Effect of Occlusion on Deaf and Hard of Hearing Users' Perception of Captioned Video Quality

Anonymous for Review

Anonymous for Review

Commented [MH1]: I will share a new Word Document with you in a moment. Please upload it to Sharepoint, then share it with VIEW and EDIT access to all of us -- share with me at both emails: mphics@rit.edu and matt.huenerfauth@rit.edu

Commented [MH2R1]: EMAIL SENT

ABSTRACT

While the availability of captioned television programming has increased, the quality of this captioning is not always acceptable to Deaf and Hard of Hearing (DHH) viewers, especially for live or unscripted content broadcast from local television stations. Although some current caption metrics focus on textual accuracy (comparing caption text with an accurate transcription of what was spoken), other properties may affect DHH viewers' judgments of caption quality. In fact, U.S. regulatory guidance on caption quality standards includes issues relating to how the placement of captions may occlude other video content. To this end, we conducted an empirical study with 27 DHH participants to investigate the effect on user's judgements of caption quality or their enjoyment of the video, when captions overlap with an onscreen speaker's eyes or mouth, or when captions overlap with onscreen text. We observed significantly more negative user-response scores in the case of such overlap. Understanding the relationship between these occlusion features and DHH viewers' judgments of the quality of captioned video will inform future work towards the creation caption evaluation metrics, to help ensure the accessibility of captioned television or video.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in accessibility;** Empirical studies in HCI;

KEYWORDS

Occlusion, Stimuli, Caption, Metric

1 Introduction and Background

In recent years, the transcription accuracy of captions appearing on television programming has improved, a phenomenon that some researchers attribute to the use of caption-evaluation metrics which allow efficient assessment of the accuracy of captioned television broadcasts [29]. However, beyond the issue of whether the captions are an accurate transcript of the words spoken by individuals in the video, there are many other factors that are known to negatively affect DHH viewers' experience with captioned video, including whether captions occlude other visual content [10, 41]. The placement of captions can pose unique challenges for DHH viewers, such as reducing the overall amount of information viewers can perceive from the visual content [20, 21] or making it difficult to use speechreading if the face of the person onscreen is blocked by a caption [37]. Captions can also block other important visual information content, e.g. non-verbal behaviors that indicate a speaker's emotional state; in addition, captions can block other onscreen text, e.g. headlines or scrolling "news tickers" on television news broadcasts.

Providing captions for spoken content is essential for providing full access to information contained in television programming, e.g. from TV news, talk shows, classes, meetings, and other sources. For instance, real-time captioned news is vital for providing DHH viewers access to critical information about their local communities, nationwide events, or emergencies [3]. Many specialized software and commercial vendors provide real-time captioning services for live television programming spanning news, current affairs, and sports [28]. There are many users of captioned programming, including people who are Deaf or Hard of Hearing (DHH), who constitute a large proportion of society. Over 360 million people worldwide experience hearing loss [9], and 15% of the U.S. adults are Deaf or Hard of Hearing (DHH) [8]. However, there has been little prior research studies with DHH participants to investigate how various visual properties of captions influence their judgment of video caption usability.

Throughout this paper, we use the term "features" to refer to the aspects or properties of captioned video that may contribute to its quality. For instance, some prior research has investigated how DHH individuals' judgements of the

quality of captions may be influenced by: incorrect transcription of speech into text [32], the latency of the caption relative to the timing of speech [33], font size or color in captions [5, 7], and other features. Furthermore, we use the term "metrics" to refer to some formula or algorithm that can produce a numerical score to represent the quality of a captioned video, whether it requires some human judgements or is calculated in a fully automatic manner. Thus, a metric may consider various features, and research on the relationship between features and the judgements of DHH viewers is foundational to deciding to incorporate particular features into a metric.

While there exist standards and regulations in many countries about providing captioning during television programming, e.g. Federal Communication Commission guideline [15], there is evidence that DHH viewers are not fully satisfied with the quality of the captioning, e.g. for live or unscripted television programming in smaller U.S. television markets [4, 27]. To enable regulatory agencies or others to monitor the quality of captioning in various settings, metrics are needed that can efficiently and accurately evaluate television captioning quality. While evaluation studies with DHH participants can be seen as a gold standard for such assessment, more automatic metrics would enable more frequent and pervasive monitoring of quality, as long as these metrics are well-correlated with the judgements of people who are DHH.

Prior automatic metrics for evaluating captioned television programming have largely focused on features relating to transcription accuracy, e.g. Word Error Rate (WER), Named Entity Recognition (NER), Closed-Caption Evaluator, Automatic Caption Evaluation (ACE) [1, 32, 4, 23]. Some metrics have considered latency issues, i.e. detecting when the timing of the appearance of caption text does not align temporally with the timing of spoken words [30-36]. However, there are emerging trends in the field of computer vision, which may enable such metrics to consider a new set of features, which we investigate in this paper. As the accuracy of automatically identifying people or text in videos increases, it will become possible to automatically calculate occlusion features, i.e. whether a caption blocks information that appears at a particular location and time in the video, e.g. a speaker's face or some onscreen text.

Prior research has suggested that such occlusions are a concern among DHH viewers of captions [10], and in this paper, we conduct a two-part experimental study to examine whether two such occlusion features (whether captions block portions of a speakers face, whether captions block onscreen text) influence DHH viewers' judgements about caption quality. Prior to incorporating such features into existing caption-evaluation metrics, basic research of this nature is necessary, to determine how they may affect DHH viewers' judgements of caption quality. Specifically, in our study, DHH participants indicated "how useful" captions were and how much they "enjoyed watching the video with the caption," when viewing videos that varied according to these features. Thus, the contributions of this paper are empirical: We provide evidence that both factors have a significant effect on DHH viewer's judgements of caption quality. These findings provide motivation for future work into how to calculate such features automatically, for incorporation into caption-quality metrics.

2 Related Work

As discussed above, there are regulations in many countries about the provision and quality of captioning for television programming. For example, in the U.S., the Federal Communication Commission (FCC) oversees the execution of laws and creates regulation on quality standards for television closed captioning. However, many local television broadcasters face challenges in providing high-quality captioning to meet guidelines provided by regulators for all their live or near-live programming, with some using cost-savings approaches based on experienced caption providers or semi-automated workflows [15]. The resulting errors in caption text, the resulting latency in caption appearance, and other factors may negatively affect the experience of viewers. Thus, this situation motivates the need for metrics, which can be efficiently calculated, to evaluate the quality of captioned television content, to predict how DHH viewers may perceive the overall quality of the captioned video.

2.1 Existing Metrics of Caption Quality

A variety of metrics (both automatic and some which require human judgements) have been proposed and used for the evaluation of television captioning quality, but these metrics have largely focused on the issue of text transcription

accuracy and certain related features. For instance, Word Error Rate (WER) is the standard approach for evaluating automatic speech recognition systems [1], and this metric simply penalizes individual words that have been incorrectly inserted, deleted, or replaced -- when comparing what was actually spoken (the "reference" text) and what the captions displayed (the "hypothesis" text). The Named Entity Recognition (NER) metric [32] is a semi-automated metric that requires human experts to label the severity of individual errors in the text, to calculate an overall error score for a text. National Center for Accessible Media (NCAM) introduced a semi-automatic caption evaluation metric called the Closed Caption Evaluator, which is another weighted version of WER [4]. A recently proposed version of this metric is fully automatic, and it uses automatic speech recognition to analyze the speech in video broadcasts and then to compute the caption error using a statistical model [4]. While not proposed for evaluating television captioning (but rather for real-time captioning of live meetings), Kafle and Huenerfauth introduced the Automatic Caption Evaluation (ACE) metric, which uses an automatically calculated word-importance model [23]. This model considered the predictability of individual words, as well as the semantic distance between the reference and hypothesis word [23]. Furthermore, ERRICSON and BBC research unveiled an approach to reduce latency in live captioning. Specifically, they focused on reducing encoding and compensating time during broadcast live program [31]. On the other hand, a machine learning model has been introduced to detect latency between audio and caption [30].

Although various metrics like those above have been proposed for evaluating the quality of captioned television programming, regulations often include provisions that captions should not only have high transcription accuracy but also have other desirable properties. For instance, regulation from the U.S. Federal Communications Commission has included provisions that captions be not only textually verbatim but also visually sound: Specifically, the caption should be complete, should be synchronously displayed with the speech, and should not conflict with any salient visual information [16] -- in other words, it should not block other important visual content. However, these issues have not previously been included in prior proposed automatic metrics of caption quality.

While teams of human judges could view samples of captioned television programming to determine when some of these issues may be occurring, recent advances in several fields have made it possible to create software that can automatically process videos to identify which person in a video is speaking, or when captions may be blocking the faces of people in the video. For instance, in computer-vision field, researchers have created technologies which can be used for detecting human faces and onscreen salient text in videos. Some of these technologies include multi-frame fusion-based face recognition [39], natural scene text detection [40], and onscreen caption detection and type recognition [38]. Since it seems possible to soon identify these features in videos, research is now needed on whether these properties do influence DHH viewer's judgements of the quality of videos – and to what degree.

2.2 Features Affecting DHH Viewer's Perception of Caption Video Quality

In addition to the features of text transcription accuracy and latency (discussed in the context of automatic metrics above), there has been significant prior experimental research to investigate how various other aspects of captions may affect DHH viewer's judgements of their quality. For instance, previous work has investigated the effect of latency between caption appearance and speakers' speech, in real-time captioning circumstances [24, 25] and identifying the current speaker in a panel discussion [17]. Prior work has also examined how inserting correct punctuation or pauses during the captioned video can benefit DHH viewers and increase readability [19, 36]. Other researchers have investigated how DHH users' subjective impression of the readability and quality of captions is influenced by aspects of caption appearance, e.g. styles, font, and background [5, 12]. Prior experimental studies have also revealed that proper segmentation (caption boundaries aligning with syntactic boundaries) can improve caption readability [28, 36]. Finally, caption speed has also been found to affect DHH viewers' comprehension of captions [11]. In addition to research on these various features above, some prior work has been even more closely related to the focus of our study, i.e. on captions visually occluding salient video content, as discussed below:

2.2.1 Captions Occluding different part of onscreen Speakers' Face

As discussed above, prior research has found that text-visibility in captions and video-content being blocked by captions are common concerns among DHH viewers [10]. While some recent televisions support users re-positioning captions to a different location upon request [12, 41], this is a relatively new feature, and it is unclear how often DHH

users would actively use their remote control to change caption locations while viewing television programming. Other recent work has examined dynamically varying the placement of captions onscreen [22], in accordance with the underlying video content; however, this technology does not yet avoid occlusions with important video content and is still being evaluated with DHH viewers [35, 2]. While even captions that remain in one location on the screen have the potential to block important content onscreen, as new dynamic placement technologies emerge, there may be an even greater possibility for visual occlusions across a wide range of the video region.

The concern here is that prior research has found that while captions are essential for providing access to spoken content for DHH viewers, they have the potential to reduce the amount of information DHH viewers perceive from other visual content on the video, e.g. facial movements of the speaker or onscreen text [20, 21]. Captions blocking the face of the current speaker is a concern, as some DHH viewers may use speech and oral-based communication, e.g. performing speechreading while looking at the mouth of the speaker [37, 34]. In addition, a prior experiment showed that even when an onscreen interpreter is present, DHH users still focus their gaze on a speakers' mouth for 12% of total television program time [37]. Thus, captions that block the mouth of the speaker may hinder the understandability and enjoyability of captioned videos for such viewers. In addition, the facial expressions of the speaker may enable the viewer to understand the speaker's emotional state, as prior work has established that emotions are expressed through verbal and non-verbal forms, including body posture, facial expressions (e.g. raising or lowering the eyebrows), eye gaze, and etc. [26]. Thus, if captions block any of these portions of the body, the emotional state of speakers may be less apparent to DHH viewers. Given this prior work, in our study described below, we investigate the impact of captions blocking the eyes or mouth of the speaker.

2.2.2 Captions Occluding different onscreen text

Researches have looked at the reception of integrated titles among viewers stressed on the analysis of the content of a video before deciding on where to place a caption. They emphasized that together with other factors such as the speed of speech acts, static versus active nature of the scenes, and visibility and number of speakers, the existence, number, and layout of onscreen text elements should also be considered before placing integrated titles [13]. Captions overlapping with text onscreen can be particularly problematic for live news broadcasts, as captions can hinder the ability of DHH viewers to read textual information transmitted as part of the video itself. This text content may include the name of the person who is speaking during a news interview, the headline of the current story, or the news ticker at the bottom of the screen, which often features additional facts or headlines for other stories. A prior experimental study revealed that DHH users focused on onscreen text 7% of total TV program time [37]. We found no prior research that had investigated the effect of captions occluding onscreen news text on the DHH viewers' judgments of captioned video quality. Recently, researchers have proposed methods that can detect text that appears in a video, either when this text appears in the real world and is simply captured by the video camera (as in the case of a real-world sign that is within the video frame) as well as text that has been added to a video image digitally (whether static or horizontally scrolling) as in a live news transmission [38, 40]. Given these advancements, future metrics that assess the quality of a captioned video could penalize captions that block onscreen text, but research is needed to understand how such occlusion affects the judgements of DHH viewers.

Commented [AAA(S3]: This paper https://dl-acm-org.ezproxy.rit.edu/doi/10.1145/2745197.2745204, reveals caption placing dynamic placement pose User Experience challenges for viewers.

Commented [AAA(S4R3]:

3 3 Research Questions

4 As discussed above, prior work on automatic metrics of caption quality has not yet integrated occlusion features, i.e. information about the degree to which captions block other onscreen visual content that appears, potentially ephemerally, at a specific place and time in a video. As technologies emerge for automatically identifying speakers' faces or onscreen text in videos, there is a need to understand how occlusion of these forms of visual information may affect DHH viewer's judgement of the quality of captioned video. Therefore, in this study, we experimentally evaluate how variations in two such features may affect viewers; judgments of video quality, in the following research questions:

RQ1: Are DHH viewers' subjective judgments about the usefulness and enjoyability of captioned videos affected by (a) whether captions overlap with the onscreen speaker's eyes and (b) whether captions overlap with the onscreen speaker's mouth?

RQ2: Are DHH viewers' subjective judgments about the usefulness and enjoyability of captioned videos affected by (a) whether captions overlap with onscreen text containing the current news headline and (b) whether captions overlap with onscreen text about other news headlines?

5 4 Methodology and Results

Our experiment consisted of one-hour appointments with a set of DHH participants, with each appointment partitioned into two time-segments: In the first segment, we conducted a study to investigate RQ1, and in the second, we conducted a study to investigate RQ2. This section provides an overview of both studies, beginning with details that were common across both studies. Later, individual sub-sections below focus on the details that are unique to each of these studies.

4.1 Study Design and Question Items

For both studies, a website was developed to display to participants several videos with different variations of occlusion features, and participants responded to questions, to provide their subjective impression of the quality of the videos displayed. For our first study, participants viewed three stimuli videos on a webpage in a side-by-side manner, in which captions: (1) overlapped with speaker's eyes, (2) overlapped with speaker's mouth, or (3) did not overlap with speakers face at all. For the second study, three stimuli videos were shown, in which captions: (1) overlapped with onscreen text displaying the headline of the current news story, (2) overlapped with a scrolling news ticker displaying headlines for other news stories, and (3) did not overlap with onscreen text at all.

After participants watched all three videos on the web page individually, they responded to two subjective scalar questions, of which one (Question 1) was adapted from [23].

How useful did you find the captions? Participants were asked to respond to this question on a five-point scale from 'Not Useful' to 'Very Useful.' For the remainder of this paper, this question may be briefly referred to as the "Usefulness" question.
Did you enjoy watching the video with the caption? Participants responded to this question on a standard five-point Likert-scale from 'Strongly Disagree' to 'Strongly Agree.' For the remainder of this paper, this question may be referred to as the "Enjoyability" question.

At the end of each sub-study, two open-ended items, which were adapted from [23], were asked:

- **Tell us what you liked about the captions in the videos.**
- **Tell us what you disliked about the captions in the videos.**

Our analysis in this study was primarily quantitative in nature, with a focus on statistical difference testing of responses to the two scalar items. Responses to these final open-ended items was considered at the conclusion of our quantitative analysis, to shed additional light on why participants may have responded to scalar questions as they did. From this perspective, some of this open-ended feedback from participants is presented in the Discussion (section 6), where we review how our results address each research question.

4.2 Data Collection Procedure

This study was originally planned to be conducted as an in-person study, in which a researcher would sit with participants to introduce the study and answer questions, in American Sign Language or spoken English, depending upon the participant's communication preference. The participants viewed the stimuli videos on a web page on a computer, and they responded to questions by writing responses on a paper answer sheet. However, partway through collecting data from our 27 participants, the experiment had to move to an online remote format, due the need to maintain social distancing during the COVID-19 pandemic. Therefore, while for the first 9 participants, the responses to these questions were taken on paper, for following 18 participants, these questions were embedded together with video stimuli in a survey hosted on SurveyMonkey. Our study had been approved by the university Institutional Review Board (IRB), and the modification of the study for online remote participant was also approved by the IRB prior to the final 18 appointments.

We conducted the **in-person segment of the study** in our lab. A researcher started the experiment with participants by obtaining signatures on the informed consent form, and then participants filled out a demographic questionnaire. Next, the researcher showed participants the website (containing stimuli videos) that we had created and provided brief instructions about the study procedure. Then, a questionnaire form was handed over to the participants, consisting of the scalar and open-ended questions (described above), and participants responded to each set of questions after watching each video stimulus.

For the **remotely conducted segment of the study**, a researcher sent an informed consent form to our participants through email, which participants read and reviewed, prior to a videoconference meeting between the researcher and the participant. Participants responded to a demographic questionnaire, which was presented as a Google Form. The researcher then sent the participant a link to the experiment, hosted on SurveyMonkey, which contained both the stimuli videos and the corresponding questions for each. We added a sample page at the start of the survey to familiarize our participants with the format of the study, to facilitate the researcher explaining the study procedure, which had been easier in the in-person format, since the researcher could point to on-screen elements of the survey.

4.3 Recruitment and Participants

Participants were recruited by posting an advertisement on social media websites. The advertisement included two key criteria: (1) identifying as Deaf or Hard of Hearing and (2) regularly using captioning when viewing videos or television. Participants received $40 cash compensation for either the in-person or the remotely conducted hour-long study conducted using a video-conferencing. A total of 27 people participated in the study including 12 females, 14 men, and one non-binary, aged 18 to 55 (median = 25). 19 of our participants identified as deaf and 8 identified as hard of hearing. All our participants except 2 reported regularly using American Sign Language at home or work. 18 of our participants reported that they began learning ASL when they were 9 years old or younger. The remaining participants reported using ASL for at least 2 years and that they regularly used it at work or school.

4.4 Study 1: Face Occlusion

For the first time-segment of the experiment appointment, we conducted our "face occlusion" study, to investigate how captions overlapping with the onscreen speakers' face during live captioned TV programming may affect DHH viewer's judgment of caption quality. There were three different placements of the caption shown during this study:

captions overlapped with the speaker's mouth,
captions overlapped with speaker's eyes, and
caption not overlapped with the speaker's eyes or mouth.

For this study, we created nine video stimuli, based on video sources collected from the YouTube distribution channels from mainstream television news agencies. Each of these video stimuli was 30 seconds long, and it consisted of a news broadcast with a single individual speaking. We avoided videos related to any sensitive, trending, or polarizing issues, in an effort to keep the content as neutral as possible. Our rationale for this selection was that videos containing

these issues might lead to divergent reactions among participants. We truncated each video to the desired length using FFMPEG [14], an open-source video-editing tool. We extracted the caption files for each video, which consisted of Advanced Substation Alpha (".ass") files. We manually inspected each caption file to ensure that there were no word omissions or other errors in the text, to prevent errors in text quality from influencing DHH viewers' judgements of the captioned video. We manipulated the settings within the caption file, to adjust the placement of the caption on the video. For the condition in which the captions overlapped with the speaker's mouth, we ensured that the overlapping occurs for the entire length of the video. Similarly, for "overlapped with eyes" condition, we ensured that the caption overlaps with speaker's eyes throughout the duration of the stimuli video. Finally, we embedded the caption file in the stimuli video, and we created three sample videos, for each condition, using the same source video.

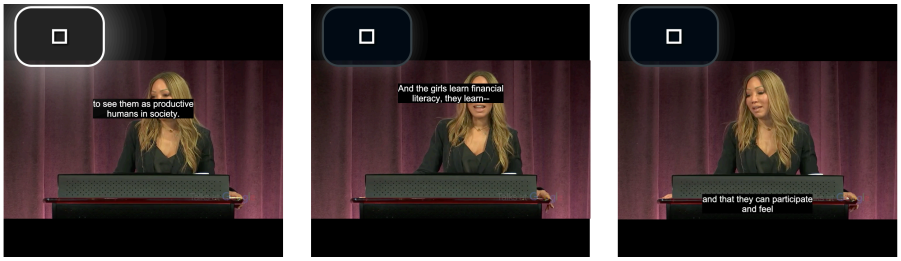
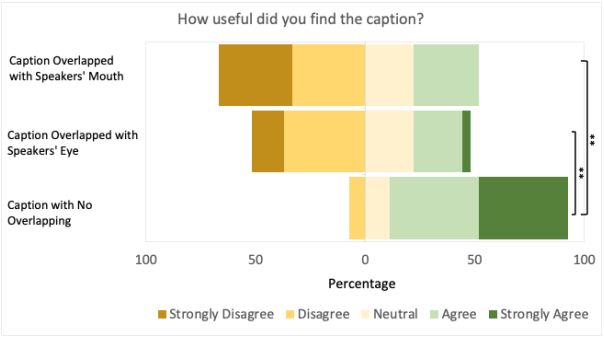


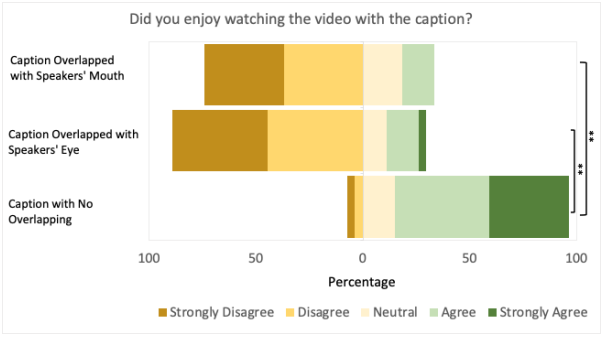
Figure 1: Sample images of video stimuli which are shown to the participants during Study 1.

Participants viewed three videos, with captions placed on different parts of the screen (speaker's eyes, speaker's mouth, and at the bottom), side-by-side, and they answered two scalar questions for each video. Figure 1 shows the

placement of captions on the screen in each of the three videos. A Greco-Latin schedule was used to determine the left-to-right placement of the videos and their assignment to conditions to video stimuli.



(a)



(b)

Figure 2: Participants' subjective scalar responses for videos in each of the three conditions (no overlap, overlap with speaker's mouth, overlap with speaker's eyes) in the Face Occlusion study, for (a) How useful did you find the caption? (b) Did you enjoy watching the video with the caption? Double asterisks ** mark significant pairwise differences (p<0.01).

Figure 2 displays a divergent stacked bar graph (with the neutral response item plotted on the midline of the x-axis) for responses to the "usefulness question" (Q1) and "enjoyability question" (Q2), for the Face Occlusion study, across the three conditions. All significant pairwise differences are indicated with double asterisk (**) in the figure if the p-value is less than 0.01. The statistical analysis performed for the two questions is described below.

To evaluate the responses to the "usefulness question," a Wilcoxon Signed-Ranks test was used. The results indicated that participants found captioned videos with no overlapping (Median=4) more useful than the videos in which the caption overlapped with speaker's eye (Median=2), (Z= -4.014, p<0.0001). In addition, participants found videos with no overlapping to be more useful than videos in which the caption overlapped with the speaker's mouth (Median =2.5), (Z= -3.695, p<0.001). For the "enjoyability question," a Wilcoxon Signed-Ranks test revealed that DHH participants found captioned videos with no overlapping (Median =4) more enjoyable than videos in which the caption overlapped with speaker's eye (Median =2), (Z=-4.106, p<.0001), or than videos in which captions overlapped with

the speaker's mouth (Median =2), (Z= -4, p<.0001). For either question item, there was no significant pairwise difference between responses for the "overlap with eyes" condition and the "overlap with mouth" condition.

4.5 Study 2: Text Occlusion

For the second time-segment of the experiment appointment, we conducted our "text occlusion" study, to investigate how captions overlapping with onscreen text during live captioned TV programming may affect DHH viewer's judgment of caption quality. There were three different placements of the caption shown during this study:

Captions overlapped with onscreen 'current news' text,
Captions overlapped with onscreen 'other news' text, and
Captions not overlapped with any onscreen text.

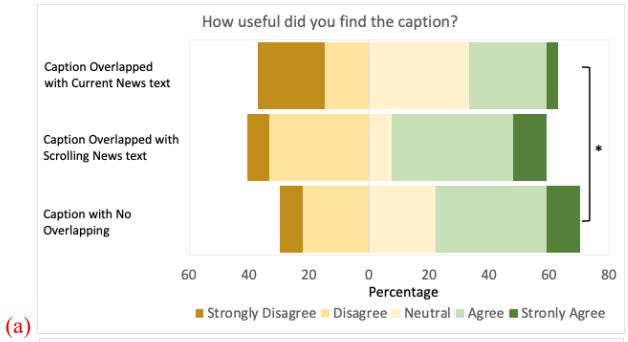
For this study, we created nine video stimuli which were collected from the same video source as previous "face occlusion" study. These videos consisted of TV news and panel discussions which are identified by prior work as the video genres of the highest captioning importance [6]. In all of these videos, a region of the screen included some text indicating the topic of the current news story, and another region contained a scrolling text area that presented headlines for other news stories. Both regions were near the bottom of the video image. As in the earlier "face occlusion" study, we cautiously selected videos in which speakers discussed a topic, which was not related to any political or trending issues, so that participants would not have strong emotional reactions to the content.

As before, to modify the caption transcript provided by the broadcaster, we extracted the caption file (.ass file) from the video source. Then we manually positioned the caption in such a way that a small fraction of 'current news' region (shown as black text on a white background in Figure 3) was visible to participants. The rationale behind this decision to reveal a tiny amount of the 'current news' text was that if in our caption completely overlapped the text, then participants might not be aware that this text existed in the first place. We also ensured that there were no errors in the caption transcript, as we did in the first sub-study. Using FFMPEG, we a total of nine videos, i.e. 3 sample videos produced under three different 3 caption-overlap conditions.

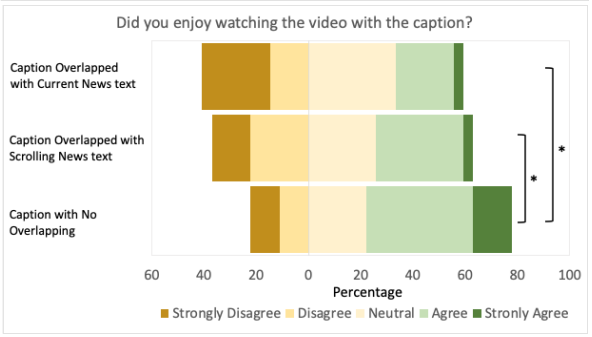
Figure 3: Sample images of video stimuli that were shown to participants during Sub-Study 2

Participants viewed videos, with captions placed on different positions such that (1) it overlapped with the current news text, (2) overlapped with scrolling news text, or (3) did not overlap with any text onscreen. On a screen that displayed three video stimuli side-by-side, participants viewed each video individually, and then answered the scalar

questions after each. The left-to-right placement of the videos and the assignment of condition to each video was again determined using the Greco-Latin square method.



(a)



(b)

Figure 4: Participants' subjective scalar responses for videos in each of the three conditions (no overlap, overlap with scrolling news text, overlap with current news text) in the Text Occlusion study, for (a) How useful did you find the caption? (b) Did you enjoy watching the video with the caption? Asterisks * mark significant pairwise differences ($p<0.05$).

Figure 4 displays participants' responses to the "usefulness question" (Q1) and "enjoyability question" (Q2) for the Text Occlusion study across the three conditions. All significant pairwise differences are indicated with an asterisk (*), if the p-value is less than 0.05. The details of the statistical analysis performed for each question is described below.

For the "usefulness question," a Wilcoxon-Signed Rank test revealed that participants found captioned videos more useful when the caption does not overlap with any onscreen text (Median=4) as compared to videos in which the caption overlaps with onscreen current news text (Median=3), ($Z= -2.291$, $p<0.022$). However, no significant difference was observed in the response to this question, for any other pairs of conditions. For the "enjoyability question," a Wilcoxon-Signed Rank test revealed that participants preferred captioned videos in which the caption does not overlap with any text on screen (Median=4) as compared to videos in which the caption overlapped with onscreen current news text (Median=3), ($Z=-2.688$, $p=0.007$), or as compared to videos in which the caption overlapped with onscreen scrolling news text (Median=3), ($Z=-2.112$, $p=0.03$). However, for this question, there was no significant pairwise difference between responses for the "overlap with scrolling news" condition and the "overlap with current news" condition.

6 5 Discussion

Our experimental studies had examined each of our research questions, which concerned how variation in occlusion features in a captioned video may affect the judgements of DHH viewers about the video's quality, specifically in regard to whether users found the captioned video useful or enjoyable to watch.

RQ1: Are DHH viewers' subjective judgments about the usefulness and enjoyability of captioned videos affected by (a) whether captions overlap with the onscreen speaker's eyes and (b) whether captions overlap with the onscreen speaker's mouth?

In this Face Occlusion study, we observed that participants found captioned videos in which the caption did not overlap with the speakers face, to be more useable and enjoyable, as compared to videos in which the caption overlapped with the speaker's eyes or mouth. These quantitative findings provide an answer to both part (a) and (b) of research question RQ1. This finding aligns with prior work, which had highlighted DHH viewers' concerns with the captions overlapping with critical onscreen content [5]. This finding also establishes that the onscreen speaker's face (their eyes and mouth, specifically) should be considered essential onscreen content, which should be visible to DHH viewers. Our findings also integrate with results of prior gaze-tracking studies with DHH viewers, which had revealed that users spend a significant amount of time focusing on a speakers' mouth [37] for speechreading.

As discussed at the end of section 4.1, we had also received some open-ended feedback from our participants at the end of the study, which sheds light on some of our earlier quantitative findings. Specifically, when participants were asked what they liked or disliked about the captions used in videos during this study, participants often talked about it is important for DHH viewers to be able to see a speaker's face. For instance, P13 explained their preference for "captions at the bottom of the screen, and not covering any useful information, was most helpful. It is critical for me, as a deaf individual, to also be able to see other people's faces, lips, and expression when speaking." When considering P13's comment, it is useful to note that no important visual information was present at the bottom of the screen (e.g. no onscreen text), which would could have been blocked when caption was at the bottom of the screen in the "no face overlap" condition. P2 commented about several of the video stimuli in which face occlusion had occurred, saying, "Placement was awkward for two of the three [videos]. Captions blocked off speaker's face, making it difficult to follow the speaker." In addition to P2, the majority of participants identified in their comments how some of the videos, in which there was no overlap with speakers' face, were better than the rest.

RQ2: Are DHH viewers' subjective judgments about the usefulness and enjoyability of captioned videos affected by (a) whether captions overlap with onscreen text containing the current news headline and (b) whether captions overlap with onscreen text about other news headlines?

In Text Occlusion study, participants' responses revealed that captioned videos in which the caption did not overlap with any onscreen text were more enjoyable, as compared to videos in captions occluded text with the current news headline or scrolling text displaying headlines for other news stories. Participants had also found videos in which the caption overlapped with current news text to be less useful than videos with no overlapping. Thus, our study provided an answer to part (a) of research question RQ2, but we had only measured an effect on "usefulness" scores in regard to part (b) of this question.

Open-ended comments from the participants revealed some of the possible reasons behind these findings. Participants, pointed out that the lack of visibility of current news text makes them lose track of the information that is being delivered in news. For instance, P15 commented, "I do not like that the captions cover the important headlines. It makes me guess what topic they are talking about. Covering that important information makes me to lose a track of information. It makes me little frustrated when reading." In addition, some participants reported that the placement of captions on top of scrolling news or current news text also made it harder to read the caption itself, with P3 explaining, "Placement of captions on top of headlines or directly below make them difficult to follow/read."

7 6 Limitations and Future Work

There were several limitations in our study, some of which may suggest directions for future research, e.g. to generalize these findings to other groups of users or other video genres: Our study focused on DHH individuals that were recruited at a university, which reflects only a particular subset of the diverse DHH community. Future studies should examine the judgements of other potential users of captions, e.g. DHH individuals recruited from broader geographic settings or educational backgrounds, hearing individuals who may become DHH later in life, people with situationally induced hearing loss due to environmental noise, or specific sub-groups of the DHH community who prefer particular modes of communication (sign-language users vs. speech communication). In addition, a variety of educational factors or other demographic characteristics may affect how variations in these features may affect users' responses.

The motivation for our work has been to identify potential features for inclusion in new metrics for evaluating live television programs, where the near-real time context makes it more challenging for television broadcasters to provide high-quality captions. Given this focus, when preparing stimuli for our studies, we used videos of news broadcasts, including news anchors speaking to the camera, conducting interviews, panel discussions, or televised segments of speeches. Additional research would be needed to investigate a wider range of television genres, e.g. sports. A different set of critical onscreen content might be present in such videos, e.g. current scores during a sports competition, and such research may broaden the list of critical onscreen content that captions should not block.

Our study investigated DHH viewers' judgements of caption quality in regard to two specific questions, the usefulness of the caption (a question adapted from [23]) and how enjoyable the captioned video was. However, there are a variety of other measurement instruments that could be used to assess caption quality, either to gather further dimensions of subjective judgements, or to evaluate captions through some objective measure. For instance, our study did not explore the effect of occlusion on the understandability of onscreen text, which could be measured using objective comprehension questions.

While our study investigated two specific occlusion features, there may exist many other types of high-importance or high-information regions of videos which could be defined, which would further influence users' judgements of quality. More broadly, there are other features of caption quality that may have interaction effects with these occlusion features. For instance, in our "face occlusion" study, we observed the effect of one feature (captions overlapping with the speaker's face), for a set of videos that consisted of a single speaker in the video, and with perfect transcription accuracy. There is a potential for interaction effects between face-occlusion and other captioned-video qualities, e.g. the number of speakers in the video segment, the latency in the captioning, etc. In future work, a multi-factor study design would be needed to investigate whether such interactions may exist among caption properties.

8 7 Conclusion

Although the transcription error in live television captioning has reduced in recent years, due in part to the availability of caption evaluation metrics [18], DHH users are still not fully satisfied with the quality of caption provided by the television broadcasters. Prior literature has suggested other important features of captioned video, such as whether captions occlude a speaker's face and other critical onscreen content. However, no prior empirical studies had examined the effect on DHH users' judgement of captioned videos from occlusion features, i.e. features relating to the overlap of captions with onscreen content that appears at particular times and locations in a video.

As discussed in our Introduction, to enable regulatory agencies responsible for monitoring the quality of captioned television programming to evaluate larger samples of captioned broadcasts, improved metrics are needed for automatically analyzing a captioned video to predict how it would correlate with DHH viewers' judgements. However, the currently available metrics do not account for whether captions occlude important video content, despite some regulations on caption quality, including this feature, e.g. [16, 15].

Herein lies the key contribution of this study, namely: This research has identified properties of captioned video content – occlusion features – and has quantify the extent to which they affect DHH users' judgement of captioned

videos. Basic empirical research, to examine how such features influence DHH viewer's quality-judgements, is foundational to the creation of high-quality, automatic metrics for efficiently evaluating captioned television programming. Having demonstrated this effect on DHH viewers' judgments, future research can investigate this relationship and examine how such features can be incorporated into existing metrics, which focus predominantly on text transcription accuracy. Since being able to measure something is often the first step toward improving it, by supporting the creation of metrics that consider a broader set of captioned video properties, this work may lead to more accessible television and video experiences for DHH viewers.

ACKNOWLEDGMENTS

Redacted for anonymous review.

REFERENCES

[1] Ahmed Ali and Steve Renals. 2018. Word Error Rate Estimation for Speech Recognition: e-{WER}. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 20-24.

[2] Wataru Akahori, Tatsunori Hirai, and Shigeo Morishima. 2017. Dynamic Subtitle Placement Considering the Region of Interest and Speaker Location. *In Proceedings of 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*

[3] Tom Apone, Brad Botkin, Marcia Brooks, and Larry Goldberg. 2011. Research into Automated Error Ranking of Real-time Captions in Live Television News Programs. Caption Accuracy Metrics Project. National Center for Accessible Media, 2011. Retrieved from http://ncam.wgbh.org/file_download/136

[4] Tom Apone, Brad Botkin, Marcia Brooks, and Larry Goldberg. 2011. Caption Accuracy Metrics Project Research into Automated Error Ranking of Real-time Captions in Live Television News Programs The Carl and Ruth Shapiro Family National Center for Accessible Media at WGBH (NCAM)

[5] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred Appearance of Captions Generated by Automatic Speech Recognition for Deaf and Hard-of-Hearing Viewers. *In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, Paper LBW1713, 6 pages. DOI: https://doi.org/10.1145/3290607.3312921

[6] Larwan Berke, Matthew Seita, and Matt Huenerfauth. 2020. Deaf and hard-of-hearing users' prioritization of genres of online video content requiring accurate captions. *In Proceedings of the 17th International Web for All Conference (W4A '20).* ACM, New York, NY, USA, Article 3, 1–12. DOI:https://doi.org/10.1145/3371300.3383337

[7] Larwan Berke. 2017. Displaying confidence from imperfect automatic speech recognition for captioning. *ACM Special Interest Group on Accessible Computing (SIGACCESS)*, 117 (2017), 14–18. DOI: https://doi.org/10.1145/3051519.3051522

[8] Debra L Blackwell, Jacqueline W Lucas, Tainya C Clarke. 2014. Summary health statistics for U.S. adults: National Health Interview Survey, 2012. National Center for Health Statistics. Vital Health Stat 10(260).

[9] Bonnie B. Blanchfield, Jacob J. Feldman, Jennifer L. Dunbar, & Eric N. Gardner. 2001. The severely to profoundly hearing-impaired population in the United States: prevalence estimates and demographics. *Journal of the American Academy of Audiology,* 12(4), 183–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11332518

[10] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic Subtitles: The User Experience. I*n Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '15)*. ACM, New York, NY, USA, 103–112. DOI:https://doi.org/10.1145/2745197.2745204

[11] Denis Burnham, Greg Leigh, William Noble, Caroline Jones, Michael Tyler, Leonid Grebennikov, and Alex Varley. 2008. Parameters in Television Captioning for Deaf and Hard-of-Hearing Adults: Effects of Caption Rate Versus Text Reduction on Comprehension. *The Journal of Deaf Studies and Deaf Education*, Volume 13. Issue 3, Summer 2008. Pages 391–404, https://doi.org/10.1093/deafed/enn003

[12] Michael Crabb, Rhianne Jones, Mike Armstrong, and Chris J. Hughes. 2015. Online News Videos: The UX of Subtitle Position. *In Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 215–222. DOI:https://doi.org/10.1145/2700648.2809866

[13] Tessa Dwyer, Claire Perkins, Sean Redmond, and Jodi Sita. 2018. Seeing into Screens: Eye Tracking and the Moving Image, USA: Bloomsbury.

[14] FFMPEG Developers. 2016. ffmpeg tool (Version be1d324) [Software]. Available from: http://ffmpeg.org/

[15] Federal Communications Commission. 2014. Closed Captioning Quality Report and Order, Declaratory Ruling, FNPRM. Retrieved from: https://www.fcc.gov/document/closed-captioning-quality-report-and-order-declaratory-ruling-fnprm
[16] Federal Communications Commission. 2012. Closed Captioning of Internet Protocol-Delivered Video Programming: Implementation of the Twenty-First Century Communications and Video Accessibility Act of 2010. Adopted rules governing the closed captioning requirements for the owners, providers, and distributors of video programming delivered using IP, and governing the closed captioning capabilities of certain apparatus on which consumers view video programming. MB Docket No. 11-154. FCC 12-9.
[17-18] Abraham Glasser, Edward Mason Riley, Kaitlyn Weeks, and Raja Kushalnagar. 2019. Mixed Reality Speaker Identification as an Accessibility Tool for Deaf and Hard of Hearing Users. *In Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology (VRST '19).* ACM, New York, NY, USA, Article 80, 1–3. DOI:https://doi.org/10.1145/3359996.3364720
[18] Government of Canada, Canadian Radio-television and Telecommunications Commission, & Crtc. 2019. Broadcasting Regulatory Policy CRTC 2019-308. Retrieved from https://crtc.gc.ca/eng/archive/2019/2019-308.htm
[19] Michael Gower, Brent Shiver, Charu Pandhi, and Shari Trewin. 2018. Leveraging Pauses to Improve Video Captions. *In Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '18).* ACM, New York, NY, USA, 414-416. DOI: https://doi.org/10.1145/3234695.3241023
[20] Stephen R. Gulliver and Gheorghita Ghinea. 2003a. How level and type of deafness affect user perception of multimedia video clips. Inform. Soc. J. 2, 4, 374--386.
[21] Stephen R. Gulliver and Gheorghita Ghinea. 2003b. Impact of captions on hearing impaired and hearing perception of multimedia video clips. *In Proceedings of the IEEE International Conference on Multimedia and Expo.*
[22] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. 2010. Dynamic captioning: video accessibility enhancement for hearing impairment. *In Proceedings of the 18th ACM international conference on Multimedia (MM '10)*. ACM, New York, NY, USA, 421–430. DOI:https://doi.org/10.1145/1873951.1874013
[23] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the Usability of Automatically Generated Captions for People who are Deaf or Hard of Hearing. *In Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17).* ACM, New York, NY, USA, 165–174. DOI:https://doi.org/10.1145/3132525.3132542
[24] Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2014. Accessibility Evaluation of Classroom Captions. *ACM Special Interest Group on Accessible Computing (SIGACCESS)*. 5, 3, Article 7 (January 2014), 24 pages. DOI:https://doi.org/10.1145/2543578
[25] Walter S. Lasecki, Christopher D. Miller, and Jeffrey P. Bigham. 2013. Warping time for more effective real-time crowdsourcing. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13).* ACM, New York, NY, USA, 2033–2036. DOI:https://doi.org/10.1145/2470654.2466269
[26] Daniel G. Lee, Deborah I. Fels, and John Patrick UDO. 2007. Emotive captioning. *Computers in Entertainment.* 5, 2, Article 11, 15 pages. DOI:https://doi.org/10.1145/1279540.1279551
[27] S. Nam, D. I. Fels and M. H. Chignell. 2020. Modeling Closed Captioning Subjective Quality Assessment by Deaf and Hard of Hearing Viewers. *In Proceedings of IEEE Transactions on Computational Social Systems*, DOI: https://doi.org/10.1109/TCSS.2020.2972399.
[28] NIDCD. 2017. National Institute of Deafness and Other Communication Disorder, 2017: Captions for Deaf and Hard-of-Hearing Viewers. Retrieved from https://www.nidcd.nih.gov/health/captions-deaf-and-hard-hearing-viewers.
[29] Ofcom, 2015. Measuring live subtitling quality, UK. Retrieved from: https://www.nidcd.nih.gov/health/captions-deaf-and-hard-hearing-viewers
[30] Oskar Olofsson. 2019. Detecting Unsynchronized Audio and Subtitles using Machine Learning (Dissertation). Retrieved from:http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-261414
[31] Press Release. 2016, World-first approach to reduce latency in live captioning. Ericsson. Retrived from: https://www.ericsson.com/en/press-releases/2016/6/world-first-approach-to-reduce-latency-in-live-captioning
[32] Pablo Romero-Fresco and Juan Martínez Pérez. 2015. Accuracy Rate in Live Subtitling: The NER Model. *Audiovisual Translation in a Global Context*. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan, London

[33] James Sandford. 2015. The impact of subtitle display rate on enjoyment under normal television viewing conditions, *In Proceedings of IET Conference Proceedings*, 2015, DOI: 10.1049/ibc.2015.0018
[34] Olaf Strelcyk and Gurjit Singh. 2018. TV listening and hearing aids. PLoS ONE. https://doi.org/10.1371/journal.pone.0200083
[35] Quoc. V. Vy and Deborah. I. Fels. Using Placement and Name for Speaker Identification in Captioning. *In Proceedings of Computers Helping People with Special Needs*.
[36] James M. Waller and Raja S. Kushalnagar. 2016. Evaluation of Automatic Caption Segmentation. *In Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16)*. ACM, New York, NY, USA, 331–332. DOI:https://doi.org/10.1145/2982142.2982205
[37] Jennifer Wehrmeyer. 2014. Eye-tracking Deaf and hearing viewing of sign language interpreted news broadcasts. *Journal of Eye Movement Research*. Retrieved from https://core.ac.uk/download/pdf/158976673.pdf
[38] Ibrahim A. Zedan, Khaled M. Elsayed, and Eid Emary. 2016. Caption Detection, Localization and Type Recognition in Arabic News Video. *In Proceedings of the 10th International Conference on Informatics and Systems (INFOS '16)*. ACM, New York, NY, USA, 114–120. DOI: https://doi.org/10.1145/2908446.2908472
[39-39] Zhang Zhaoxiang, Wang Chao, Wang Yunhong. 2011. Video-Based Face Recognition: State of the Art. In *Biometric Recognition* (pp. 1–9). Springer Berlin Heidelberg.
[40] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: An Efficient and Accurate Scene Text Detector. *In the Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
[41] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic Subtitles: The User Experience. In Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '15). Association for Computing Machinery, New York, NY, USA, 103–112. DOI:https://doi-org.ezproxy.rit.edu/10.1145/2745197.2745204