# Watch It, Don't Imagine It: Creating a Better Caption-Occlusion Metric by Collecting More Ecologically Valid Judgments from DHH Viewers

Akhter Al Amin*
Saad Hassan*
aa7510@rit.edu
sh2513@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Sooyeon Lee
Rochester Institute of Technology
Rochester, New York, USA
slics@rit.edu

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

## ABSTRACT

Television captions blocking visual information causes dissatisfaction among Deaf and Hard of Hearing (DHH) viewers, yet existing caption evaluation metrics do not consider occlusion. To create such a metric, DHH participants in a recent study imagined how bad it would be if captions blocked various on-screen text or visual content. To gather more ecologically valid data for creating an improved metric, we asked 24 DHH participants to give subjective judgments of caption quality after actually watching videos, and a regression analysis revealed which on-screen contents' occlusion related to users' judgments. For several video genres, a metric based on our new dataset out-performed the prior state-of-the-art metric for predicting the severity of captions occluding content during videos, which had been based on that prior study. We contribute empirical findings for improving DHH viewers' experience, guiding the placement of captions to minimize occlusions, and automated evaluation of captioning quality in television broadcasts.

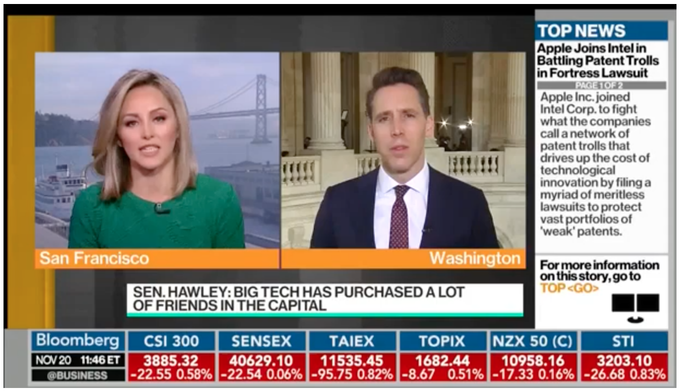
## CCS CONCEPTS

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

## KEYWORDS

Accessibility, Caption, Metric, Regression

**ACM Reference Format:**
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 (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3517681



Apple Joins Intel in Battling Patent Trolls in Fortress Lawsuit

Apple Inc. joined Intel Corp. to fight what the companies call a network of patent trolls that drives up the cost of technological innovation by filing a myriad of meritless lawsuits to protect vast portfolios of 'weak' patents.

For more information on this story, go to TOP <GO>

SEN. HAWLEY: BIG TECH HAS PURCHASED A LOT OF FRIENDS IN THE CAPITAL

| CSI 300 | SENSEX | TAIEX | TOPIX | NZX 50 (C) | STI |
|---|---|---|---|---|---|
| 3885.32 | 40629.10 | 11535.45 | 1682.44 | 10958.16 | 3203.10 |
| -22.55 0.58% | -22.54 0.06% | -95.75 0.82% | -8.67 0.51% | -17.33 0.16% | -26.68 0.83% |

**Figure 1: Sample screen layout of a TV news program, showing multiple information content regions.**

## 1 INTRODUCTION

More than 360 million people across the world experience hearing loss [7], and 15% of US adults who are Deaf and Hard-of-Hearing (DHH) rely on captioning services while watching television programming [2, 4]. To ensure DHH viewers' access to spoken information, captioning is required for television programming – whether the program has been pre-recorded, live broadcast, or nearly live. Broadcasters typically employ human captioning services to generate verbatim captions or make use of semi-automatic captioning approaches [63]. During live television programming in particular, such as news, weather, or sports programs, broadcasters tend to include a great deal of visual information content on the screen, as illustrated in Figure 1 [47, 48]. With this onscreen information density, captioners may find it challenging to place a caption without blocking other onscreen textual or graphical information, e.g., text of a news headline, an onscreen speakers' name, sports scores, or charts with numerical data. Although in pre-recorded TV programs, captioners may have the opportunity to strategically place captions in salient locations, due to limited time for production and placement, especially for live or near-live programming, captions generally appear in standard locations, e.g., the lower or upper region of the TV screen [45, 54]. Prior research has revealed DHH viewers' dissatisfaction with captioning when there is occlusion, even if the caption is an accurate transcription of what has been spoken [4, 6, 44].

*Both authors contributed equally to this research.

*CHI '22, April 29-May 5, 2022, New Orleans, LA, USA*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9157-3/22/04...$15.00
https://doi.org/10.1145/3491102.3517681

In order for regulators or organizations to inexpensively monitor the quality of captioning provided in television broadcasts, it is valuable to have automatic metrics that can assess the quality of the captioning, e.g. based on the accuracy of the text or latency relative to the speech [5, 13, 14]. Several such metrics have been proposed or are in common use, e.g. [1, 9, 32, 49], yet most focus only on the text of the caption, rather than its placement on the screen. We, therefore, investigate how captions **occluding onscreen information** influence the quality of captioning from DHH viewers' perspective, as some initial work has revealed how such occlusion can pose problems for DHH viewers, e.g. [2, 23, 24]. In particular, there is a need for quantitative measures of how captions blocking specific on-screen elements may affect DHH viewers' satisfaction, since such data could serve as a basis for metrics to evaluate caption placement in television programs. Such metrics could also guide the work of someone who needs to select a placement for a caption when there are trade-offs about what is blocked, especially during television programs with dense on-screen content.

One prior study had attempted to gather such quantitative information about how DHH viewers' subjective judgments are affected when captions block various types of on-screen information content during television programming [3]. That prior work focused on six popular live-television genres (news, weather, sports, emergency announcements, interviews, and political debates), and for each genre, DHH participants were asked to consider a diagram indicating various regions of the television screen, e.g. the eyes of the speaker, text indicating a news headline, etc. Notably, participants were not shown a video and asked to rate its quality. Instead, for each component region of the screen shown in the diagram, participants were asked to imagine how bad it would be if it were blocked by a caption. The researchers then created a metric for calculating the severity of caption occlusion in an entire video, based on the severity values that participants had imagined for each component of the screen. This metric, which is the current state of the art, is referred to as the **Component Judgment Model** in this paper. A concern with this approach to data collection is that participants may not be able to introspect how bad it would be to block a piece of information, especially when viewing a static diagram and imagining a dynamic video.

In this paper, we investigate whether it would be more ecologically valid to instead display videos to DHH participants, with various caption placements, some of which may block elements of the screen, and to simply ask participants to give a holistic, subjective rating as to the quality of the caption placement in the video. Regression modeling can then be used to examine how captions blocking specific regions of the screen relate to DHH viewers' overall holistic judgment of the quality of the captioning. We refer to the metric resulting from this alternative approach to gathering quantitative subjective judgments from users as the **Holistic Judgment Model**. In fact, a similar methodology has been used successfully to gather subjective feedback from Blind and Visually Impaired (BVI) individuals, to build a predictive regression model of what factors influence their overall judgment of the quality of an online video search experience [40]. However, this method has not been employed previously to create models of video captioning quality among DHH viewers, nor specifically for the issue of predicting caption-occlusion severity.

To address this gap, in this paper, we have conducted a study to gather judgments from DHH participants using the **holistic** methodology described above. Our 24 DHH participants rated the caption-placement quality of a set of video stimuli we created, across the set of television genres identified in that prior study [3], e.g., weather or sports, with each video containing the set of on-screen information regions identified in that prior work, e.g. the speaker's eyes, the current news headline, or the current sports game score. We conducted linear regression modeling to predict participants' subjective rating of each video, based on features that included the time-duration or degree to which each information region was occluded by a caption. Using a feature engineering approach [33], we have derived a best-fit regression model for predicting DHH viewers' subjective judgements of caption-occlusion severity, for each television genre. A relative-importance analysis of our models' features revealed that our holistic data-collection approach yielded differences in which features were judged to be important by DHH participants, as compared to those found in prior work [3]. Further, an evaluation revealed that a caption-occlusion severity metric based on our new model outperformed the metric produced in that prior study [3].

The contributions of our work are empirical:

- We identify best-fit models for predicting how DHH viewers would subjectively rate the quality of caption placement, for television videos of a variety of genres, based on considering the degree to which text, people, or visual content in the videos are occluded by captions. We provide evidence that our models are capable of explaining a significant amount of the variance in DHH viewers' judgements of caption-placement quality.
- Our relative-importance analysis reveals that our holistic data-collection and modeling approach led to different insights than the component/imagination approach in prior work [3]. Specifically, the degree to which occlusion of various regions of the screen was important to each model differed.
- We present evidence that our new metric for calculating a caption-occlusion severity score for a video out-performed the prior state-of-the-art model. We thereby contribute a superior tool for retrospectively evaluating television caption-placement quality or for prospectively guiding caption placement in television videos. We disseminate a software implementation of our caption-occlusion metric for use by practitioners, or for use by the research community in replicating our work.

## 2 BACKGROUND AND RELATED WORK

To evaluate the quality of TV captioning services, several automatic and semi-automatic caption evaluation **metrics** have been introduced. Most of these metrics focus on caption text itself, that is, how accurately it has transcribed all of the spoken words in the video. Some fully automatic metrics, such as Word Error Rate (WER) [1] penalize each insertion, deletion, or replacement error, while other metrics, such as Weighted Word Error Rate (WWER) [9] or Automatic Caption Evaluation (ACE) [32], weight errors based on linguistic factors. Other approaches require humans to make judgements about the severity of each text error, e.g. [49].

Beyond the text accuracy of the caption, some guidelines for humans who are assessing caption quality discuss other factors that should be considered, including: synchronicity between caption and audio, uniformity in style and presentation, readability, and avoiding occlusion with onscreen information [5, 13, 14]. While prior studies investigating the preferences of DHH viewers have often focused on appearance and style aspects of captions, e.g. [2, 6, 16, 21, 36, 46, 58, 60], there has been some research with DHH participants that has examined how their viewing experience is diminished when captions occlude other onscreen information [6, 10, 37, 62].

Section 2.1 will discuss how some research has examined how to automatically select caption placements to, in part, reduce occlusion; however, such work has considered occlusion of relatively few elements, e.g. faces. Section 2.2 describes the aforementioned study [3] that collected judgements from DHH participants to produce a metric to automatically calculate the severity of caption occlusion in a video. Section 2.3 describes a more ecologically valid methodology for collecting judgements from participants when designing a predictive metric of their subjective preferences. The goal of our new study is to investigate how utilizing this methodology may yield a higher-quality dataset of DHH participants' judgments—and ultimately an automatic metric of occlusion severity that out-performs the current state of the art.

## 2.1 Prior Work on Automatic Selection of Caption Placement

Researchers have investigated various approaches for automatically selecting where captions should appear on screen. Some have focused on placing captions close to person who is currently speaking [25, 26, 55, 55]. While changing the location of captions too often can place a burden of viewers, who must visually seek the caption on screen [34], such dynamic placement technologies generally improve DHH viewers' experience [10, 35]. Most relevant to our current study, some prior research has detected a few important regions of the screen, e.g., the face of the person speaking, and attempted to avoid blocking those when automatically placing a caption [30].

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 [45, 62]. 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 [62]. Moreover, DHH viewers spend a significant amount of time looking at the caption itself while watching captioned videos [45], therefore collecting gaze information from DHH viewers may not be an effective approach for determining preferred caption location – since their gaze may naturally be drawn to wherever the caption is located in the video, rather than revealing other important regions of the screen that captions should not block.

Overall, while researchers have proposed several methods of automatically placing captions, relatively little work has explored how to avoid captions occluding other on-screen content. While eye-tracking data has guided some work on determining salient regions of video, there are challenges in utilizing this approach to determine, from DHH viewers' perspective, which regions of the screen should not be occluded by captions.

## 2.2 Prior Work on Caption Evaluation Metrics that Consider Occlusion

While the caption-placement research above has examined *prospectively* where a caption should be placed, there has also been some work on how caption-occlusion could be incorporated into *retrospective* metrics to evaluate caption-placement quality during a previously-captioned video. Such work has been motivated by focus-group and experimental research that has revealed that DHH viewers are concerned about captions blocking other on-screen content [4].

In the most closely related prior work, researchers collected a dataset of judgements from DHH participants, to create a metric to automatically assess the severity of captions occluding other onscreen content [3]; as mentioned above, we refer to this prior state-of-the-art metric as the Component Judgement Model in this paper. That prior work focused on several popular **genres** of live television programming [41], including: news, weather news, political debates, interviews or talk-shows, emergency announcements, and sports. The researchers described how television programs in each genre make use of typical **layouts** of on-screen **information regions** [3], e.g., news headline text, the eyes and mouth of anyone on-screen who is speaking, temperature numbers on a map during weather news, game score or player statistics during a sports broadcast, of a news presenter's face, name, job title, or location. For example, in a television news broadcast, a common layout may include a news presenter who looks at the camera while presenting news, with text content along the bottom of the screen indicating the headline and an information graphic appearing above the presenter's shoulder. Another common camera view and layout may be a reporter who presents information from a remote location, again with text content on the screen that may indicate their location or name [17].

As discussed in section 1, the researchers [3] asked DHH participants to view static line-drawing diagrams of typical layouts of information regions. Participants were asked to consider each component of the screen and imagine that they had been watching a video similar to that diagram. For each information region, the participant was asked to provide a numerical score indicating how bad it would be if a caption were to block that part of the screen. The researchers repeated this process for several diagrams, illustrating typical layouts and information regions, for all six of the live-television genres. These set of judgements were subsequently used as penalty weights within a metric that considered when a caption occluded one of these information regions – with their metric based on both the **occlusion percentage** (area of the region blocked by the caption) and **occlusion time** (the amount of time the caption occluded that region). For instance, if participants had given a high score to indicate that it would be very bad for a caption to block the mouth of the person who is speaking, then the metric would give a high penalty weight when considering the occlusion time and percentage of any caption blocking that region in a video.

While the Component Judgement Model is the current state-of-the-art metric for assessing the severity of caption-occlusion in videos, there are several key limitations of that prior study [3]. Some limitations include:

(1) Participants had been asked to imagine watching a dynamic video when shown a static diagram.
(2) Participants had been asked to make judgements about individual components (captions blocking specific information regions) yet the aim of the research was to create a metric of DHH viewers' holistic judgement of caption-placement in a video.
(3) The dataset contained judgments about information regions only for specific layouts of 6 television genres.

As discussed below, the goal of our current study is to address the limitation (1) and (2) from the above list by utilizing an alternative data-collection methodology that may be more ecologically valid [52]; however, for comparison purposes in this paper, we will focus on the same set of television genres, screen layouts, and information regions that had been established in the prior study [3].

## 2.3 Prior Work on Data-Collection among Specific User Groups for Factor Analysis

There are a variety of common methodological approaches used within the HCI research literature for conducting analysis and modeling of how component factors may influence a holistic quantitative value [19, 38, 51], e.g., for modeling a score of usability or users' subjective preference [12, 31, 40]. Some work has investigated how to gather data on the factors that influence the subjective judgements of specific user groups [40, 42, 53]. For instance, Xingyu et al. [40] identified factors that affect how accessible an online-video search experience would be for Blind and Visually Impaired (BVI) users. The researchers first identified a large set of potential factors and then conducted a study in which BVI users performed tasks with a prototype system and reported their overall subjective judgment about how accessible their experience had been. Finally, the researchers utilized a regression-modeling approach to create a metric that could predict BVI users' subjective assessment of the system's accessibility, based on various factors. Notably, unlike the data-collection among DHH participants for the prior Component Judgement Model [3], Xingyu et al. [40] had asked their BVI participants to provide a holistic judgement about the system–specifically, a judgement which was the output of the predictive model they were creating.

## 3 RESEARCH QUESTIONS

No prior study on caption-occlusion metrics among DHH participants has made use of this holistic-judgment approach; our study will address this gap in the literature and determine whether this alternative methodology will yield a more accurate metric. To address this primary research question (RQ3 below), we first investigated two foundational research questions: to understand how much variance in DHH viewers' subjective judgments was explained by our new modal and whether the resulting model had actually learned a different weighting among the independent variables that the older component-judgment model. With that foundation, we finally conducted an extrinsic comparison between the new holistic-judgment model and the prior component-judgment model. Our research questions included:

**RQ1: How much variance in participants' holistic judgment of caption quality can we predict using regression based on occlusion features?** We sought to understand whether the regression modeling itself had been successful at explaining a significant amount of the variance.

**RQ2: Does this holistic approach to learning a model of users' subjective preferences through regression analysis differ in which factors are important to the model, as compared to the prior component-judgement model?** We conducted a relative-importance analysis to investigate whether the features within our new best-fit regression model differed from the weights within the prior state-of-the-art model.

**RQ3: Does the score produced using our new Holistic Judgment Model correlate significantly better with DHH viewers' subjective preference than the score generated from the prior Component Judgment Model?** Finally, as the primary measure of whether this new approach yielded a superior model, we compared our new caption-occlusion severity metric to the prior state-of-the-art metric. For this analysis, we made use of an existing dataset [3] of captioned videos whose caption placement had been subjectively rated by DHH viewers, to determine which metric was better correlated with human judgements.

## 4 CREATION AND EVALUATION OF REGRESSION MODELS

To collect judgments from DHH participants about the quality of caption-placement in videos, it was first necessary for us to assemble a set of video stimuli. Section 4.1 describes how videos of multiple television genres, with a variety of information regions on the screen, were selected, as well as how captions were placed in a variety of locations so that they sometimes occluded these information regions. Section 4.2 describes how each video was annotated to identify the amount of time and the percentage of the area of each information region that was blocked by a caption. Section 4.3 describes the conduct of a data-collection study with DHH participants who viewed the stimuli videos and provided subjective ratings of the holistic caption-placement quality of each video. Finally, section 4.4 describes how we conducted a regression analysis using the occlusion annotations for each video from Section 4.2 as *input* features and the numerical judgements from participants from Section 4.3 as the *output* prediction. Following the approach of prior work [3], this modeling was performed separately for each genre of television programming, and we analyzed the variance explained by each model to address research question RQ1.

## 4.1 Methods

*4.1.1 Construction of Video Stimuli.* Our goal in assembling a set of video stimuli for our data collection process was to ensure that our videos satisfied several criteria:

(1) We must include videos from all 6 popular live-television genres [57], matching those studied in prior work [3].
(2) In each video, a set of information regions should be present on the screen. To enable comparison, the set of information regions should match those in prior work [3].

(3) Videos should not include contentious or emotionally upsetting topics that may affect participants' preferences.
(4) Multiple versions of each video should be produced, with the caption at different placement locations, such that different subsets of the on-screen information regions are blocked by the caption.

For the comparison purposes between two models, in our study, we have used the same set of information regions as in prior work [3], which are enumerated in Appendix A. Some of these regions are relatively fine-grained, e.g., the eyes and the mouth of the person speaking are two distinct information regions in that dataset. We speculate that prior researchers [3] had made this choice based on the various information conveyed by the eyes and mouth of a speaker. For instance, DHH viewers who use speechreading focus on a speakers' mouth to perceive spoken information [2], and human emotion is highly correlated with eye and eyebrow movements [59].

We selected 104 videos to include in our stimuli bank, by searching several online sources, including YouTube and Vimeo, which included recordings of live-television broadcasts across multiple national and local TV channels, e.g. CNN, ABC News, Fox News, ABC 8, NBC 26, 10 news, and sports-related TV channels, e.g. Fox Sports, ESPN. Similar to considerations described in prior work [3], when selecting the proportion of videos from each of the six genres, we considered the diversity of screen layouts used within each genre. For instance, if the overall arrangement of information regions on the screen is relatively homogeneous within some genre, then relatively fewer examples of that genre were included our stimuli bank, which included: 24 news videos, 16 emergency-announcement videos, 16 interview videos, 8 political-debate videos, 24 weather-news videos, and 16 sports videos.

Prior work revealed that a 30-second video stimulus was sufficient for obtaining judgements from viewers about caption-occlusion severity [2, 4]. Also, we speculate that prior research has done this to maximize the number of stimuli they can display during the study. Therefore, we trimmed each video in the stimuli bank to approximately 30 seconds. We carefully produced accurate text captions for each video. When multiple speakers appeared in a video, changes in speaker were indicated with "»" (double chevron) in the caption, following standard guidelines and recent research [13]. Captions were generated using standard colors (white font on a solid black background), as recommended by prior work [2]. To simulate typical live-television captioning latency, with text appearing approximately 3-6 seconds after the speaker [50], we set the latency for our captions to approximately 3 seconds. To segment longer spans of text across multiple captions, we followed guidelines in recent research [61]. In this way, we prepared the caption files, and then captions were burned into the video stimuli.

In support of our research objective, multiple versions of each video had to be produced with captions in different locations, such that a diverse range of information regions were blocked, across the stimuli. The stimuli included a variety of videos in which captions blocked subsets and combinations of information regions, e.g., with some blocking both the speakers' eyes and mouth while other videos with captions only blocking one region. At the same time, to maintain ecological validity, we wished to avoid absurd placements of captions in unusual locations on screen. We, therefore, added captions to each video in four static locations (nearly top of screen, upper third, lower third, nearly bottom of screen), as illustrated in Figure 2. In this way, captions remained within either the lower 20 vertical lines or upper 20 vertical lines, leaving the center 60 vertical lines unblocked, as recommended by existing TV broadcast guidelines [54]. Given that we produced 4 captioned versions of each of the 104 videos in our stimuli bank, a total of 416 video stimuli were created for subsequent evaluation by DHH participants.

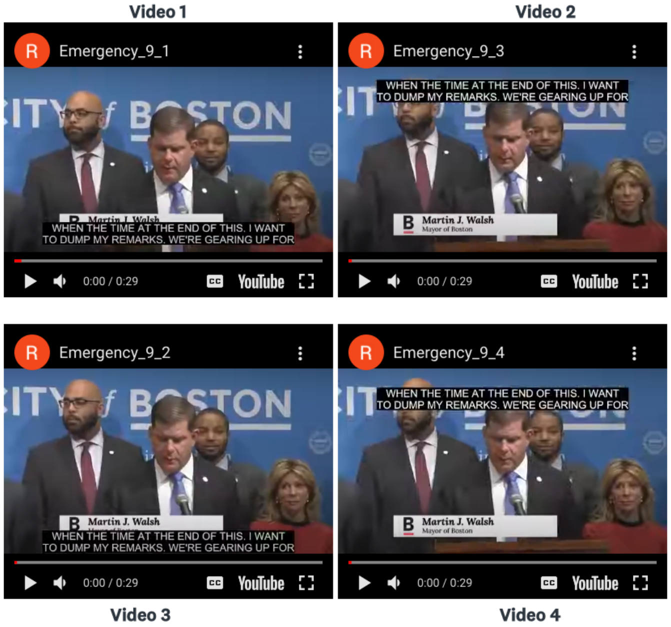


**Figure 2: Four versions of a video, with varied caption placement, as presented on a single screen of our data-collection website.**

*4.1.2 Video Stimuli Annotation.* Since a goal of our current study was to understand the relationship between captions occluding regions of the screen and the overall subjective judgement of DHH viewers of the quality of caption placement, it was necessary for us to examine each stimulus video to determine which regions were blocked by captions. For comparison purposes, our annotation process was adapted from prior work [3]. Specifically, a member of our research team annotated, for each information region, within each video stimulus, the following information:

- **Occlusion Percentage** is a value in the range of 0% to 100%. If the total area of an information region, e.g., the current news headline text, is blocked by a caption, then the occlusion percentage will be 100%. If a portion of the headline text is blocked by a caption, then the occlusion percentage will be a value between 0% and 100%, and if none is blocked, the percentage is 0%. Because the portion of an information region that was blocked by captions may vary throughout a video, for completeness, we included both the minimum and maximum percentage of occlusions as candidate variables.

Whereas maximum occlusion may indicate the greatest degree to which something was blocked (thus detracting from viewing experience), minimum occlusion may represent the degree to which something was actually visible at some point during the video (thus giving someone an opportunity to view the information).

- **Occlusion Time** is also a value on the range between 0% to 100%. We counted the number of video frames in which captions blocked an information region, divided by the total number of frames. For instance, out of 900 frames, if an information region is blocked by captions for 450 frames, the occlusion time is 50%.

*4.1.3 Data-Collection Study.* Pilot testing had revealed that a participant was able to view and provide numerical subjective judgements of caption-placement quality for approximately one quarter of our stimuli videos during a one-hour appointment. To avoid fatigue, we randomly divided our video stimuli into four partitions (maintaining the proportional mix of genres from our entire dataset within each partition). Each participant made four one-hour appointments to view and provide judgements on each partition of the stimuli set, such that at the conclusion of their fourth appointment they had provided judgements for every video in our stimuli set.

Due to COVID-19 safety guidelines, we conducted this study remotely using video conferencing, with questions hosted on the SurveyMonkey website. Video stimuli were embedded within the questionnaire using private YouTube links. Each screen of the website displayed four versions of a video (with four different caption placements), as shown in Figure 2. For each of the four videos, the participant was asked to respond to a scalar item **"How happy are you with location of the caption in the video?"** using a 10-point response scale with the end-points labeled as "Extremely Unhappy" and "Extremely Happy." This question item had been used successfully in prior work [3].

During the study, participants viewed all videos of a particular genre, e.g. sports, contiguously. However, the following were randomized for each participant in the study: (a) the order in which each genre was displayed, (b) the order in which individual videos were displayed within each genre, and (c) the arrangement of particular videos on the screen on a single page.

Participants were recruited by sending out Institutional Review Board-approved advertisements to social network groups and university-related student groups. Each ad included two screening questions: (1) "Do you identify as Deaf or Hard of Hearing?" (2) "Do you use captioning when viewing videos or television?" Participants were identified as qualified to participate in this experiment if they responded with yes to both questions. In this study, we decided to recruit 24 participants, guided by prior research [28] on how to select an appropriate sample size when conducting a multiple regression analysis with user data. Participants included 14 men, 7 women, and 3 individuals who identified as non-binary. Their mean age was 29.33 years (SD=9.16). Nineteen participants identified as deaf, and 5 identified as hard of hearing. Participants indicated spending an average 3.4 hours per week watching captioned TV programming.

In this remote study during COVID-19, although we did not specifically require all participants to use monitors of a particular size, we did control characteristics of the hardware and video display properties: Participants were required to use a personal computer or laptop for the study; participation through tablet or smartphone was disallowed. We also required participants to display the study website in a full-screen manner, with the video displayed at a 4:3 aspect ratio. Finally, it is important to note that since we had "burned in" captions into the video (meaning that the caption text was displayed as actual pixels of the video image), we retained precise control over what information regions were blocked by captions in each stimulus, regardless of the screen or monitor size of the participants.

At the beginning of the study, after connecting with a researcher who was a hearing ASL signer on a Zoom video conference session. the participant completed an informed consent form. Then, the researcher explained the goal of the study and provided instructions to the participant, who was provided the link to the survey website. The researcher remained available on the Zoom session in case of any questions. To avoid fatigue during the one-hour appointment, participants were encouraged to take a short break at the middle of the hour. Participants received compensation of $40 for each one-hour appointment session. Participants were instructed to watch each video on the page beginning with the top left video and ending with the bottom right, and they were asked to fill out the scalar response question after watching each individual video. If a participant wished, they were permitted to view a video more than once. At the conclusion of the one-hour session, the participant responded to some demographic questions.

*4.1.4 Multiple Regression Analysis.* We conducted multiple regression analysis to examine how factors related to the occlusion of information regions in each video could predict participants' overall subjective judgment of the caption-placement quality. For comparison with prior work [3], we created a model for each of the 6 television genres.

Prior to our analysis, we calculated the pair-wise correlation between all three annotations (maximum occlusion percentage, minimum occlusion percentage, and occlusion time) for each each information region, and we observed high collinearity ($> 0.7$). **Collinearity** in multi-variate regression analysis refers to the phenomena in which several independent variables used within a model are correlated with one another [18]. When creating regression models, it is generally considered undesirable for there to be collinearity among the independent variables within a model, because it makes the model less interpretable, i.e., it is not possible to determine how much variance in the dependent variable could have been predicted by each independent variable alone. Our collinearity analysis motivated us to avoid including more than one form of occlusion measurement (maximum percentage, minimum percentage, or time) for any single information region within a single model. Thus, for each model, for each information region, we conducted individual correlation analyses between DHH participants' overall judgement for that video and each of the three forms of occlusion measurement (maximum percentage, minimum percentage, or time). We thereby determined, for each information region, which occlusion measurement explained the most variance, and only this measurement was considered during model creation. For instance, for a specific genre, if maximum occlusion percentage of "speaker's eyes" explained more variance in DHH participants' judgements than minimum occlusion percentage or occlusion time of "speakers' eyes" did, then for our modeling, we included only the maximum occlusion percentage for "speakers' eyes" for potential inclusion within that model.

## 4.2 Findings for RQ1: Variance Explained by Each Model

Our regression analysis yielded a best-fit model for each of the six television genres:

- **News:** Adjusted $R^2 = 0.132, F(15, 568) = 6.895, p < 0.001$
- **Weather News:** Adjusted $R^2 = 0.282, F(13, 444) = 14.81$, $p < 0.001$
- **Sports:** Adjusted $R^2 = 0.2213, F(8, 359) = 14.03, p < 0.001$
- **Emergency Announcements:** Adjusted $R^2 = 0.176$, $F(15, 328) = 5.873, p < 0.001$
- **Interviews:** Adjusted $R^2 = 0.095, F(11, 172) = 2.752$, $p < 0.01$
- **Political Debates:** Adjusted $R^2 = 0.0482, F(10, 173) = 1.927$, $p < 0.05$

The Adjusted $R^2$ score refers to the proportion of the variation in the dependent variable (DHH viewers' judgement of the caption-placement quality of a video) that is explained by the independent variables (the amount of occlusion of various information regions by the caption). Prior research has discussed how even when Adjusted $R^2$ values are small, if they are significantly different from 0 (as indicated by the p-values above), the regression model has statistically significant explanatory power [27]. In this case, while occlusion of information regions explains a significant portion of DHH viewers' judgement of the quality of caption-placement, there may be other factors that also contribute to their judgement, e.g., perhaps whether captions appear close to the person who is speaking.

In regard to research question RQ1, all six regression models revealed a significant relationship between captions occluding information regions and DHH viewers' subjective judgements of the overall caption-placement quality for that video. Our models were most effective at explaining the variance in participants' subjective judgements for the Weather-News and Sports genres, which had the highest Adjusted $R^2$ scores, as listed above.

Table 2 in Appendix A provides a detailed summary of the coefficients for each information-region-occlusion feature in the best-fit regression model for each genre. The set of features within each model is discussed in more detail in section 5 below, which presents the results of a relative-importance analysis.

# 5 COMPARISON OF FEATURES IN THE NEW HOLISTIC MODEL AND PRIOR MODEL

The results in section 4.2 suggest that the holistic-judgment models for each genre explained a significant portion of the variance of DHH viewers' subjective judgements about caption-placement quality in videos. However, to address research question RQ2, we must examine whether the occlusion features used within our new models differ from those in the prior state-of-the-art metric [3]. Such an analysis would reveal whether the imagination of participants in that prior study about which occlusions would most severely affect their viewing experience differed from the actual relationships revealed through our new data-collection and regression modeling approach.

## 5.1 Methods

The coefficients for each feature in our best-fit regression models are sensitive to the order in which features were added to each model during its construction. Therefore, for more meaningful interpretation, we have conducted a relative-importance analysis of the contribution of each information-region occlusion, using the Linderman-Merenda-Gold (LMG) metric [12, 39]. The results of this analysis for each model is displayed in Figure 3(a-f), which also lists the occlusion features included in each model. The LMG metric identifies a percentage of the R-squared that had been explained by each predictor variable in the model [22], thus the bars shown in each graph sum to 100%. We employed bootstrap to estimate the variability of each relative-importance value, to calculate 95% confidence intervals for each (displayed as thin whisker lines for each bar), to reveal which features contribute significantly to each model.

## 5.2 Findings for RQ2: Comparison of Features in New and Prior Models

The goal of this analysis was to determine whether the set of information-region-occlusion features most important in our new model differed in comparison to those which had been most important in the prior state-of-the-art caption-occlusion metric [3]. Therefore, for each genre, we list below the top 3 most predictive information-region features in our new models (i.e., the longest bars in the graphs in Figure 3) and the 3 most important information-regions for each genre in the prior component-judgement model, as had been reported in prior work [3]:

- **News:** As shown in Figure 3(a), occlusion of the scrolling news (text listing other news stories), speakers' eyes, over-the-shoulder text (appearing behind the news presenter), and current discussion (text headline of the current news story) explained 15%, 15%, 12% and 10% of total variance respectively. In comparison, the three most important occlusion features in the News model in prior work [3] had been occlusion of: the speakers' mouth, the current discussion topic, and the listeners' face. **All of the top-3 features differed.**
- **Weather News:** Figure 3(b) reveals that 60% of the total variance was explained by top-3 predictor variables: discussion topic (25%), speakers' eyes (23%), over the shoulder text (19%). On the other hand, the top-3 contributing variables of the prior Weather-News model [3] were occlusion of: over-the-shoulder text, current time and temperature, and the speakers' mouth. **Two of the top-3 features differed.**
- **Sports:** Figure 3(c) indicates that 60% of total variance was explained by: current score of the game (27%), text displaying players' statistics (21%), the current quarter of the game (17%). In contrast, the top-3 features of the prior model [3] were:

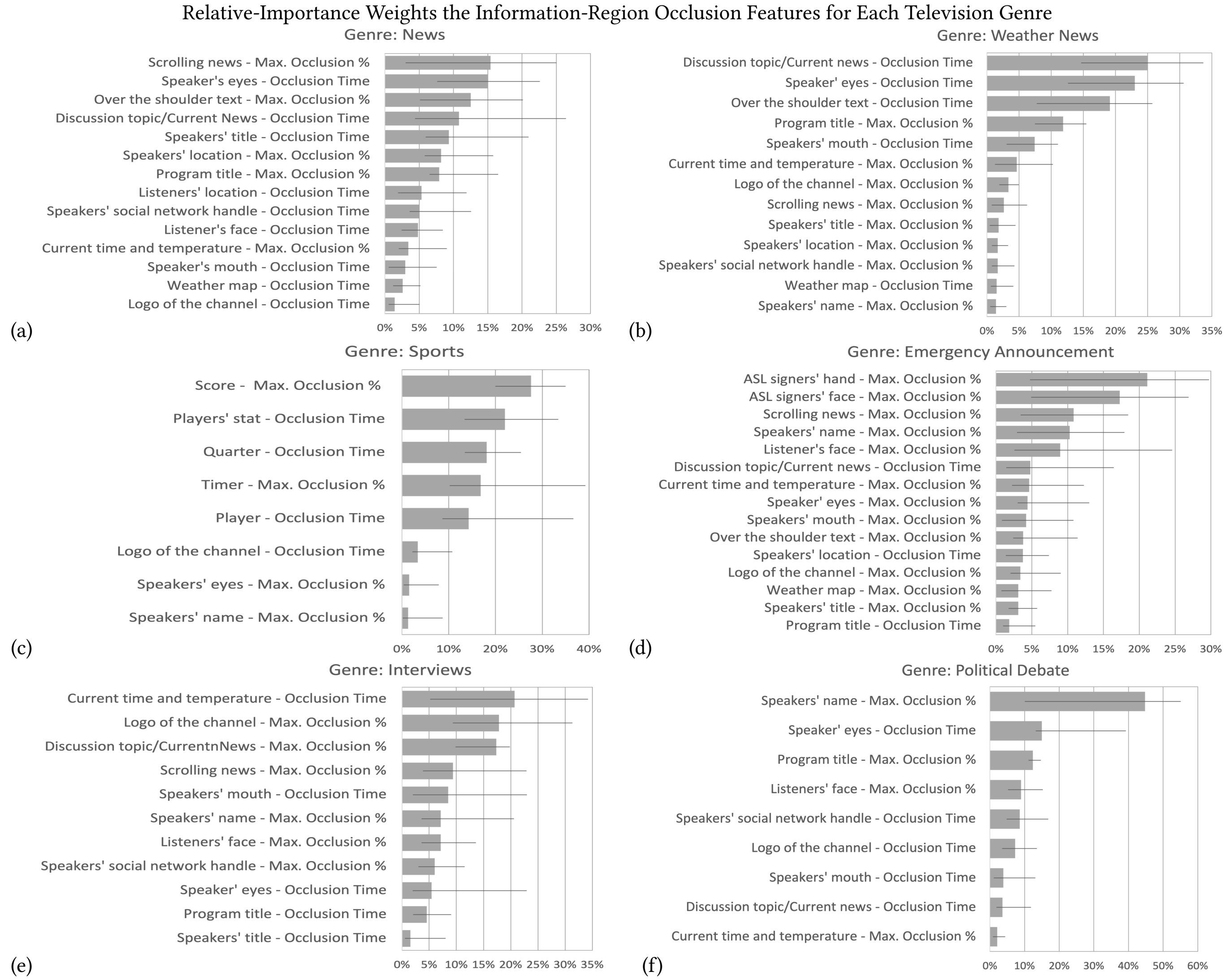


**Figure 3: Relative importance, based on percentage of total adjusted $R^2$, for each information-region-occlusion feature, for each genre: (a) News (b) Weather News (c) Sports (d) Emergency Announcement (e) Interviews (f) Political Debate**

the current score, the current game timer, and the view of the player during the game. **Two of the top-3 features differed.**

- **Emergency Announcement:** Figure 3(d) illustrates how the top-3 features: the ASL signer's hand (many emergency announcement videos include an ASL interpreter), the ASL signers' face, and scrolling news (text displaying other related information or headlines) explain 21%, 17%, and 10% of total variance, respectively. In contrast, the top-3 contributing features in the prior model [3] were: current discussion topic, over-the-shoulder text, and the speakers' job title. **All of the top-3 features differed.**
- **Interviews:** Figure 3(e) displays the top-3 features and their contribution to total variance: current time and temperature (20%), logo of the channel (17%), and discussion topic (17%). On the other hand, the top-3 features in the prior model [3] were: the current discussion topic, the speaker's mouth, and the speaker's eyes. **Two of the top-3 features differed.**
- **Political Debate:** Finally, Figure 3(f) shows how more than 60% of total variance was explained by the top-3 features: the speaker's name (44%), the speaker's eyes (14%), and the title of the program (12%). In the prior model [3], the top-3 variables had been: the current discussion topic, the speaker's mouth, and the speaker's name. **Two of the top-3 features differed.**

# 6 COMPARISON OF THE PERFORMANCE OF THE NEW AND PRIOR MODEL

While the findings in section 4.2 revealed that a significant amount of variance was explained by the holistic-judgement regression model, and the findings in 5.2 revealed that the weighting of factors in the new models differed from that of the prior component-judgement model [3], our analysis thus far has not revealed whether the new model is actually *better* at predicting DHH viewers' judgement of the caption-placement quality in videos. To investigate research question RQ3, we need to evaluate how well a metric based on the new holistic model would compare to the metric based on the prior component model.

## 6.1 Methods

Both the new holistic-judgement model and prior model [3] had a goal of predicting DHH viewers' subjective judgement of the quality of caption placement during a video. As a basis for this evaluation, we made use of an existing dataset of 33 video stimuli, with a mix of videos across six television genres, which had been made publicly available by researchers in a prior study [3]. The dataset also included 1-to-10 scalar subjective judgements of the caption-placement quality of each video, which had been collected from 23 DHH participants [3].

To apply our new metric and the prior metric to this dataset, we began by annotating the occlusion percentage and occlusion time, using the same methodology described in section 4.1.2, to identify the degree to which captions occluded various information regions in each video. Next, for each video, we calculated the predicted caption-placement quality score using both our new metric and the prior metric [3]. Assuming the DHH participants' subjective judgment as ground truth, we calculated the following two correlations: (a) prediction of the new model as compared to the ground truth, and (b) prediction of the prior model as compared to the ground truth.

## 6.2 Findings for RQ3: Comparison of Performance of the New and Prior Model

| Genre | $\rho$ **for Component Judgment Model** | $\rho$ **for Holistic Judgment Model** |
|---|---|---|
| News | 0.717 | 0.741 |
| Weather News | 0.299 | 0.590 |
| Sports | 0.226 | 0.439 |
| Emergency Announcement | 0.233 | 0.293 |
| Interviews | 0.619 | 0.608 |
| Political Debate | 0.553 | 0.518 |

**Table 1: A comparative illustration of Pearson Correlation Coefficients ($\rho$), across 6 genres, between DHH viewers' judgements of video quality, as compared to: the Component Model and the Holistic Judgment Model.**

Table 1 illustrates the Pearson correlation coefficients, for each genre, between participants' subjective judgments of video quality and the value predicted by two metrics: 1) the quality score predicted by the prior state-of-the-art component-judgement model [3], and 2) the new holistic-judgement models' predictions.

To address research question RQ3, a significance test was conducted to determine which model correlated better with DHH users' subjective scores. A Fisher r-to-z transformation revealed that the occlusion score generated from the new holistic model was significantly better correlated with DHH viewers' feedback than the prior model [3] for two genres: weather news ($z = 2.36, p < 0.05$) and sports ($z = 2.58, p < 0.01$).

# 7 DISCUSSION

## 7.1 Making Use of Our Findings

To conduct our comparison in Section 6, we had to implement a metric for predicting DHH viewers' judgements of caption-occlusion severity in videos. As discussed in section 1, we foresee that such metrics could be used prospectively by TV broadcasters who employ human-powered captioning services: In the case of a video with multiple information regions on the screen, with trade-offs to consider in what a caption may block, our metric could guide selection of an optimum caption placement. Our metric may also be useful as a basis for future fully-automated methods for caption-placement selection, e.g., similar to prior work in section 2.1.

Furthermore, our metric may be useful in the context of retrospectively evaluating the quality of how captions were placed during a television broadcast. Rather than asking a human judge to subjectively evaluate television programs to evaluate caption placement, the use of an automated metric may enable greater efficiency and replicability. Furthermore, by making such evaluation easier and less expensive, such evaluation of the quality of captioning in broadcast television programs may ultimately contribute to improvements in captioning quality over time.

## 7.2 Making Use of the Holistic Metric Tool

To provide a concrete illustration for future researchers as to how our holistic model could be used to evaluate the quality of captioning, we are distributing a software implementation of our metric discussed in Section 6. This software implementation makes use of the genre-specific importance weights or coefficients of each information region to build the model we introduced in section 4.2. However, if someone wants to use this tool to predict the quality of a captioning, there are two ways to prepare the input for this tool, which requires content regions in a video frame to be labeled:

- **Automatically identifying information content regions:** Several modern computer-vision libraries could be used to identify faces [29, 56] or onscreen text regions [43, 64] in a given video frame automatically.
- **Manually identifying information content regions:** Various tools enable someone to manually identify and label dynamic information regions that appear in a video, e.g., rectangle function of cv2 [8] or Qt5 [15].

Given the placement of captions and the location of other information regions on the screen, this software tool, available in the electronic supplementary materials that accompany this paper, will consider the placements of captions and content regions, calculate the caption-occlusion percentage and occlusion time for each content region in the video, and use the pre-trained models created in this paper to generate a total caption-quality score. A *Readme.md* file has also been attached that describes how to use this tool in practice.

## 7.3 Superiority of the Model based on Holistic Judgements vs. Imagination

A key contribution of this paper is the creation of a superior model for predicting caption-occlusion severity, based on a dataset of DHH viewers' judgements that had been obtained in a more ecologically valid manner than had been in prior work. The prior state-of-the-art metric for prediction of caption-occlusion severity in videos had been based on a dataset consisting of DHH participants viewing a static diagram and imagining how information might be blocked by captions if they had watched a video with a layout similar to the diagram [3]. Thus, our study has demonstrated the efficacy of building a model based on subjective judgements from DHH participants that had been collected in a different manner: We asked participants to actually view a captioned video and then to provide a single, holistic, subjective judgement of the overall quality of caption placement during the video. Through analysis of which information regions had been blocked by captions during each video, a regression modeling approach was used to determine how such occlusions may relate to the participants' holistic score.

We speculate that the dataset collected in our study was more ecologically valid, given that participants no longer had to imagine the experience of watching a video. Further, we asked participants to give a single score for each video, rather than introspecting about how bad it would be if captions blocked each component of the screen. We speculate that the many component judgements provided by participants in prior work [3] may not have been well-calibrated with each other. The result of such mis-calibration would be that, in a resulting metric based such a dataset, the weights assigned to occlusion of specific regions of the screen may not be proportionate. Essentially, rather than asking participants to give us the weights/coefficients for each occlusion feature in our model directly, as had been done in prior work [3], we used regression modeling to determine these coefficients.

Our comparison in Section 6.2 revealed that our new model outperformed the prior model [3] in predicting DHH viewer's overall judgements of caption-placement quality, specifically in the case of videos in the Weather-News and Sports genres — with neither model significantly better for the other four genres. We speculate that for these two genres, there is an especially high density of onscreen text and graphical information, e.g., detailed weather map information or numerical sports data and graphics. In addition, for these two genres, the people who appear in the video tend to move across the screen, e.g., the weather presenter walking to point to elements of the map or sports players running while participating in a game. For these complex visual environments, especially with people moving in the video, it may have been difficult for participants to imagine the experience of watching a video.

As discussed in section 2.4, prior HCI research studies had made use of a methodology similar to our study—i.e., holistic-judgement-collection combined with regression modeling—to create predictive models of users' subjective assessment of a system, including some work among BVI users [40]. However, no prior work had employed such methodology for collection of subjective judgements among DHH participants to determine quality judgements about captioning, let alone any prior work specifically on caption-occlusion severity.

## 7.4 Differences between Features Important to the New and Prior Models

As discussed in section 5.2, the information-region-occlusion features that were most important to our best-fit regression models in this study differed from those in prior work [3]. We speculate that differences in the data-collection procedure used in each study may help to explain some of these differences, as discussed below:

### 7.4.1 Influence of Dynamic Presentation of Information Regions.

A limitation of the prior component-based model [3] was that participants had judged the importance of information regions from a static diagram, rather than from watching a dynamic video. This difference in how judgements were collected may help to explain differences in the most important features in the new model, as compared to prior work:

- **Slowly Changing Information:** In the Weather-News genre, participants in the prior study had believed it would be very bad if the "current time and temperature" had been blocked by a caption. For the News genre, participants in the prior study had believed that it would be bad if the "current news headline" had been blocked. However, these regions were not among the top-3 in our new models. We speculate that these regions of the screen contain information that changes relatively slowly. If a participant were asked to imagine a caption blocking these regions, they may imagine the caption blocking that region of the screen during the entire video. However, during an actual video, captions appear and disappear, with a short interval in between [11], which may be of sufficient duration for the viewer to briefly see the content that had been blocked.
- **Rapidly Changing Information:** Similarly, some information content on the screen changes rapidly, such as the hands and face of an ASL interpreter who appears in an Emergency-Announcement video or the continuously moving scrolling news at the bottom of the News video. While participants in the prior study had not included the "ASL signer's face", "ASL signer's hands" and "scrolling news" among the top-3 most important information regions, our regression analysis of the new holistic-judgement data revealed that these were among the top-3 most important features in the model. We speculate that when asked to actually watch a video, participants appreciated how detrimental it would be if these regions of the screen were blocked, as compared to participants who had simply imagined watching a video.

While the list above suggests a binary classification of content regions as to whether they contain slow- or fast-changing information, this could instead be conceived of as a scalar property of any particular content region. Rather than a slow vs. fast boolean distinction, one could envision characterizing each information region within a video as to the speed of information change. For example, a headline scrolling at a faster or slower speed should be measured on a different scale.

*7.4.2 Influence of Multiple Onscreen Speakers.* Among several of the genres in the prior component-model study [3], participants had indicated that it would be detrimental if a caption were to block a speaker's mouth; however, occlusion of the speaker's mouth was never among the top-4 most important features in our new models. In prior research, DHH individuals had explained their desire that captions not block a speaker's mouth as due to it interfering with speechreading [2], which is also known as lipreading. However, in a heated political debate or an interview among multiple people, it is common for people to speak in a rapid and overlapping manner, which makes speechreading more difficult. We speculate that when actually watching videos with multiple speakers onscreen in rapid conversation, participants relied more on captions than they imagined, rather than relying on speechreading.

*7.4.3 Influence of Speaker Orientation and Camera Angle .* In the prior component-model study [3], participants had also imagined that the speakers' mouth would be important during Weather-News videos, but in our new holistic-judgement model, we observed otherwise. We speculate that the way in which weather presenters often turn their head to look at the weather map when speaking [20] may contribute to this difference: When actually watching a Weather-News video, the presenter may have turned away from the camera more than participants imagined.

## 8 LIMITATIONS AND FUTURE WORK

There were several limitations in our study, that may suggest future avenue of research:

- While evaluating the model with a set of captioned video stimuli in section 6.2, the total number of participants recruited for evaluating those stimuli had been relatively small, and the DHH individuals recruited were relatively young adults, which only represents a subset of the DHH community. There is a need for future research to obtain a large dataset of user feedback from a greater number of DHH individuals.
- Of course, television programs that people watch might generally be much longer than 30 seconds. Therefore to truly understand if these models are generalized to longer TV programs, a future study is needed. The reason why we selected 30-second videos in this study was because our work was intended to compare with the prior research study that had created the earlier component judgment model [3], and researchers, in that earlier work, had also built and tested a model based on judgements focusing on 30-second video samples.
- In preparing the video stimuli dataset for evaluation by DHH viewers, we had selected video stimuli from 6 television genres, again to enable comparison with prior work [3]. However, a future study that investigates videos from additional genres may reveal other information regions that are a concern for caption occlusion. In future work, researchers who investigate an even wider range of genres may be able to create an even more generalizable model of caption occlusion.
- In this research study on caption quality, we have specifically focused on the issue of caption occlusion, but there are other visual properties of captions that may affect viewers' perception of quality [13, 14]. Future research may investigate these other properties, e.g., synchronicity of the captions with the spoken audio signal, methods of indicating who is speaking, etc.
- The caption-occlusion severity metric based on our new model, evaluated in Section 6, required as input both the location of captions in the video and the location of information regions on the screen. In this paper, we have identified the location of information regions on the video stimuli manually, i.e., with a researcher identifying the location on screen of headlines or faces. In future work, that manual step could be automated through use of modern computer-vision tools for detecting text or faces in video images. Our findings thereby motivate future advances in computer-vision algorithms for detecting specific information regions of the screen during television video.
- Since we conducted a remote study due to COVID-19, we did not have the complete control over the devices or monitors used by our participants. While participants were required to use a computer or laptop (rather than a smartphone or tablet), were required to maximize the window displaying our stimuli, and were shown videos at a fixed aspect ratio with captioned burned into the video image (to control what regions of the screen were blocked by the caption), we were not able to fully control the viewing conditions of our remote participants, who may have used monitors of different sizes. A future study could be conducted an in-person modality, e.g., in a laboratory setting, to enable more complete control of the monitor or screen size.
- In this study, participants viewed videos on a computer, but it has become increasingly popular to consume media on a variety of devices, including smartphones, tablets, or other smaller screens. Future research would be needed, potentially using data collection methodologies analogous to those in this study, in order to understand how viewing television content on smaller screens or other form factors, e.g., large TVs in a home setting, may affect the relative importance of various information content regions.

## 9 CONCLUSION

The key contribution of this research is the creation of a superior metric for assessing the quality of caption-placement in videos based on whether the caption occludes other information in the video. This advancement beyond a prior state-of-the-art model [3] was enabled through the use of a more ecologically valid data collection approach: Rather than asking DHH participants to give imagination-based estimations of the importance of regions of the screen when looking at static images, we asked them to provide a subjective judgement of the caption-placement quality for videos they watched. Based on this data, regression modeling was used to understand the relationship between caption occlusion and participants' overall quality scores.

Beyond this main contribution, a relative-importance analysis of the features within our regression models revealed how the importance-ranking of occlusion feature weights in our new model differed from prior work [3]. This finding further supported the main premise of this study, i.e., the way in which occlusion affects DHH viewers' experience of watching videos differs from how viewers imagine that it would.

This improved metric has benefits for both prospective placement of captions in videos (to minimize the negative impact of occlusion) and for retrospective evaluation of broadcast television programs (to assess the quality of caption placement). To allow HCI researchers to replicate our work, and to provide an example for captioning practitioners of how to make use of our new model, we disseminate a software implementation of our metric in our electronic supplement.

Our work contributes more broadly to the HCI research literature, as further demonstration of the effectiveness of collection of holistic subjective judgments and regression modelling, for creating metrics to predict users' subjective ratings—rather than asking participants to provide individual judgments of the importance of component factors.

## ACKNOWLEDGMENTS

This material is based on work supported 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.* Association for Computing Machinery, New York NY USA. 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] BBC. 2018. *BBC Subtitle Guidelines, 2018.* British Broadcasting Corporation, Portland Place, London, United Kingdom. https://bbc.github.io/subtitle-guidelines

[6] 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

[7] 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

[8] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[9] 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.

[10] 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

[11] Wim De Bruycker and Géry d'Ydewalle. 2003. Chapter 31 - Reading Native and Foreign Language Television Subtitles in Children and Adults. In *The Mind's Eye*, J. Hyönä, R. Radach, and H. Deubel (Eds.). North-Holland, Amsterdam, 671–684. https://doi.org/10.1016/B978-044451020-4/50036-0

[12] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. *Investigating the Impact of Gender on Rank in Resume Search Engines.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi-org.ezproxy.rit.edu/10.1145/3173574.3174225

[13] 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

[14] 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

[15] The Qt Company. 2020. *Qt 5 tool (Version 5.15) [Software].* https://doc.qt.io/qt-5.15/index.html

[16] 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* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 215–222. https://doi.org/10.1145/2700648.2809866

[17] S. Cushion. 2015. *News and Politics: The Rise of Live and Interpretive Journalism.* Taylor & Francis, 5 Howick Place, London, SW1P 1WG. https://books.google.com/books?id=d8kqBwAAQBAJ

[18] Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 1 (2013), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0587.2012.07348.x

[19] A. Edwards and I. Curator of Charleston Museum J Long. 1995. *Extraordinary Human-Computer Interaction: Interfaces for Users with Disabilities.* Cambridge University Press, 1 Liberty Plaza New York, NY, USA 10006. https://books.google.com/books?id=3KBOAAAAIAAJ

[20] David Fairbairn and Milad Niroumand Jadidi. 2013. Influential Visual Design Parameters on TV Weather Maps. *The Cartographic Journal* 50, 4 (2013), 311–323. https://doi.org/10.1179/1743277413Y.0000000040 arXiv:https://doi.org/10.1179/1743277413Y.0000000040

[21] Olivia Gerber-Morón, Agnieszka Szarkowska, and Bencie Woll. 2018. The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research* 11, 4 (Jun. 2018), 18 pages. https://doi.org/10.16910/11.4.2

[22] Ulrike Groemping. 2006. Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software, Articles* 17, 1 (2006), 1–27. https://doi.org/10.18637/jss.v017.i01

[23] Stephen R. Gulliver and Gheorghita Ghinea. 2003a. How level and type of deafness affect user perception of multimedia video clips. *Inform. Soc. J. 2* 2, 4 (2003a), 374–386.

[24] Stephen R. Gulliver and Gheorghita Ghinea. 2003b. *Impact of captions on hearing impaired and hearing perception of multimedia video clipsb.* In Proceedings of the IEEE International Conference on Multimedia and Expo., USA.

[25] 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

[26] 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.

[27] Neter J., Kutner M. H., Nachtsheim, C. J., and W. Wasserman. 1996. *Applied Linear Statistical Models. (1996).* Chicago: Irwin, Chicago, USA.

[28] David G. Jenkins and Pedro F. Quintana-Ascencio. 2020. A solution to minimum sample size for regressions. *PLOS ONE* 15, 2 (02 2020), 1–15. https://doi.org/10.

1371/journal.pone.0229345

[29] Yingyu Ji, Shigang Wang, Yang Lu, Jian Wei, and Yan Zhao. 2018. Eye and mouth state detection algorithm based on contour feature extraction. *Journal of Electronic Imaging* 27, 5 (2018), 1 – 8. https://doi.org/10.1117/1.JEI.27.5.051205

[30] 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

[31] Hernisa Kacorri, Eshed Ohn-Bar, Kris M. Kitani, and Chieko Asakawa. 2018. *Environmental Factors in Indoor Navigation Based on Real-World Trajectories of Blind Users.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi-org.ezproxy.rit.edu/10.1145/3173574.3173630

[32] 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

[33] Joos Korstanje. 2021. *The Linear Regression.* Apress, Berkeley, CA, 149–157. https://doi.org/10.1007/978-1-4842-7150-6_11

[34] 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

[35] 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

[36] Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2014. Accessibility Evaluation of Classroom Captions. *ACM Trans. Access. Comput.* 5, 3, Article 7 (Jan. 2014), 24 pages. https://doi.org/10.1145/2543578

[37] English language Working Group on Closed Captioning Standards. 2018. *English-language Working Group. 2008. Closed Captioning Standards and Protocol for Canadian English Language Television Programming Services. Retrieved from:.* Canadian Association of Broadcasters. https://www.cab-acr.ca/english/social/captioning/captioning.pdf

[38] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2010. *Research Methods in Human-Computer Interaction.* Wiley Publishing, Hoboken, NJ, USA.

[39] R.H. Lindeman, P.F. Merenda, and R.Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis.* Scott, Foresman, Northbrook, IL. https://books.google.com/books?id=-hfvAAAAMAAJ

[40] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. *What Makes Videos Accessible to Blind and Visually Impaired People?* Association for Computing Machinery, New York, NY, USA. https://doi-org.ezproxy.rit.edu/10.1145/3411764.3445233

[41] Obach M, Lehr M, and Arruti A. 2007. Automatic speech recognition for live TV subtitling for hearing-impaired people. *Challenges for Assistive Technology: AAATE 07* 20 (2007), 286.

[42] Fabio Masina, Valeria Orso, Patrik Pluchino, Giulia Dainese, Stefania Volpato, Cristian Nelini, Daniela Mapelli, Anna Spagnolli, and Luciano Gamberini. 2020. Investigating the Accessibility of Voice Assistants With Impaired Users: Mixed Methods Study. *J Med Internet Res* 22, 9 (25 Sep 2020), e18431. https://doi.org/10.2196/18431

[43] Ali Mirza, Ossama Zeshan, Muhammad Atif, and Imran Siddiqi. 2020. Detection and recognition of cursive text from video frames. *EURASIP Journal on Image and Video Processing* 2020, 1 (Aug. 2020), 34.

[44] Ofcom. 2015. *Measuring live subtitling quality, UK.* The Office of Communications. https://www.ofcom.org.uk/__data/assets/pdf_file/0019/45136/sampling-report.pdf

[45] 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

[46] Anni Rander and Peter Olaf Looms. 2010. The Accessibility of Television News with Live Subtitling on Digital Television. In *Proceedings of the 8th European Conference on Interactive TV and Video* (Tampere, Finland) *(EuroITV '10)*. Association for Computing Machinery, New York, NY, USA, 155–160. https://doi.org/10.1145/1809777.1809809

[47] Rui Rodrigues, Ana Veloso, and Oscar Mealha. 2016. Influence of the graphical layout of television news on the viewers: An eye tracking study. *Observatorio* 10 (03 2016), 67–82.

[48] Rui Rodrigues, Ana Veloso, and Óscar Mealha. 2012. A Television News Graphical Layout Analysis Method Using Eye Tracking. In *2012 16th International Conference on Information Visualisation*. IEEE Computer Society, USA, 357–362. https://doi.org/10.1109/IV.2012.66

[49] 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

[50] J. Sandford. 2015. The impact of subtitle display rate on enjoyment under normal television viewing conditions. *IET Conference Proceedings* 1 (2015), 8 .–8 .(1). https://digital-library.theiet.org/content/conferences/10.1049/ibc.2015.0018

[51] Martin Schmettow. 2015. Tutorial: Modern Regression Techniques for HCI Researchers. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 651–654.

[52] Mark A. Schmuckler. 2001. What Is Ecological Validity? A Dimensional Analysis. *Infancy* 2, 4 (2001), 419–436. https://doi.org/10.1207/S15327078IN0204_02

[53] Elizabeth C. Smith, Mary Nell McNeese, Lin Harper, and Sherry Finneran. 2006. Factors Which Predict Compliance with Accessibility Guidelines for Disabled Users By Higher Education Institutions. In *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006*, Thomas Reeves and Shirley Yamashita (Eds.). Association for the Advancement of Computing in Education (AACE), Honolulu, Hawaii, USA, 922–929. https://www.learntechlib.org/p/23819

[54] Society of Cable Telecommunications Engineers. SCTE. 2012. *Standard For Carriage Of VBI Data In Cable Digital Transport Streams.* Technical Report. CableLabs.

[55] 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

[56] S. V. Tathe, A. S. Narote, and S. P. Narote. 2016. Human face detection and recognition in videos. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2200–2205. https://doi.org/10.1109/ICACCI.2016.7732378

[57] The Nielsen Company (US), LLC. 2020. *The Nielson Total Audience Report: April 2020.* Technical Report. Nielson.

[58] Toinon Vigier, Yoann Baveye, Josselin Rousseau, and Patrick Le Callet. 2016. Visual attention as a dimension of QoE: Subtitles in UHD videos. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Piscataway, NJ, 1–6. https://doi.org/10.1109/QoMEX.2016.7498924

[59] Caroline Wagenbreth, Julia Rieger, Hans-Jochen Heinze, and Tino Zaehle. 2014. Seeing emotions in the eyes – inverse priming effects induced by eyes expressing mental states. *Frontiers in Psychology* 5 (2014), 1039. https://doi.org/10.3389/fpsyg.2014.01039

[60] 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* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 331–332. https://doi.org/10.1145/2982142.2982205

[61] 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* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 331–332. https://doi.org/10.1145/2982142.2982205

[62] 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.

[63] WGBH. 2019. *Closed Captioning on TV in the United States 101.* Media Access Group (WGBH). https://blog.snapstream.com/closed-captioning-on-tv-in-the-united-states-101

[64] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. 2017. EAST: An Efficient and Accurate Scene Text Detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 2642–2651. https://doi.org/10.1109/CVPR.2017.283

# A APPENDIX A

| | News | Weather News | Sports | Emergency Announcement | Interviews | Political Debate |
|---|---|---|---|---|---|---|
| Constant (Y-intercept) | $6.55^{***}$ | $7.15^{***}$ | $6.79^{***}$ | $6.98^{***}$ | $6.90^{***}$ | $6.68^{***}$ |
| Speaker' Eyes | $-1.56^{**}$ | $-1.31^{***}$ | -0.78 | 0.39 | -0.90 | $-1.92^{**}$ |
| Speakers' Name | -0.67 | -0.97 | 0.29 | $-2.34^{**}$ | -0.51 | $-2.75^{**}$ |
| Speakers' Title | $-1.74^{*}$ | 1.10 | | -0.72 | -0.66 | |
| Discussion Topic/Current News | $-1.14^{**}$ | $-1.53^{***}$ | | -0.86 | 0.16 | 0.57 |
| Logo of the channel | -0.56 | 0.41 | -0.91 | $1.79^{*}$ | -1.64 | -1.15 |
| Over the shoulder text | $-2.89^{***}$ | $-0.97^{*}$ | | 0.51 | | |
| Scrolling News | $-1.98^{***}$ | $-1.20^{*}$ | | $-4.74^{***}$ | -0.49 | |
| Current Time and Temperature | $1.06^{*}$ | $-2.24^{**}$ | | 3.20 | -2.50 | 0.92 |
| Speakers' Social Network Handle | $2.88^{*}$ | 0.38 | | | -2.03 | -2.08 |
| Speakers' Mouth | -0.33 | -1.88 | -5.79 | 1.4028 | $-3.94^{**}$ | |
| Weather Map | -3.38 | 0.03 | | $16.73^{*}$ | | |
| Listeners' Face | $-1.17^{**}$ | | | $-2.16^{***}$ | -1.67 | -1.17 |
| Speakers' Location | $-1.63^{*}$ | 1.22 | | 0.49 | | -0.7497 |
| Listeners' Location | -2.31 | | | -1.01 | | |
| Program Title | -0.51 | $-1.21^{**}$ | | | 0.70 | 0.47 |
| ASL signers' face | | | | -2.16 | | |
| ASL signers' hand | | | | $-3.28^{*}$ | | |
| Score | | | $-1.61^{*}$ | | | |
| Player | | | $-0.81^{*}$ | | | |
| Timer | | | -0.73 | | | |
| Quarter | | | 0.51 | | | |
| Players' Stat | | | $-1.31^{***}$ | | | |

**Table 2: Coefficients from regression models in Section 4 indicating how occlusion-measurement of various information regions contribute to DHH viewers' perceived captioned video quality. Significance codes:** $ifp < 0.001" * * * "$, $ifp < 0.01" * *"$, $ifp < 0.05" * "$**.**