking rate and automatic speech recognition rate. We are in the middle of a disruptive transition to captions that can be viewed anywhere, anytime. These new technologies create different types of caption errors, compared with human captioning techniques that have evolved over 40 years. As a result, there has been much consumer frustration.

There has been research on how fast human eyes can read, in particular the timing of vision. Rapid serial visual presentation (RSVP) is one such methodology. Instead of a passage of text, words are shown one-by-one in quick succession, and the words are always in the exact same location. When viewers concentrate, they can read at a much faster rate measured in words per minute (WPM) as the viewer does not have to move their eyes down lines or passages.

There has not been any prior work on assessing viewer preferences for caption single-word lines for TV broadcasts or online videos. Research has shown presenting one word lines allows viewers to read much more quickly than traditional reading, while other research has shown that the optimal number of words per caption line is around 6 words per line.

We investigated the impact of caption width on viewers' preferences. Our study showed that DHH viewers had no preference difference between 6 and 12 word lines. Furthermore, they significantly preferred 6 and 12 word lines over single word lines due to the need to split attention between the captions and video.

W4A '21, April 19–20, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8212-0/21/04…$15.00
https://doi.org/10.1145/3430263.3452435

2 RELATED WORK

Deaf and Hard of Hearing community have to obtain information through reading text, or they will not be able to understand any content in detail or get confused about what information they are providing. One of the reasons for this confusion is because there is a mismatch between speech and text perception, in that speech perception is speaker synchronized, while text perception is reader synchronized. The reader's experiences in following the captions is grounded in their life experiences, and hence their preferences are shaped by those experiences. Given this mismatch between the speaking and reading speeds and prior reading experiences, captions are usually shown two lines at a time. If the captions consist of two lines with different lengths, the upper line should be shorter in order to ensure that as much of the screen as possible is free so that the viewer does not have to move the eyes unnecessarily.

We investigate alternative methods to present textual information to expedite reading speed, while preserving capacity to follow the captions. Traditional captions can be shown either as a list of short lines, but sometimes also as longer lines shown sentence-by-sentence, and the rapid serial visual presentation or RSVP, first proposed in the 1950s [7] and adapted to reading in the 1970s [5], and that four words per second may elicit performance comparable to traditional text presentation formats [10]. Also, less fluent readers may benefit from RSVP displays [4]. RSVP consists of displaying in sequential order one or more words at a time, and the elimination of saccades should reduce visual fatigue and improve comprehension. The display of one or more words at a time and in sequential order minimizes the eye movements generated during reading, and increasing the attentional focus.

Most speech is in the range of 120-240 words per minute (2-4 words per sec). If captions are shown over two lines then each pop-up line should have around 3-6 words, and each line is shown for 0.5-2 seconds.

Captioning guidelines recommend that caption display time follow the "the six-seconds rule", i.e., a full two-line subtitle should be displayed for six seconds in order for an average viewer to be able to read it [1, 6]. This six-seconds guideline translates to approximately 140–150 WPM or 12 characters per second (CPS), and applies across languages [9]. The display of caption lines for a longer time can increase the viewer's reading time, but decrease the time on following video details. Longer caption display times may benefit viewers by giving them more time to follow video details. But viewers, especially those who are fluent may be able to read the captions more efficiently and not require longer caption display times.

Given that caption guidelines stick with the most common preferences for all readers, they limit the set of solutions that will be considered. Thus the current guidelines for around 6 words per line may miss out on presenting an optimal, elegant solution.

The Described and Captioned Media Program (DCMP) Captioning Key guidelines suggest that each line be limited to 32 characters [1]. A recent questionnaire study with responses from 237 professional captioners from 27 countries around the world shows that captioners work with a maximum of 37 to 42 characters per line. Furthermore, the data show that the number of characters for a full two-lined subtitle has increased from maximum 64 characters in the 1980s to maximum 84 characters [11].

Table 1: Distribution of the 27 stimuli videos.

140 WPM	Video 1	1, 6, 12 words per line
	Video 2	1, 6, 12 words per line
	Video 3	1, 6, 12 words per line
180 WPM	Video 4	1, 6, 12 words per line
	Video 5	1, 6, 12 words per line
	Video 6	1, 6, 12 words per line
240 WPM	Video 7	1, 6, 12 words per line
	Video 8	1, 6, 12 words per line
	Video 9	1, 6, 12 words per line

3 RESEARCH QUESTION

For DHH college educated viewers who rely on captioning, what is the most comfortable number of characters or words per line for captions? That is, what is the most comfortable number where it does not feel like too many or too little words on a line?

4 RESEARCH METHODS

We collected different videos and calculated the speeds. Each video is approximately 30 seconds long. We used Aegisub software [2] to measure the CPS for each line, and then used that to estimate the average WPM for the video. We have three different speeds: 140 WPM, 180 WPM, and 240 WPM. For each of those, we have three different videos with the same average WPM.

Next, we need to have various caption width for the videos. We re-recorded each video, and edited the subtitles so that we had 3 variations of each video: 1 word per line, 6 words per line, and 12 words per line.

Variables such as font size and typeface were not manipulated and were kept constant during the whole experiment and across the conditions. So the only difference resided in the way the text was displayed.

In total, we had 27 different videos: three videos for each of the three WPM-levels, and 3 instances of each of those videos, with different numbers of words per line. Table 1 illustrates a summary of this. In the figures and this paper, "140_1" refers to a video at 140 WPM with the one-word captions, and so forth.

We used an between-subjects design, in which the participants were arbitrarily allocated to each condition. We counterbalanced the order in which we would show these to participants, using a Latin-square method [3]. We adopted this approach to minimize the influence of other factors on the results, such as the content or speaking differences in the videos themselves. The only factors we focused on were the caption speed and caption line width.

We recruited participants from the Gallaudet University through flyers and word-of-mouth. After explaining the basic purpose and principles for the experiment, we asked the participants to complete a preliminary survey to ensure that they were eligible by meeting the requirements of regularly watching live TV with captions for speech accessibility. Then the participant signed a consent form and filled out a demographic questionnaire and then took the study.

During the study, each participant watched nine different clips, to ensure that they saw each possible combination of the three speeds (140, 180, and 240 WPM), and each caption width (1, 6, and 12 words per line). After each clip, which were 30 seconds long, participants answered five Likert-scaled questions:

(1) How easy was it to follow the video action?
- (1=Too difficult, 5=Very easy)

(2) How easy was it to follow the video captions?
- (1=Too difficult, 5=Very easy)

(3) What do you think about the width of the captions on this video?
- (1=Too short, 5=Too long)

(4) How fast was the captioning in this video?
- (1=Too slow, 5=Too fast)

(5) What do you think about the caption style on this video?
- (1=Very uncomfortable, 5=Very comfortable)

5 RESULTS AND DISCUSSION

We recruited a total of 14 participants for the study. All were college students who were fluent English caption readers. The average age across all participants was 27 years old, with a standard deviation of five years. Seven participants identified as Male and seven identified as Female. Thirteen participants identified as Deaf, and one identified as Hard of Hearing. The majority of the participants watched live TV at least every other day, and at least 1-2 hours a day on average.

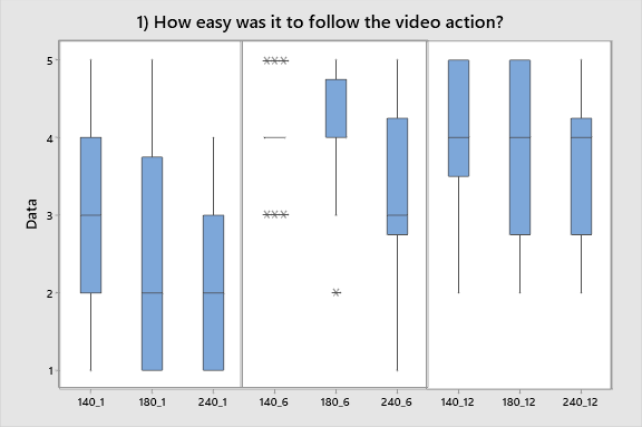

Figure 1: Boxplot summaries of participants' responses to the likert-scaled question, "How easy was it to follow the video action?"

Figure 1 shows the participant responses to the first Likert-scaled question, which is "How easy was it to follow the video action?" A response of 1 means it was too difficult to follow, while 5 means it was very easy to follow. For one-word length, it was harder for participants to follow the video action. Participants indicated it was neutral at 140 WPM, but shifted to difficult at 180 and 240 WPM. Since the participants were not used to the one-word length, perhaps at 180 and 240 WPM, the captions were distracting from the video action since they basically are flashing single words.

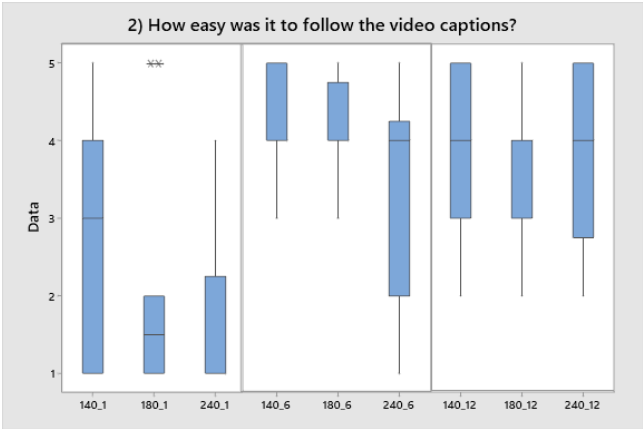

Figure 2: Boxplot summaries of participants' responses to the likert-scaled question, "How easy was it to follow the video captions?"

In Figure 2, the participant responses to the second question, "How easy was it to follow the video captions?" also indicate that it was difficult to follow the video captions themselves. For 140 WPM, the median was also 3, and the spread was similar to figure 1. However, at 180 and 240 WPM, responses indicated it was more difficult to follow the video captions. At 140 WPM, the one-word caption lines take longer to change, so it is easy to glance at the captions, read, and go back to the video action quickly. The responses were very spread out at this speed. However, at 180 WPM, every participant responded with either 1 or 2, which indicated that it was too difficult to follow the video captions. This suggests that, if one-word captions are used, it would be better at slower speeds.

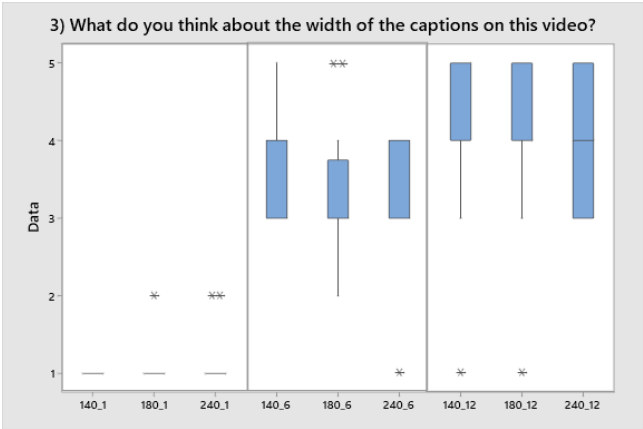

Figure 3: Boxplot summaries of participants' responses to the likert-scaled question, "What do you think about the width of the captions on this video?"

The same pattern is clear in the rest of the questions. In figure 3, it is strongly evident that participants thought the width of the captions was too short, as all the responses were 1 or 2 across all

speeds, while only a couple of responses were below 3 for each the six-word and twelve-word lines.

We believe that RSVP might be more suitable for very short texts when split attention is not an issue; future studies should investigate this.

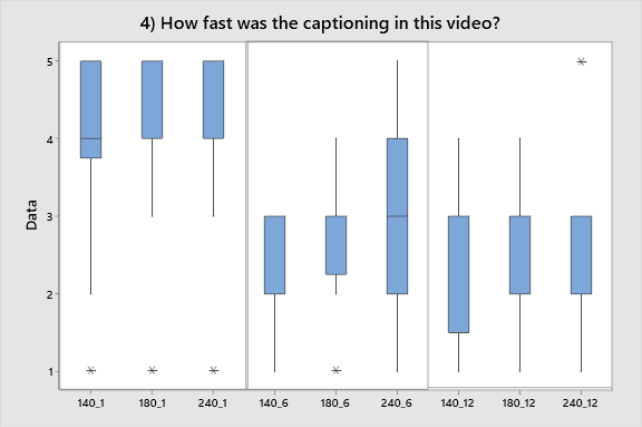

Figure 4: Boxplot summaries of participants' responses to the likert-scaled question, "How fast was the captioning in this video?"

In Figure 4, it is also evident that participants thought the captions were too fast. At each of the WPM levels, a majority of participant responses to the question "How fast was the captioning in this video?" were 4 or 5 at the one-word level, which means they were too fast, however at the six and twelve-word level, just about all the responses were at or under 3, which indicates they were not fast. The one-word style of the captions seemed too fast for participants to follow while they also had to split their attention to the video content. At the six or twelve-word level, participants thought the captions were slow, even though they followed the same speed in WPM. Other things equal, understanding of captions is maximized when audiovisual issues are minimized.

In figure 5, participants respond to the question "What do you think about the captioning style on this video"? There is a statistically significant difference between 140_1 with both 140_6 and 140_12 ($p<.001$ and $p<.01$ respectively), 180_1 with both 180_6 and 180_12 ($p<.001$ and $p<.01$), and 240_1 with both 240_6 and 240_12 ($p<.01$ and $p<.0001$).

This shows evidence that the one-word length captioning style is more uncomfortable than the six and twelve-word lengths. For the six and twelve word lengths, however, there was no significant difference in either one; participants were equally, roughly neutrally comfortable with both.

6 CONCLUSION

Caption presentation width and time of presentation influence the time spent for viewers to comfortably follow both the captions and video. Based on the responses to Questions 1 through 5, it seems that viewers adjust time in reading the captions according to caption width while leaving sufficient time to follow the video. For 12-word caption lines that were displayed for a longer time,

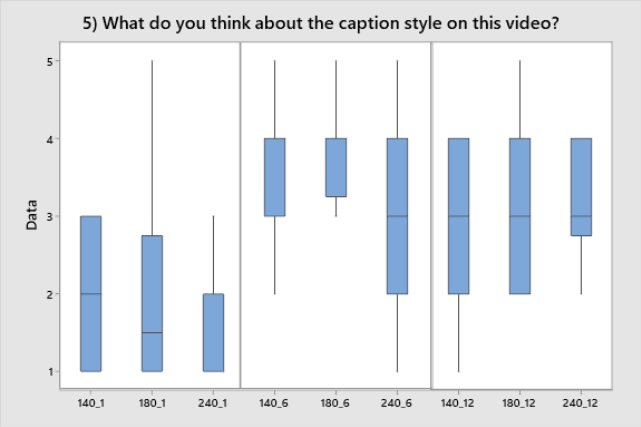

Figure 5: Boxplot summaries of participants' responses to the likert-scaled question, "What do you think about the caption style on this video?"

the viewer comments suggested that the viewers could read more slowly, while being able to quickly switch to video and back. Conversely, it appeared that when the viewers followed the 6-word caption lines that were displayed for a shorter time, the viewers commented that they could read them more quickly while also dividing attention on the video. This is consistent with Luyken's findings: If the ratio between the amount of text and the time of exposure remains constant, the resulting reading speed will also remain constant. [8].

7 ACKNOWLEDGMENTS

We thank Norman Williams for his help in engineering the study stimuli.

We thank the National Science Foundation, grant #1757836 (REU AICT) and the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR #90DPCP0002).

NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this paper do not necessarily represent the policy of NIDILRR, ACL, HHS, and you should not assume endorsement by the Federal Government.

REFERENCES

[1] 2021. Described and Captioned Media Program (DCMP) Captioning Key. https://dcmp.org/learn/captioningkey
[2] Aegisub. 2021. Aegisub. https://github.com/Aegisub/Aegisub
[3] James V. Bradley. 1958. Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528. http://www.jstor.org/stable/2281872
[4] Hsuan-Chih Chen. 1986. Effects of reading span and textual coherence on rapid-sequential reading. *Memory & Cognition* 14, 3 (1986), 202–208. https://doi.org/10.3758/bf03197693
[5] Kenneth I. Forster. 1970. Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics* 8, 4 (1970), 215–221. https://doi.org/10.3758/bf03210208
[6] Pablo Romero Fresco. 2009. More haste less speed: edited versus verbatim respoken subtitles. *Vial-vigo International Journal of Applied Linguistics* (2009), 109–133.
[7] Luther C. Gilbert. 1959. Speed of processing visual stimuli and its relation to reading. *Journal of Educational Psychology* 50, 1 (1959), 8–14. https://doi.org/10.1037/h0045592

[8] Georg-Michael Luyken, Thomas Herbst, Jo Langham-Brown, Helen Reid, and Hermann Spinhof. 1991. *Overcoming language barriers in television dubbing and subtitling for the European audience*. Vol. 13. Europ. Inst. for the Media.
[9] José Martí Ferriol. 2013. Subtitle reading speeds in different languages : the case of Lethal Weapon. *Quaderns* 20 (01 2013), 201–210.
[10] Alexandra B. Proaps and James P. Bliss. 2014. The effects of text presentation format on reading comprehension and video game performance. *Computers in Human Behavior* 36 (2014), 41–47. https://doi.org/10.1016/j.chb.2014.03.039
[11] Agnieszka Szarkowska, Izabela Krejtz, Olga Pilipczuk, Łukasz Dutka, and Jan-Louis Kruger. 01 Dec. 2016. The effects of text editing and subtitle presentation rate on the comprehension and reading patterns of interlingual and intralingual subtitles among deaf, hard of hearing and hearing viewers. *Across Languages and Cultures Across Languages and Cultures* 17, 2 (01 Dec. 2016), 183 – 204. https://doi.org/10.1556/084.2016.17.2.3