Understanding How Deaf and Hard of Hearing Viewers Visually Explore Captioned Live TV News

Akhter Al Amin
Rochester Institute of Technology
Rochester, New York, USA
aa7510@rit.edu

Saad Hassan
Rochester Institute of Technology
Rochester, New York, USA
sh2513@rit.edu

Sooyeon Lee
New Jersey Institute of Technology
Newark, New Jersey, USA
sooyeon.lee@njit.edu

Matt Huenerfauth
Rochester Institute of Technology
Rochester, New York, USA
matt.huenerfauth@rit.edu

ABSTRACT

Captions blocking visual information in live television news leads to dissatisfaction among Deaf and Hard of Hearing (DHH) viewers, who cannot see important information on the screen. Prior work has proposed generic guidelines for caption placement but not specifically for live television news, and important genre of television with dense placement of onscreen information regions, e.g., current news topic, scrolling news, etc. To understand DHH viewers' gaze behavior while watching television news, both spatially and temporally, we conducted an eye-tracking study with 19 DHH participants. Participants' gaze behavior varied over time as measured by their proportional fixation time on information regions on the screen. An analysis of gaze behavior coupled with open-ended feedback revealed four thematic categories of information regions. Our work motivates considering the time dimension when considering caption placement, to avoid blocking information regions, as their importance varies over time.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in accessibility**.

KEYWORDS

Accessibility, Caption, Attention, Area of Interest

ACM Reference Format:
Akhter Al Amin, Saad Hassan, Sooyeon Lee, and Matt Huenerfauth. 2023. Understanding How Deaf and Hard of Hearing Viewers Visually Explore Captioned Live TV News. In *20th International Web for All Conference (W4A '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3587281.3587287

W4A '23, April 30–May 01, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0748-3/23/04…$15.00
https://doi.org/10.1145/3587281.3587287

1 INTRODUCTION

More than 360 million people across world, and 15% of US adults, are Deaf or Hard of Hearing (DHH) who rely on captioning to access auditory information in television programming [2, 4, 10]. However, captions can also block important information on a screen, which can lead to DHH viewers missing crucial visual information and lead to dissatisfaction with their viewing experience [2, 3, 5, 9, 24]. This onscreen visual information may be graphical or textual, e.g., the face of a person who is speaking or some onscreen text that provides important information. This paper provides insights about how DHH viewers' preference for these information regions change over time and in-depth guidance for TV captioners on where to place captions during a live TV news broadcast based on this dynamic visual demand.

In particular, during live television news programs, common screen layouts tend to contain a dense amount of visual information, including various onscreen text or graphics (e.g., a headline text, a scrolling news ticker) and potentially multiple individuals who are speaking (e.g., during a multi-person interview or when a news presenter speaking with a reporter) [3, 5]. During video with such information density, it is not possible to place caption without blocking something, and so a consideration of tradeoffs is necessary about which region of the screen is least harmful to occlude with a caption. If a captioner is making this decision during a live TV news broadcast, there is little time available for them to make a rapid decision about where to place the caption in a to block a region of lowest importance. Research is needed to help captioners with making this decision or to support the creation of tools that can automatically place captions atop the least important region.

While there are several existing guidelines for caption appearance and placement [8, 16, 17], these guidelines often contain only general recommendations to avoid blocking content on the screen. They do not address trade-offs necessary when placing captions during information-dense video, nor are these guidelines specific to news, an important genre of television programming that supports viewers awareness of emergency events, political and societal engagement, and other important information. In addition, while there are various metrics for evaluating the quality of captioning during a video, e.g., [1, 11, 26, 41], most do not penalize captions occluding onscreen content.

Some prior work has sought to identify which information regions on the screen during live television programming DHH viewers believe would be most important to avoid blocking with captions [3, 5], with a goal of producing caption-placement guidelines or evaluation metrics. However, a limitation of that prior work was the *source* of this information about the importance of various regions of the screen. Some work asked DHH individuals to subjectively rate the importance of various information regions of a video [3] or asked DHH participants to watch a video in which captions blocked something and rate their experience [5]. However, a common approach within the HCI and accessibility literature is for researchers to consider more direct behavioral measures of attention, rather than depending solely upon indirect, subjective assessments from participants, e.g., [7, 35, 47].

This work elicits DHH viewers' attention distribution in temporal space while watching live captioned video. We utilize eye-tracking technology to capture which information regions on a TV screen layout DHH participants looked at different points in time in news videos. We compare the importance of various on-screen regions based on attention distribution over time with the importance of various regions captured using only subjective judgments in prior research [3]. We create four thematic groupings of on-screen regions with similar gaze distributions. For each group, we also present a thematic analysis of open-ended feedback that uncovers **how participants perceived their division of attention over time** and what **factors led them to do so** for the sets of onscreen regions. Our findings inform better captioning guidelines that take into account the relative importance of different onscreen regions over time and the design of more sophisticated caption evaluation metrics in the future that make use of behavioral data.

There are two empirical contributions of this research:

(1) Our analysis of the proportion of time DHH participants viewed various regions of the screen during television news video revealed a different prioritization of these regions' importance, as compared to prior work that depended upon subjective ratings from participants. These findings reveal that direct behavioral measures of attention can shed new light on DHH viewer's use of information regions.
(2) Our analysis of attention over time on various information regions revealed clusters of regions with similar attention patterns, e.g., some which were most important during the first few seconds of a video, some that drew sustained attention over time, and some with occasional bursts. Participants' responses revealed factors that affected how attention was distributed, e.g., whether regions had static or dynamic content, whether the content was textual, and whether the content provided essential context for the news story. These findings inform the work of broadcasters or captioning professionals, e.g., in suggesting groups of information regions which vary in their importance over time and thereby would suggest prioritizing how to place caption text to avoid blocking specific regions during specific times.

2 BACKGROUND AND RELATED WORK

As motivation for why we seek to understand which information regions on the screen during live television videos are most important for DHH viewers, some background is presented about prior work on captioning guidelines and evaluation metrics, which generally do not consider this issue of caption placement and occlusion of onscreen content. Next, we discuss prior work that has investigated how DHH viewers would prioritize various regions of a screen; a limitation of this prior work is that it had made use of indirect, subjective ratings from DHH viewers when determining the importance of regions. Finally, we consider prior work on the use of eye-tracking to analyze how viewers, including DHH viewers in particular, distribute their attention when watching video; however, that prior work had not specifically considered how DHH viewers would prioritize various information regions, nor had that work specifically considered television news video.

2.1 Prior Work on Guidelines or Metrics for Captioning Placement

Prior work on caption placement had considered both a prospective (where to place captions) and retrospective (were captions in a video placed well) perspective:

- Researchers have investigated various approaches for selecting where captions should appear on screen. Some have focused on placing captions close to person who is currently speaking [22, 23, 42]. While changing the location of captions too often can place a burden of viewers, who must visually seek the caption on screen [29], such dynamic placement technologies [12, 30] generally improve DHH viewers' experience, e.g. by putting captions near the person speaking.
- Existing caption evaluation metrics, employed in commercial settings, mainly measure the quality caption transcriptions [1, 11, 26, 41]. While these metrics improve the quality caption transcriptions, several research have shown that captions blocking onscreen information may reduce DHH viewers' ability to perceive vital information that appear on the screen [3, 4, 6, 9].

A limitation of these prospective guidelines and retrospective metrics is that it had not substantially considered the issue of captions occluding other visual information content during videos.

Most relevant to our current study, some prior research has intuitively detected a few important regions of the screen, e.g., the face of the person speaking, and avoid blocking those while placing a caption using their software [6, 25]. However, this work had considered a relatively limited set of visual elements on the screen to avoid blocking with a caption, and it had not specifically considered the broad range of information regions present on the screen in information-dense videos, such as television news.

2.2 Prior Work on Identifying Important Regions to Avoid Occluding with Captions

In an effort to produce more sophisticated guidelines or metrics that considered this issue of captions blocking information regions during live video, there has been some prior work that has sought to understand DHH viewers' subjective judgments about the importance of information regions that appear on a TV screen. Amin et al. [4] have itemized a list of information regions that in often appear across 6 television genres. They conducted a study in which

participants gave their subjective judgement of how important each of these regions were for videos in each genre, and they used this dataset to develop an occlusion-based metric to evaluate the quality of caption placement in videos.

In later work, Amin et al. [5] asked participants to view videos in which captions blocked information regions on the screen and assign a quality score to each video; a regression modeling approach was used to develop a caption-placement quality metric in videos, which penalized placements of captions based on which information regions were blocked.

A limitation of this related work on understanding the importance of information regions that appear in videos is that it had relied upon subjective ratings from DHH participants, rather than direct behavioral measurements of visual attention. Participants' recollection of how their attention was divided across information regions may not match their actual behavior. In addition, because this prior research had relied upon overall subjective judgements, rather than behavioral measures, this work had not explored whether the importance of regions of the screen may vary over time during a video. For instance, not all information may require someone's continuous attention: only a few seconds might be sufficient to comprehend a short piece of text on the screen, whereas some other information may require users' longer or more continuous attention.

2.3 Prior Work on Measuring Visual Attention of DHH Viewers During Video

There have been several research studies that have attempted to predict users' regions of interest within a video frame or image [3, 14, 20, 33, 42, 46], and many researchers have employed users' eye-tracking data in modelling regions of interest within an image [46]. Other work has utilized eye-tracking to measure users' distribution of attention while performing tasks, e.g., [37, 44]; such gaze-pattern-based task modelling demonstrates how a users' gaze can reveal the saliency of information [13, 31, 43].

For DHH viewers specifically, some research has employed users' eye-tracking measurements to train machine learning models to estimate where DHH viewers might distribute their attention given certain types of video or content appearance [14, 15, 38]. In fact, by collecting users' gaze measurement, Zheng et al. attempted to estimate salient frames within a video [33]. Most relevant to our research, Kuno et al. and Hu et al. have proposed a gaze adaptive caption placement technique that follows users' gaze and try to place caption near to that location [23, 30]. However, such technologies require the viewer's gaze to be tracked with eye-tracking while they are viewing a video, which is not practical in most television-viewing contexts.

To understand what regions of a video are most salient, some researchers have collected datasets using eye-tracking technology, to determine where (non-DHH) viewers tend to focus their gaze, e.g., [36, 45]. However, such datasets may not generalize to DHH viewers, as prior work has revealed significant differences in gaze behavior between DHH viewers and hearing individuals [45]. Therefore, it is important to determine how DHH viewers' gaze distribution on an information region dataset varies over time and what are the key issues that influence these gaze behavior. No prior work had collected eye-tracking data to investigate this attention distribution over time among DHH viewers for the various information regions in television news video.

No prior study has conducted an in-depth analysis of DHH viewers' gaze behavior when watching information-dense live television news; our study addresses this gap in knowledge, which would be valuable for providing guidance for broadcasters or captioning professionals who must decide where to place captions for such video. Whereas some prior work on caption placement and occlusion of information regions during video had asked participants to give subjective judgements, direct behavioral measures of attention distribution may reveal how attention shifts over time.

3 RESEARCH QUESTIONS

In our first research question, we compare an importance ranking of information regions based on users' gaze patterns with a prior ranking of the importance of information regions based on DHH viewers' subjective judgments [5].

RQ1 Is there any difference between what DHH viewers believe they are paying attention to and what their actual gaze behavior is when they are looking at a captioned video?

An additional advantage of analysis of eye-tracking recordings of visual attention is that it can also reveal shifts in attention over time, an issue which was not investigated in prior work on caption placement [4, 22, 23, 42]. Our next research question considers the patterns in how attention changes over time for various information regions in news video, as well as whether information regions fall into groups based on this analysis. We then triangulate participants' open ended feedback with the behavioral data to formulate a clustering of information regions in news video and to reveal factors that participants believe led to their attention patterns:

RQ2 What will be DHH viewers' gaze behavior over time while watching live captioned videos?
- (a) Is it possible to define categories of information regions based on an analysis of the attention curves?
- (b) What factors do participants believe explain their attention over time?

4 METHODS

4.1 Construction of Video Stimuli Dataset

We assembled a set of stimuli videos from various news TV channel sources, to satisfy several criteria:

(1) Since we aim to investigate how DHH viewers distribute their attention toward different onscreen information areas in TV news videos, we restrict ourselves to such videos only.
(2) Videos had to include information regions that are common in the TV news genre. We considered a set of information regions that had been enumerated in prior work that had analyzed such videos [5].
(3) The videos we selected did not include contentious or emotionally disturbing topics, which could have affected participants' preferences.
(4) Unlike video stimuli sets assembled in prior research on caption occlusion [2, 5], we placed captions so that they did not block other information regions. The rationale for this

choice was that our data analysis will focus on how DHH viewers' gaze moves across various information regions on the screen, and we sought to remove the potential confound of occlusion from captions in this data collection.

We reviewed 100 video samples from 15 TV channels and selected a total of 28 video stimuli from 9 TV channels. While selecting these videos, we ensured the screen layouts of these videos align with [3]. The layouts are as follows:(1) only news presenter appear on the screen; (2) both the speaker and listener appear on the screen; and (3) a news reporter is speaking from a location outside the television studio.

From each video, we created two versions of stimulus, by placing captions onto the video in one of two screen locations commonly used in the broadcasting industry. The locations are (1) the lower third of the screen and (2) the upper third. In this way, we created a total of 56 video stimuli from our selected 28 videos. All videos were resized to a uniform height of 720 pixels, and the width was adjusted according to the aspect ratio of the original video source. Notably, the length of each video stimulus was between 40 and 50 seconds. A previous eye-tracking study with participants drawn from a specific sub-population had demonstrated how 40-second videos were sufficient for analyzing gaze behavior [27]. Each participant watched 28 videos in a randomized non-repetitive manner.

In our video stimuli, we engineered the captions so that they accurately transcribe the spoken content. Also to maintain the standard captioning properties, we set the caption font size to 14, the font color to white, the caption background to black, and the font style to Arial. To simulate natural latency(3-6 seconds) of live captioning scenario, we have retained the time duration of visibility of each caption to be at least three seconds. After preparing a caption file following this protocol, we have burned the captions into the video, to produce stimuli for our study, so that regardless of the player software or platform used to display the videos, the captions would appear in a controlled location on the screen, occupying at most 10% of the total area of the video screen.

4.2 Area of Interest Annotation

After constructing this video stimuli dataset, we examined each video to annotate the location and timing of each information region that appeared on the screen, using as a basis the set of information regions from a prior study on live television video [3]. First, we extracted the individual frames from each 30-frames-per-second video; our annotation occurred on each individual frame of video. For each frame, we drew rectangular shape [18, 32] around each information region to annotate. After this initial annotation performed by one researcher, two other researchers reviewed every frame to ensure that the rectangular boxes sufficiently contain each information region in that image, while remaining as tight as possible. The two researchers performed this task together and discussed their work. The set of information regions annotated in each video consisted of:

(1) Current Discussion Topic (text displaying the headline of the news story),
(2) Listeners' Face (when there is a person onscreen who is listening to the current speaker),
(3) Scrolling News (moving text describing various news story headlines),
(4) Speakers' Face (the face of the person speaking),
(5) Logo of the Channel (graphic identifying the news network),
(6) Speakers' Location (text describing the geographic location of the person who is speaking),
(7) Speakers' Name (text identifying the person speaking),
(8) Current Time and Temperature (text displaying this information),
(9) Program Title (text name of the television program),
(10) Over the Shoulder Text (text appearing behind the shoulder of the person, common in news broadcasts), and
(11) Over the Shoulder Video or Animation (videos or animation appearing behind the shoulder of the person).

4.3 Technical Setup

We used a Tobii Pro Nano remote eye tracker with a 19-inch screen (resolution 1920×1200), and participants sat with an eye distance of 65cm from the screen [29, 40]. The distance was based on the recommended distance for the device. After calibration, participants were instructed to remain relatively still until each video stimulus segment finished playing. We displayed the 28 videos to each participant, in random order.

The eye-tracking system recorded the horizontal and vertical screen coordinates where the eye is aimed. Human eye gaze tends to move rapidly from one location to another, during movements called **saccades**. Moments when the eye is relatively stationary are called **fixations**. The recorded data was preprocessed using the iMotion software's fixation filter, with the following settings: velocity threshold = 30 pixels/samples, distance threshold = 30 pixels [29]. Because the videos contained primarily dialogue scenes, smooth-pursuit eye movements (in which a user visually tracks a moving item across the screen) were not analyzed in this study, as they are uncommon for this video genre.

A semi-structured interview session was conducted twice during the study: (1) after participants finished watching half of the video stimuli, and (2) at the end of the study. We asked participants to share their experience with watching the videos, and we asked what aspects of the videos and the information regions might have influenced their gaze behavior. This approach was motivated by prior work that had revealed how participants' qualitative responses complement eye-tracker-based gaze behavior data [28]. Our questions are shared assupplementary electronic files with this paper.

4.4 Participants

Participants were recruited through advertisements on social network groups and university student groups, with two screening questions: (1) "Do you identify as Deaf or Hard of Hearing?," and (2) "Do you use captioning when viewing videos or television?". 19 participants (6 men, 10 women, 3 non-binary) who responded yes to both questions were eligible to participate, with mean age 27.33 years (SD=6.46). Sixteen participants identified as D/deaf, and 3 as hard of hearing. Participants indicated spending on average 3 hours per week watching captioned news programs on TV.

5 ANALYZING DHH VIEWERS' ATTENTION DISTRIBUTION ACROSS INFORMATION REGIONS

5.1 Data Analysis

Eye-tracking data is usually processed into a list of the fixations that occur during a study, each with a: start-time, end-time, horizontal and vertical screen coordinates, and other information. To facilitate analysis, we perform one more step of processing on the fixation list. We had previously defined regions of the screen (during specific time durations of each video stimulus) that are important to consider; such regions are called "Areas of Interest" or AOIs. Each information region (the speaker's face, the text on the screen displaying the current news headline, etc.) had been defined as an AOI, consisting of the shape and location of the region and the time duration when it was visible. Each fixation in the fixation list can thus be labeled as to whether it was within an AOI. For each AOI, we generate a "proportional fixation time," which is the sum of the duration of all fixations on this AOI, divided by the total time of the recording segment. This data reveals the distribution of DHH viewers' gaze on each AOI.

5.2 Findings for Research Question 1

Figure 1 (a) displays the proportional fixation time of our participants across the various information regions in the news videos displayed in this study and Figure 1 (b) represents ranking of weights across various information regions in the news videos, as had been identified in a prior study [5]. In that prior work, researchers had asked DHH participants to report their subjective numerical rating of how bad it would be if captions blocked various information regions on the screen during TV news. For example, in that prior study, 'Scrolling News' was identified as top ranked whereas DHH viewers' gaze behavior in our study revealed low gaze time on that region. Furthermore, from the eye-tracking data in our study, we observed substantial gaze time on the 'Listeners' Face,' whereas 'Listeners' Face' was relatively lower ranked in that prior study.

Our comparison between the findings of a prior study based on subjective ratings and our current study based on eye-tracking data indicates the potential for direct behavioral measurements of visual attention to reveal new insights about which information regions on the screen are most used by DHH viewers during news videos. Given this, the following section will discuss how eye-tracking data can provide deeper insights, based on how attention may shift over time, and whether some information regions have similar patterns regarding this.

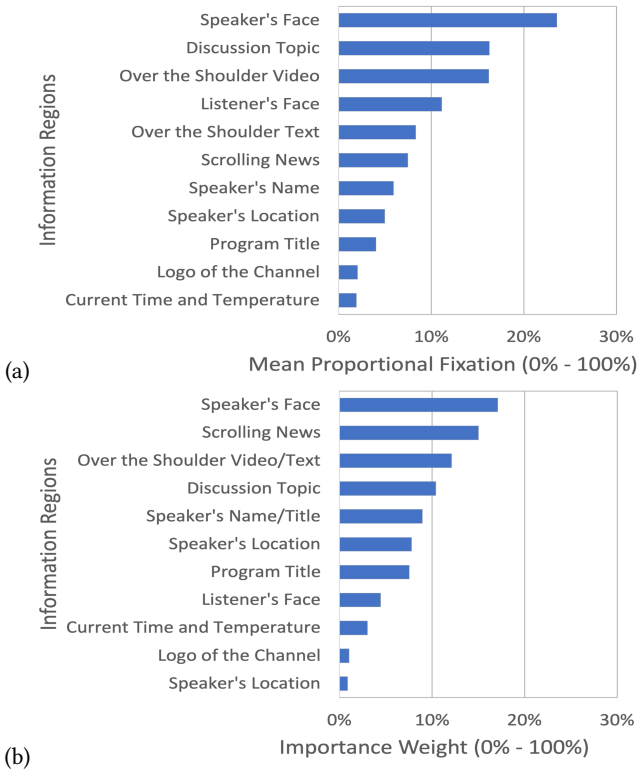

Figure 1: (a) Mean Proportional Fixation Time for information regions that appeared during news videos shown in our study, revealing an overall ranking of these regions based on the amount of time participants looked at each. (b) Data from a prior study [5], showing average of DHH participants' subjective numerical rating of the importance of information regions that appeared on the screen in news videos. Speaker's Face consists of data for eye and mouth regions, which had been considered individually in [5].

6 DHH VIEWERS' GAZE BEHAVIOR OVER TIME FOR EACH INFORMATION REGION

6.1 Data Analysis

To address the second research question, we analyzed the eye-tracking data to generate a continuous graph of the amount of fixation over time for each AOI, to determine how DHH viewers' gaze patterns and attention shifted over time across the areas of interest. In the subsections below, we include plots that display the fixation time for each information region, over time, averaging across all participants and all video stimuli in the study. This graph depicts how viewers' attention on a particular information region changes over time. The X-axis of these graphs refer to proportional fixation time scaled on a '0%' to '100%' scale where '0' represents no attention and '100' represents maximum attention.

Considering these graphs, which indicate how much participants were looking at each information region at various points in time during videos, may reveal whether some regions receive similar patterns of visual attention over time. For instance, some information regions on the screen may receive attention only in the first few seconds of a video, with a steep decline thereafter. To arrive at a categorization of the graphs for each information region, to identify regions with similar patterns of attention over time, we have employed rule-based approach by defining some subjective characteristics of these graphs [34], as follows:

- Overall heights of the curves
- Overall shapes of the curves
- Width of peaks throughout periods of sustained attention

Two researchers examined these graphs based on these shape properties, and a discussion took place between researchers to

consolidate this into the set of four groupings that correspond to four subsections below.

Next, to analyze the interview and behavioral data from the study qualitatively, we employed a mixture of deductive and inductive approaches [21] wherein these high-level groupings of graphs were used as a deductive framework. First, two authors read all 19 transcripts to build familiarity, then during a subsequent reading, they individually took notes to produce initial codes, which they collated and collapsed into two individual code-books. Each then investigated underlying patterns among their codes and formed initial categories, falling under each of the four graph-shape groupings. The authors then met to review their initial memos, to identify similarities and differences. During two three-hour meetings, the authors performed an initial thematic grouping within each category, which led to final themes, which were presented and discussed among the rest of the team to be finalized.

6.2 Findings: Group 1: Peak Followed by Slowly Decreasing Sustained Attention

Based on the shape of the attention curves for our information regions, as depicted as the jagged lines in the Figures throughout Section 6, our first grouping is characterized by a shape that consists of a peak followed by slowly decreasing sustained attention over time. The three information regions whose attention curves were categorized into this grouping were: Discussion Topic, Scrolling News, and Over the Shoulder Text. The overall shape of the graphs is highlighted by a smooth best-fit curve appearing as a solid line in Figure 2(a), based on a power-series best-fit of the average of the curves, and its shape suggests that viewers were more likely to look at these regions at the start of the video. However, they also continued to direct some attention to these regions throughout the video, as evident from the gradually decreasing attention over time. A common feature of these three information regions is that they consist of textual information, and the text is generally short and can be read quickly. (For scrolling news headlines a few seconds may be needed for the text to move along in order for it all to be read.)

In their open-ended feedback, participants discussed how their attention was distributed over time on these information regions and the factors affected their attention:

- **High attention priority:** Participants discussed the importance of not blocking these regions, e.g., P12 said *"The information on the bottom, the discussion topic, and the running headlines should be visible at any time. I want to be able to read those things and have those things not be blocked. It is fine if some of the information is blocked for a few seconds."* This suggests that while it might be okay for the captions to be blocked momentarily, these regions should stay unblocked throughout the video.
- **Initial visual scan:** Besides slowly decreasing sustained attention, another feature of these graphs was the peak at the start. Four participants' mentioned how they tended to look at these information regions for the first few seconds of a video. For example, P10 commented how they looked at the, *"discussion topic first, then ... - whatever captures my attention first."* P7 commented how at the beginning of a video, it was important to see *"the text on the bottom (current discussion topic), the scrolling text..."*
- **Providing important context:** Participants' also discussed why these regions are important for their viewing experience. Participants mentioned how the discussion topic provided them with important context needed for comprehending the content, especially at the beginning, but it was also important for it to always be available. For instance, P16 commented *"It's good to see the discussion topic stay there the whole time ... in case I need to reference it."* Similarly, P19 mentioned how any texts that appear over the shoulder of the news presenter was also crucial. *"I always look at the other text [referring to the over-the-shoulder text], and I also always look at the speakers faces, so you would never want it over the face or over the text."* (The "Speaker's Face" information region appears in grouping 2 below.)

6.3 Findings: Group 2: Sustained Attention

Based on the shape of their attention curves, the second grouping of information regions we identified were those which had a high level of sustained attention over time, throughout the video. Within this group were the following information regions: the Speakers' Face, the Listeners' Face, and the Over-the-Shoulder Animation/Video. As illustrated by the smooth best-fit curve displayed as a solid line in Figure 2(b), the overall shape of the attention curves reveals continuous attention to these information regions across the entire video duration.

Participants discussed how they perceived their their attention shifting across these information regions, and their open-ended comments also mentioned factors why they believed these regions of the screen drew their attention:

- **Human Faces Convey Emotion and Subtext:** Faces that appear on the screen provides emotional information, intent of the speaker, and how the listener is responding. P15 explained how facial expression and body language is related to the content: *"Again the person's mouth and facial expression and sometimes body language [are important]. You can really get a lot of information from body language and facial expressions about the context of the video."* Viewers tended to look not only who is speaking but also other individuals who are listening, and P12's comment reflects this: *" The face is the most important, even if someone is not talking. I would like to watch what their speaking, how their body language is."*
- **Dynamic Nature of the Information:** Three participants mentioned how the dynamic/moving properties of these information regions attracted their gaze when watching a captioned news video. For instance, P21 commented, *"We have peripheral vision. When something changes, it is easy to notice and then decide if I want to read it or keep looking at it."* Since these information regions consist of large moving visual elements (faces or video/animations), the movement of those regions may draw attention. Similarly, P12's commented how their attention was drawn due to these information regions' *"dynamic nature across screen. Yes, it's like a reflex. When I see something move, I automatically look to see*

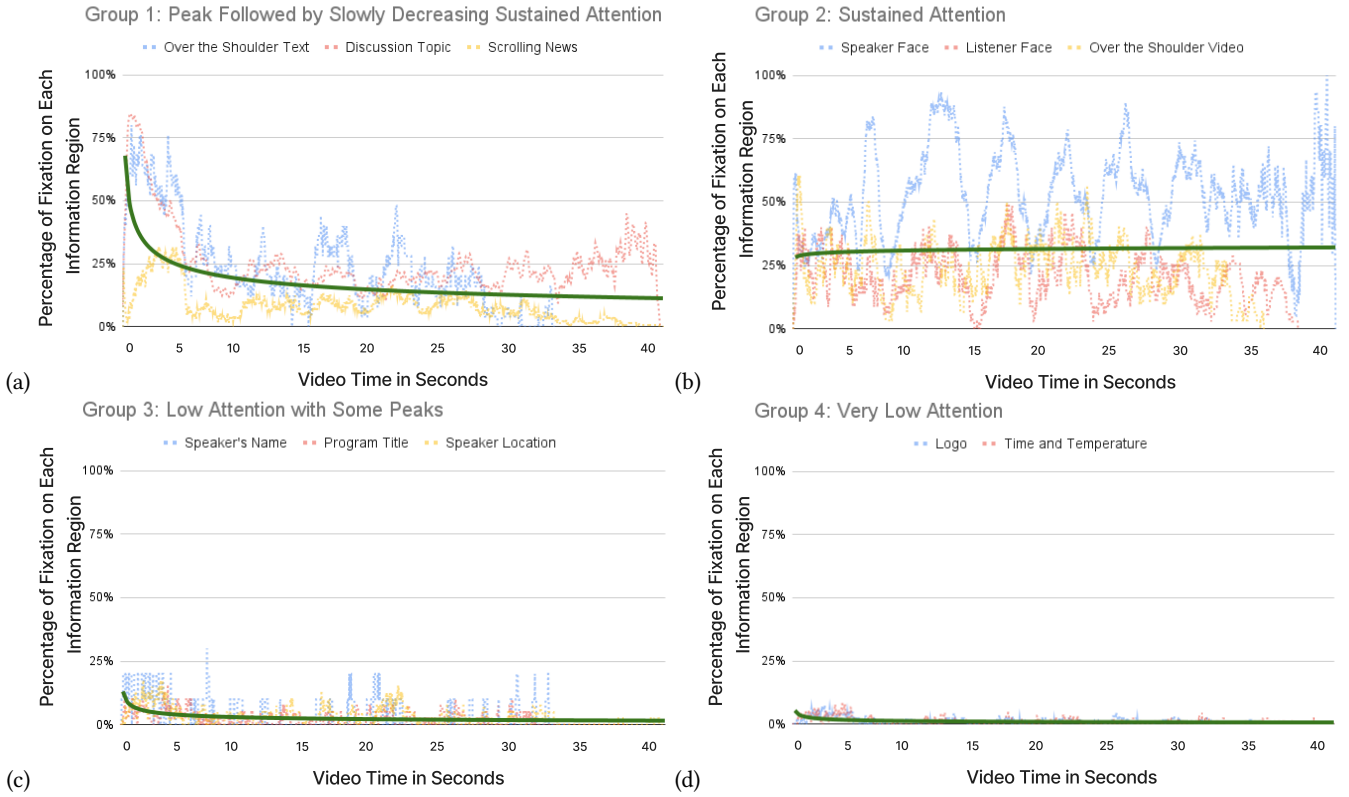

Figure 2: Curves depicting the relative percentage of attention over time for (a) three information regions (Current Discussion Topic, Scrolling News, and Over-the-Shoulder Text) across all videos and all participants, along with a smooth power-series best-fit curve for the average of all three regions, (b) three information regions (Speakers' Face, Listeners' Face, and Over-the-Shoulder Video/Animation) across all videos and all participants, along with a smooth power-series best-fit curve for the average of all three regions, (c) three information regions (Speakers' Name, Program Title, and Speakers' Location) across all videos and all participants, along with a smooth power-series best-fit curve for the average of all three regions and (d) two information regions ('Logo of the Channel' and 'Time and Temperature') across all videos and all participants, along with a smooth power-series best-fit curve for the average of both regions. Based on curve height and shape, we have categorized the information regions into four groups: "Group 1," which is characterized initial higher attention and then slowly decreasing sustained attention, "Group 2," which is characterized by sustained attention over time, "Group 3," which is characterized by low attention over time with some sharp peaks, and "Group 4," which is characterized by very low attention.

what happened on that area of the screen. It definitely impacts where you look during the video for sure."

- **Identification of Speaker:** Participants mentioned how following who the current speaker is and who is saying what is necessary when understanding the content of the video. Several participants mentioned how seeing the speakers' face allowed them to know who is speaking, e.g., P16 commented: *We need to know who is speaking the speaker or whoever is speaking on the screen. I typically look at the speaker.*
- **Providing Context:** Participants also discussed how any video/animation that appears over the shoulder of the person speaking can support their understanding of the topic of the news story, e.g., P8 mentioned, *"If there is a picture or video [over the shoulder] connected to what is being talked about, and then the discussion topic or description on the bottom, then I know the context of the video."*

6.4 Findings: Group 3: Low Attention with Some Peaks

Based on the shape of attention curves over time, our third group of information regions included: text displaying the Program Title, text displaying the Speakers' Location, and text displaying the Speakers' Name. As illustrated by the smooth best-fit curve shown in Figure 2(c), the general shape of these curves was a low amount of attention over time. There was a gentle peak at the beginning, but much less pronounced than for Group 1, and the overall graph height is lower than for Group 1. As shown in the jagged lines in Figure 2(c), which depict the proportional fixation data for each of the three information regions, there were some peaks of attention over time throughout the video, albeit at a low proportion of attention overall. Unlike the peaks in attention over time during the Group 2 information regions, the peaks in the recordings for Group 2 are sharper or more narrow, indicating brief glances at these regions.

In their open-ended comments, participants discussed some of the properties of these information regions—and commonalities between them—which may explain why there were occasional peaks in attention during the video:

- **Understanding the source:** Participants discussed how during a news broadcast, when some information or opinions were expressed, it was useful in those moments to understand who was saying this, to understand their authority or perspective. For instance P10, explained: *"I would be more interested in knowing who exactly was saying what."* Furthermore, P18 commented, *"Sometimes I see where the news is from, for example like UK or something like that."* These factors can help to explain why during some videos, participants may seek out the name or geographic location of the person who is speaking, which may help to explain the occasional peaks of attention observed for the three information regions in this group.
- **Static text requires only brief attention:** All three information regions in Group 3 consist of a piece of static text, which participants noted was short and could be read quickly. Three participants mentioned how because of this, they believe it would be OK if captions were to occasionally block these regions. During moments when the text was unblocked, it would be sufficient for them to take a quick look, to learn the name of the news program, the name of the speaker, or the geographic location of the news broadcast or speaker. For instance P11, commented, *"as long as the information is static then blocking it is fine."* P8 commented how sometimes the text for these information regions appear partway through a video, and this appearance of the text content is sufficient to draw their gaze briefly: *"It was easy to notice the additional text pops up [speakers' name or title]. When text was being switched out, I would quickly move my eyes to see and then go back to what I was watching."*

6.5 Findings: Group 4: Very Low Attention

Based on the height and shape of the attention curves, our fourth grouping consists of two information regions: (a) the Logo of the Channel and (b) the Current Time and Temperature. As illustrated in the smooth best-fit line in Figure 2(d), the general shape of attention curve for information regions in this grouping was an extremely low level of attention over time.[1] Like Group 3, there is some evidence of a small peak near the beginning, but there are relatively fewer peaks over time. During our discussions of groupings of the attention curves, we had considered whether to merge Group 3 and Group 4, but we decided to present it separately here, given the overall lower level of attention. Further, our subsequent analysis of participants' comments supported this decision to present this fourth grouping separately.

From participants' open-ended responses, we have observed some commonalities that might explain why participants' attention to these information regions are relatively lower than for other information region:

- **Does not affect understanding of the news story content:** Most of the participants explained how they tend not to look at these information regions, as these were less relevant to the news content they were watching. For instance, P13 shared that, *"for the most part, there is some information that is more important than others. Like the weather, temperature isn't as important as long as the other discussion topics and news are still able to be seen."* P9 agreed and commented, *"time, temp, logos, doesn't matter. I do not care, because I can pull my phone out and find that information."* P11 discussed how the integration between the main content of the news story and the onscreen information regions affected their importance, e.g., saying *"If the information are related to the topic, then it helps, but if it is not related to the topic being discussed then it does not add any value."*
- **Brief attention is sufficient:** Even if someone wanted to consult these information regions, participants indicated that a very brief glance would be sufficient, which would not lead to a high level of attention in the curves shown in Figure 2(d). For instance, P20 explained how for *"the date and time I only need to glance at briefly."* P7 believed these regions could be useful, but a glance is what is needed, saying *"things like logo, time, weather, are still helpful and might be OK for me as I just want to glance at the screen for that info."*

7 DISCUSSION

To address **RQ1**, we analyzed the behavioral data from our 19 DHH participants and found that importance-ranking of information regions based on users' gaze patterns was different from a prior importance-ranking that had been based on DHH viewers' subjective judgments [5]. For instance, Figure 2(a) revealed that participants directed more attention to the text showing the current Discussion Topic than Scrolling News text, yet participants in a prior study that used subjective judgments rated scrolling news as more important [5]. Further, in the current study, listener's face was ranked fourth in visual-attention priority, but it had been ranked eighth in that prior study. These differences suggest the value of direct behavioral measurements when investigating information needs of DHH viewers of videos.

To address **RQ2**, our analysis of gaze patterns of DHH viewers was coupled with an analysis of their open-ended comments. Prior work [5] based on subjective judgments had assumed that the importance of onscreen regions remained the same over time. In contrast, our eye-tracking methodology revealed how attention changed over time. Our analysis revealed four groupings of information regions, and within each, we were able to characterize this gaze change over time. Participants' comments also revealed some reasons why their gaze behavior changed over time.

7.1 Why does attention change over time?

We had grouped attention curves on three primary properties, and our analysis of users' open-ended comments revealed factors underlying the attention behaviors responsible for these patterns.

[1] We are aware that it is difficult to see the details in the graphs shown in Figure 2(d) because the values are so low, but we decided to provide all of the attention-curve graphs in this paper with an identical y-axis scaling, since graph height was a key factor in distinguishing among groups. Within our electronic supplementary files, we provide alternative versions of these graphs which are zoomed-in on the y-axis to reveal greater detail.

7.1.1 Overall heights of the curves. Our attention curves varied in terms of overall height. For example, the information regions belonging to Group 1 (speakers' face, listeners' face, and discussion topic) drew more attention than the rest, leading to higher attention curves. This finding was aligned with prior work that had simply considered an overall prioritization of various information regions [3, 5]. Furthermore, our findings aligned with prior work that had used eye-tracking to assess which areas of video drew the attention of DHH viewers. For example, prior eye tracking research has uncovered how the human face in captioned videos is a key source of information that allows DHH viewers to lipread and to understand the emotion and body-language of speaker [2, 45].

We interpret the overall heights of the curves as an indicator of how DHH viewers prioritize their attention. For instance, we observed low and generally flat lines corresponding to certain information regions, e.g., the information regions in the fourth group. Participants' comments revealed that group had less direct relevance to the video content, e.g., logo of the channel or time temperature (Group 4), and thus drew minimal attention.

7.1.2 Overall shapes of the curves. In our study, the attention curves that resulted from eye-tracking measurements had more nuance than just their overall height; our groupings of information regions also considered the overall shape of the curve. For Group 1 and Group 3, we had observed higher-than-usual amount of attention at the very beginning of the view, which then leveled off over time. This finding suggests that DHH viewers had engaged in an **initial visual scan** of these information regions at the beginning of the video. Such temporal nuance had not been revealed in prior studies that had depended upon subjective judgements, rather than direct behavioral measures.

Participants described how they had directed their visual attention towards information sources that provided them with overall context on the topic of the video. Our findings align with prior work in which DHH viewers had been observed while watching sign-language-interpreted news broadcasts [45]; specifically, that work had found that DHH viewers' attention towards textual information gradually decreased over time, whereas their attention towards human faces remained at a similar level across the video [39, 45].

Broadly, our findings revealed differences between static regions of information content, such as text displaying the headline of the news story, and dynamic information content, such as moving human faces or animated over-the-shoulder video. Generally, we observed curve shapes for static content that revealed a "higher at the beginning and then levels off" shape, with the more dynamic information content maintaining more sustained attention over time. However, this static/dynamic distinction is not binary, but rather a matter of degree. For instance, the "scrolling news text" information region is somewhere in-between. Viewersneed to linger their gaze longer on this region until the full scroll of text has begun to loop, but then it no longer required viewers' attention thereafter.

7.1.3 Width of peaks throughout periods of sustained attention. The attention curves for our information regions were not completely smooth, but instead, the recordings included "spikes" over time, whenever viewers' attention was drawn to that information region during the videos. Our participants' open-ended comments indicated that their attention was drawn to regions of the screen with movement or large dynamic changes, and this finding aligns with prior work. For instance, some prior eye-tracking studies had revealed how if there is any change or movement in a region's color within the screen, viewers tend to shift their attention to that area [19, 36]. Besides how movement or color change can draw attention [14, 20, 33, 42, 46], our participants' comments revealed why their gaze lingers longer on some regions. For instance, in Group 3, we observed some narrow/sharp peaks when users glanced at information regions that consisted of static text content, such as the speaker's name, only briefly. In contrast, in Group 2, we observed peaks that were broader/wider, indicating longer, sustained attention; the information regions in this group were dynamic information regions containing faces or over-the-shoulder video. Our observations align with prior work had found viewers' gaze lingered on human faces [43].

7.2 Design Implications

For **individual captioning professionals** who are trying to create high-quality captioned videos, our findings inform how to consider where captions should be placed. During television news broadcasts, the density of information regions means that there may be times when there is no perfect location on the screen where a caption should be placed such that it does not block something. Furthermore, during *live* television news broadcasts, time is even tighter for captioning professionals to make decisions about where to put captions on the screen. Specifically, our analysis of attention curves and our grouping of information regions suggests some design considerations for caption placement:

- Group 1: Peak followed by Slowly Decreasing Sustained Attention: During the first few seconds of a news video story, it is especially important that any information regions in this group should not be blocked. Later in the video, it is also better to avoid blocking these high-priority information regions, but not at the expense of blocking the dynamic information regions in Group 2 below.
- Group 2: Sustained Attention: These information regions generally receive continuous attention from DHH viewers; therefore, captioners should not place captions in locations that would block these information regions during a news video. Unlike Group 1 and 3, our study did not reveal any additional priority for these regions during the first few seconds of the news story video.
- Group 3: Low Attention with Some Peaks: While lower priority than information regions in Group 1 or 2, the information regions in Group 3 were higher priority than those in Group 4. However, if necessary to avoid blocking information regions in Group 1 or 2, it could be OK to block these, as long as there were some short gaps in-between caption blocks, such that there were short periods of time when a viewer could briefly see the text content in these regions.
- Group 4: Very Low Attention: These information regions were the lowest priority for DHH viewers. If it is possible to place captions without blocking any information regions, then that is always best, but if necessary, it should not be

problematic to block these regions. Since these are text content regions that can be glanced briefly, even if there are only brief durations of time in-between caption blocks when these regions are visible, this may be enough for DHH viewers to read them briefly.

For **policy makers and captioning regulatory agencies**, our findings can motivate the future development of more specific guidelines for how captions should be placed during television news programs that consider how DHH viewers' attention changes over time. Existing guidelines, e.g., from FCC or DCMP, do not specifically guide broadcasters about which information regions should not be blocked during news videos nor how to prioritize among them. Further, those guidelines do not consider how some regions may be vary in importance over time during a news video. Therefore, a human judge must watch a video and make an intuitive judgment as to how to penalize when a caption occludes other important information on the screen. Our findings suggest how specific information regions during news videos are important during specific periods of time for DHH viewers. Thus, future captioned-video-quality metrics could be invented that penalize occlusions more severely during specific times during a video. For instance, Group 1 occlusions are very bad during the first few seconds, but occasional occlusions of Group-3 regions can be OK as long as the content is briefly visible in-between each caption appearance.

8 LIMITATIONS AND FUTURE WORK

While this study provides empirical insights about how DHH viewers distribute their attention over information regions in news video, future researchers would need to investigate how to update current caption-placement guidelines or evaluation metrics to consider some the findings of this study, specifically the new findings from this study about how attention changes over time during videos.

Since this is an in-person study with eye-tracking equipment conducted at our laboratory, in this study, we had to recruit participants from one geographic location. In future work, a study can be conducted with participants with a wider range of geographic and demographic backgrounds with respect to culture and language, since that might give us a more diverse set data and help us understand whether there is a similarity or difference between various user groups. Furthermore, future research could include a control-group experiment to elicit variation in attention behavior across two groups of people who watch live news videos with and without captions.

To fit within the resource limitations of this project, in this study, we have focused on videos from the news genre, in order to obtain enough measurements to investigate DHH viewers' attention paid to information regions that appear in that specific genre. Future research could investigate viewers' attention distribution across a wider range of information regions and across a wider range of video genres.

Prior research has shown that presence of American Sign Language (ASL) interpreter on screen draws substantial visual attention from DHH viewers [45]. Since regulations in most locations do not require interpreters onscreen (and they rarely appear on local TV news broadcasts), the study in this paper has not included videos with ASL interpreters. However, future research can investigate how inclusion of an ASL interpreter affects viewers' attention distribution.

While selecting the caption properties, such as font size, color, or background, for use in this study, we have selected only a single, standard setting for these properties, based on the most common way in which captions appear in live television programming. In future work, researchers could repeat this study with a wider range of caption appearance properties, to determine whether there is impact on DHH viewers' visual attention distribution over time.

9 CONCLUSION

When captions block useful information in a video, prior work had revealed that this leads to reduced satisfaction among DHH viewers, and there is a need for greater understanding of how DHH viewers' prioritize their attention across regions of the screen, to help inform the work of captioners or broadcasters in deciding where to place captions, especially during news videos which contain a dense amount of information content on the screen. Unlike prior work that solely relied on DHH viewer's subjective judgments of captioned videos [5], our study has made use of eye tracking to investigate DHH viewers' attention across various information regions while watching news videos.

An analysis of eye tracking data revealed a prioritization which information regions should not be blocked, and notably, this ranking differed from rankings in prior work based on subjective judgements—thereby suggesting the value of considering direct observational measures of DHH viewers' attention. Our further analysis revealed a grouping of these information regions, based on DHH viewers' attention patterns over time, with open-ended comments revealing factors that lead to these attention behaviors.

Our study provides empirical evidence of the importance of considering the time dimension when investigating DHH viewers' attention during videos, with a specific focus on how to decide which information regions should not be blocked by captions during news videos. More broadly, our work contributes to the accessibility and HCI research literature on captioning, and our findings inform the work of captioners, policymakers, caption evaluators, and future researchers studying occlusion in the context of captioned videos.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. 2125362, 2235405, 2212303, and 1954284, and by the Department of Health and Human Services under Award No. 90DPCP0002-0100.

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 (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 20–24. https://doi.org/10.18653/v1/P18-2004

[2] Akhter Al Amin, Abraham Glasser, Raja Kushalnagar, Christian Vogler, and Matt Huenerfauth. 2021. Preferences of Deaf or Hard of Hearing Users for Live-TV Caption Appearance. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 189–201.

[3] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Caption-Occlusion Severity Judgments across Live-Television Genres from Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 18th International Web for All Conference* (Ljubljana, Slovenia) *(W4A '21)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3430263.3452429

[4] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Effect of Occlusion on Deaf and Hard of Hearing Users' Perception of Captioned Video Quality. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 202–220.

[5] Akhter Al Amin, Saad Hassan, Sooyeon Lee, and Matt Huenerfauth. 2022. Watch It, Don't Imagine It: Creating a Better Caption-Occlusion Metric by Collecting More Ecologically Valid Judgments from DHH Viewers. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 459, 14 pages. https://doi.org/10.1145/3491102.3517681

[6] Akhter Al Amin and Matt Huenerfauth. 2021. Perspectives of Deaf and Hard-of-Hearing Viewers on Live-TV Caption Quality. In *Illinois Digital Environment for Access to Learning and Scholarship*. Illinois Digital Environment for Access to Learning and Scholarship, Illinois, Chicago, US, 98–107. http://hdl.handle.net/2142/109692

[7] Mihai Bâce, Sander Staal, and Andreas Bulling. 2020. Quantification of Users' Visual Attention During Everyday Mobile Device Interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376449

[8] BBC. 2018. *BBC Subtitle Guidelines, 2018.* British Broadcasting Corporation, Portland Place, London, United Kingdom. https://bbc.github.io/subtitle-guidelines

[9] 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 FaWERctors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312921

[10] Bonnie B. Blanchfield, Jacob J. Feldman, Jennifer L. Dunbar, and 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 (2001), 183–9. http://www.ncbi.nlm.nih.gov/pubmed/11332518

[11] 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.* National Center for Accessible Media (NCAM) WGBH, WGBH (NCAM) One Guest Street Boston, MA 02135.

[12] Andy Brown, Rhia Jones, Michael Crabb, James Sandford, Matthew Brooks, Michael 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. *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* 1, 1. https://doi.org/10.1145/2745197.2745204

[13] Xiuli Chen, Aditya Acharya, Antti Oulasvirta, and Andrew Howes. 2021. An Adaptive Model of Gaze-Based Selection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 288, 11 pages. https://doi.org/10.1145/3411764.3445177

[14] Shun Chiba, Tomo Miyazaki, Yoshihiro Sugaya, and Shinichiro Omachi. 2015. Estimation of gazing points in environment using eye tracker and omnidirectional camera. In *2015 IEEE International Conference on Consumer Electronics - Taiwan*. 47–48. https://doi.org/10.1109/ICCE-TW.2015.7217003

[15] Bogdan Ciubotaru, Gheorghita Ghinea, and Gabriel-Miro Muntean. 2014. Subjective Assessment of Region of Interest-Aware Adaptive Multimedia Streaming Quality. *IEEE Transactions on Broadcasting* 60, 1 (March 2014), 50–60. https://doi.org/10.1109/TBC.2013.2290238

[16] Federal Communications Commission. 2010. *Captioning Key for Educational Media, Guidelines and Preferred Technique. Retrieved from:.* The Described and Captioned Media Program. http://access-ed.r2d2.uwm.edu/resources/captioning-key.pdf

[17] Federal Communications Commission. 2014. *Closed Captioning Quality Report and Order, Declaratory Ruling, FNPRM. Retrieved from:.* Federal Communications Commission., Washington, D.C., USA. https://www.fcc.gov/document/closed-captioning-quality-report-and-order-declaratory-ruling-fnprm

[18] Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2276–2279. https://doi.org/10.1145/3343031.3350535

[19] Géry d'Ydewalle and Wim De Bruycker. 2007. Eye movements of children and adults while reading television subtitles. *European Psychologist 12* 3 (2007).

[20] Hans-Peter Frey, Christian Honey, and Peter König. 2008. What's color got to do with it? The influence of color on visual attention in different categories. *Journal of Vision* 8, 14 (10 2008), 6–6. https://doi.org/10.1167/8.14.6 arXiv:https://arvojournals.org/arvo/content_public/journal/jov/933528/jov-8-14-6.pdf

[21] Saad Hassan, Akhter Al Amin, Caluã de Lacerda Pataca, Diego Navarro, Alexis Gordon, Sooyeon Lee, and Matt Huenerfauth. 2022. Support in the Moment: Benefits and Use of Video-Span Selection and Search for Sign-Language Video Comprehension among ASL Learners. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 29, 14 pages. https://doi.org/10.1145/3517428.3544883

[22] Richang Hong, Meng Wang, Xiao-Tong Yuan, Mengdi Xu, Jianguo Jiang, Shuicheng Yan, and Tat-Seng Chua. 2011. Video Accessibility Enhancement for Hearing-Impaired Users. *ACM Trans. Multimedia Comput. Commun. Appl.* 7S, 1, Article 24 (Nov. 2011), 19 pages. https://doi.org/10.1145/2037676.2037681

[23] Yongtao Hu, Jan Kautz, Yizhou Yu, and Wenping Wang. 2014. Speaker-following video subtitles, In ACM Transactions on Multimedia Computing, Communications and Applications. *ACM Transactions on Multimedia Computing, Communications, and Applications 11* 2, 32.

[24] Dhruv Jain, Bonnie Chinh, Leah Findlater, Raja Kushalnagar, and Jon Froehlich. 2018. Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems* (Hong Kong, China) *(DIS '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 7–11. https://doi.org/10.1145/3197391.3205404

[25] Bo Jiang, Sijiang Liu, Liping He, Weimin Wu, Hongli Chen, and Yunfei Shen. 2017. Subtitle Positioning for E-Learning Videos Based on Rough Gaze Estimation and Saliency Detection. In *SIGGRAPH Asia 2017 Posters* (Bangkok, Thailand) *(SA '17)*. Association for Computing Machinery, New York, NY, USA, Article 15, 2 pages. https://doi.org/10.1145/3145690.3145735

[26] Sushant Kafle and Matt Huenerfauth. 2019. Predicting the Understandability of Imperfect English Captions for People Who Are Deaf or Hard of Hearing. *ACM Trans. Access. Comput.* 12, 2, Article 7 (June 2019), 32 pages. https://doi.org/10.1145/3325862

[27] Ami Klin, Warren Jones, Robert Schultz, Fred Volkmar, and Donald Cohen. 2002. Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals With Autism. *Archives of General Psychiatry* 59, 9 (09 2002), 809–816. https://doi.org/10.1001/archpsyc.59.9.809

[28] Anida Krajina and Dušan Mladenović. 2018. Colliding the eye Tracking and Qualitative Methods in Assessing Online Consumer Behaviour. *European Conference on Research Methodology for Business and Management Studies* 13 (07 2018), 457–464.

[29] Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. *Close to the action: Eye-tracking evaluation of speaker-following subtitles.* Association for Computing Machinery, New York, NY, USA, 6559–6568. https://doi.org/10.1145/3025453.3025772

[30] Kuno Kurzhals, Fabian Göbel, Katrin Angerbauer, Michael Sedlmair, and Martin Raubal. 2020. A View on the Viewer: Gaze-Adaptive Captions for Videos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376266

[31] Michael Lankes and Florian Berger. 2021. Blind Spot: An Interactive Gaze-Aware Experience. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 205, 4 pages. https://doi.org/10.1145/3411763.3451570

[32] Hailiang Li, Bin Zhang, Yu Zhang, Weiwei Liu, Yijun Mao, Jiacheng Huang, and Linfeng Wei. 2020. A semi-automated annotation algorithm based on weakly supervised learning for medical images. *Biocybernetics and Biomedical Engineering* 40, 2 (2020), 787–802. https://doi.org/10.1016/j.bbe.2020.03.005

[33] Zheng Ma, Jiaxin Wu, Sheng-hua Zhong, Jianmin Jiang, and Stephen J. Heinen. 2019. Human Eye Movements Reveal Video Frame Importance. *Computer* 52, 5 (May 2019), 48–57. https://doi.org/10.1109/MC.2019.2903246

[34] Sheila A. McIlraith. 1989. Qualitative data modeling: application of a mechanism for interpreting graphical data. *Computational Intelligence* 5, 2 (1989), 111–120. https://doi.org/10.1111/j.1467-8640.1989.tb00320.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.1989.tb00320.x

[35] Danni Mulrennan. 2013. Deaf See TV News Differently A First Step in Finding a better Path to Knowledge. *Multimodal Communication* Vol 2(1): 85-107 (04 2013).

[36] Andrew D. Ouzts, Nicole E. Snell, Prabudh Maini, and Andrew T. Duchowski. 2013. Determining Optimal Caption Placement Using Eye Tracking. In *Proceedings of the 31st ACM International Conference on Design of Communication* (Greenville, North Carolina, USA) *(SIGDOC '13)*. Association for Computing Machinery, New York, NY, USA, 189–190. https://doi.org/10.1145/2507065.2507100

[37] Anne-Flore Perrin, Lu Zhang, and Olivier Le Meur. 2020. Inferring Visual Biases in UAV Videos from Eye Movements. *Drones* 4, 3 (2020). https://doi.org/10.3390/drones4030031

[38] Cole S. Peterson, Jonathan A. Saddler, Natalie M. Halavick, and Bonita Sharif. 2019. A Gaze-Based Exploratory Study on the Information Seeking Behavior of Developers on Stack Overflow. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312801

[39] Hamza Polat. 2020. Investigating the use of text positions on videos: An eye movement study. *Contemp. Educ. Technol.* 12, 1 (Feb. 2020).

[40] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 3 (1998), 372–422.

[41] Pablo Romero-Fresco and Juan Martínez Pérez. 2015. *Accuracy Rate in Live Subtitling: The NER Model.* Palgrave Macmillan UK, London, 28–50. https://doi.org/10.1057/9781137552891_3

[42] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia. 2019. DEEP-HEAR: A Multimodal Subtitle Positioning System Dedicated to Deaf and Hearing-Impaired People. *IEEE Access* 7 (2019), 88150–88162. https://doi.org/10.1109/ACCESS.2019.2925806

[43] Margot van Wermeskerken, Susanna Ravensbergen, and Tamara van Gog. 2018. Effects of instructor presence in video modeling examples on attention and learning. *Computers in Human Behavior* 89 (2018), 430–438. https://doi.org/10.1016/j.chb.2017.11.038

[44] Xi Wang, Andreas Ley, Sebastian Koch, David Lindlbauer, James Hays, Kenneth Holmqvist, and Marc Alexa. 2019. The Mental Image Revealed by Gaze Tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300839

[45] Jennifer Wehrmeyer. 2014. *Eye-tracking Deaf and hearing viewing of sign language interpreted news broadcasts.* Journal of Eye Movement Research, Moosgasse 16 CH-3305 Iffwil Switzerland.

[46] Jing Zhang, Li Zhuo, Zhenwei Li, and Yingdi Zhao. 2012. An approach of region of interest detection based on visual attention and gaze tracking. In *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*. 228–233. https://doi.org/10.1109/ICSPCC.2012.6335613

[47] Kaixing Zhao, Sandra Bardot, Marcos Serrano, Mathieu Simonnet, Bernard Oriola, and Christophe Jouffrais. 2021. Tactile Fixations: A Behavioral Marker on How People with Visual Impairments Explore Raised-Line Graphics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. https://doi.org/10.1145/3411764.3445578