Caption-Occlusion Severity Judgments across Live-Television Genres from Deaf and Hard-of-Hearing Viewers

Akhter Al Amin*
Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY, USA
aa7510@rit.com

Saad Hassan*
Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY, USA
sh2513@rit.edu

Matt Huenerfauth
School of Information
Rochester Institute of Technology
Rochester, NY, USA
matt.huenerfauth@rit.edu

ABSTRACT

Prior work has revealed that Deaf and Hard of Hearing (DHH) viewers are concerned about captions occluding other onscreen content, e.g. text or faces, especially for live television programming, for which captions are generally not manually placed. To support evaluation or placement of captions for several genres of live television, empirical evidence is needed on how DHH viewers prioritize onscreen information, and whether this varies by genre. Nineteen DHH participants rated the importance of various onscreen content regions across 6 genres: News, Interviews, Emergency Announcements, Political Debates, Weather News, and Sports. Importance of content regions varied significantly across several genres, motivating genre-specific caption placement. We also demonstrate how the dataset informs creation of importance-weights for a metric to predict the severity of captions occluding onscreen content. This metric correlated significantly better to 23 DHH participants' judgements of caption quality, compared to a metric with uniform importance-weights of content regions.

CCS CONCEPTS

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

KEYWORDS

Dataset, Accessibility, Caption, Metric, Genre

ACM Reference Format:
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 (W4A '21), April 19–20, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3430263.3452429

1 INTRODUCTION

Over 360 million people worldwide [5] who are Deaf and Hard of Hearing (DHH) may benefit from captions while watching live television programs, i.e. video that is not pre-recorded and is generally transmitted by local television stations, which may be affiliates of a national television network. Such programming consists of various **genres**, e.g. news about the local community, emergency announcements, weather information, live sports, etc. Especially during a public health emergency like COVID-19, local stations transmit live emergency announcements through broadcast TV or streaming video on the web. For example, this may include the current prevalence of the virus in the local communities and the safety and health precautions to be taken by those communities. To effectively disseminate such information among viewers, broadcasters use various layouts and graphical elements that include both textual and non-textual content, i.e. graphs showing health trends or the name and title of a health official who is speaking (Figure 1).

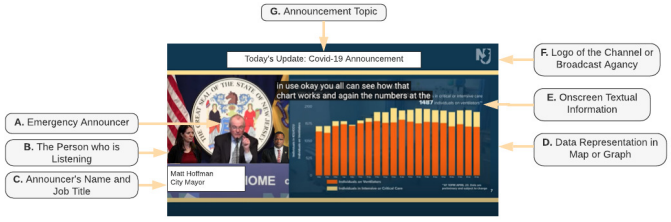

Figure 1: Example of various onscreen information content regions in an emergency announcement video.

Across genres of live video broadcasts, e.g. news or sports, there are diverse layouts and onscreen content, [14, 32, 42], e.g. dynamic scoreboards, players' statistics during sports, daily temperature-forecast graphics during weather news. While a human professional may place captions carefully during prerecorded television programming to avoid occluding important onscreen content, placement may not be adjusted manually during live broadcasts, with captions often left at static location, e.g. the lower third of the screen [38] or other locations [3, 7]. Recent research has examined algorithms for automatically selecting caption placement, e.g. by considering the speaker's location onscreen or the viewer's gaze [19, 22, 27]. Regardless, if onscreen information appears where captions are often placed, then there is a risk that captions may block important onscreen content regions of interest to DHH viewers, leading to dissatisfaction with captioning services [16, 17, 33].

To ensure quality in the captioning provided by local television broadcasters, periodic evaluation is necessary, e.g. by regulators or advocacy groups [7, 11]. Prior research has established that avoiding captions blocking other onscreen content is important for DHH

*Both authors contributed equally to this research.

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

ACM ISBN 978-1-4503-8212-0/21/04…$15.00
https://doi.org/10.1145/3430263.3452429

viewers' overall perception of the quality of a captioned live television program [16, 17]. Existing guidelines for caption placement encourage broadcasters to avoid occluding "salient" graphical elements which may be important to DHH users [7, 11], but the specifics of interpreting what are the most salient elements are left to subjective interpretation, nor do guidelines provide **genre-specific** guidance about which onscreen information is the most important not to be occluded. This is a concern since some onscreen content may differ in importance depending on the genre. For instance, the job title of a speaker during an emergency public-health announcement may be especially important for viewers, who may wish to know their authority or credentials; a reporter's job title in a news program may be less important for viewers. Therefore without empirical research from DHH users, users of such guidelines are left to subjectively prioritize which regions of the screen are most important to be visible, across various genres.

A limitation of human-powered evaluation of local television captioning quality is that it is resource-intensive, and therefore it is only possible to perform occasional spot-checks of sample video from broadcast regions. Evaluation could also be performed by a fully- or semi-automated metrics, to enable more frequent and consistent evaluation. Prior efforts to develop such metrics have largely focused on the text transcription accuracy of the captions, e.g. [1], [37], [41], [23]. These automatic metrics generally do not consider the degree to which captions block other visual information onscreen, nor whether the severity of occluding specific onscreen content may depend upon the genre of the program itself. As a result, quality scores may not reflect DHH users' judgments, since the severity of content occlusion can not be measured in existing metrics nor do current guidelines provide genre-specific criteria.

There is a lack of empirical data, gathered from DHH viewers, about their preferences and judgements about which visual elements should not be blocked by captions during live television, and there has been no empirical research as to whether these judgments may vary depending upon the television genre. Such data is necessary to support two goals: (1) creating better metrics for automatically evaluating caption placement quality to support the work of regulators, and (2) supporting efforts to place captions better in the future (either through guidelines for human professionals or the development of automatic placement algorithms).

We therefore conducted a data-collection study with **19** DHH participants, who viewed videos from six local television genres: news, weather news, sports, interviews or talk shows, emergency announcements, and political debates. For each genre, we included videos with various typical screen layouts we had identified [10], e.g. a news anchorperson looking at the camera, a new reporter speaking from a remote location, etc. We collected 3,002 judgements from participants, on an ordinal scale, about how important it would be that captions not block various regions of the screen. We found that the importance of onscreen content regions varied significantly, across different genres. Participants also provided some open-ended feedback about why they rated various onscreen regions as important, across various genres.

To demonstrate the use of our dataset toward goal "(1)" mentioned above, we next designed two prototype metrics for evaluating the degree to which captions occlude other visual content, weighted by the importance scores in our dataset. A follow-up study was conducted with **23** DHH participants who judged the quality of short videos, with various caption placements that blocked different regions of the screen. We found out that users' judgements of caption placement correlated better with a metric with importance-weights based on genre-specific data from our initial study, as compared to a baseline metric based on existing caption guidelines.

This study has both a **dataset** and **empirical** contribution: (a) We collect and disseminate a dataset containing 3,002 subjective ratings from DHH users' about how important it is that captions not block various types of onscreen text and graphic information, for various genres of live TV programming. This dataset can be used to inform guidelines about caption placement and the design of future caption evaluation metrics. (b) We empirically investigate whether the importance DHH users' ascribe to various onscreen regions varies, based on the genre of live TV programming. (c) We empirically determine whether a metric with importance-weights based on our dataset correlates with DHH users' preferences for caption placement. We leave the optimization of such a metric to future work; our prototype metric in this study is merely meant to demonstrate the potential value of this new dataset.

2 BACKGROUND

There are a variety of local television stations in the U.S., typically affiliates of major national television networks; these local stations transmit programming through over-the-air broadcast signals, through arrangements with cable or satellite services within their region, or through streaming apps or services. The programming may include content from the national television network, e.g. national news broadcasts, scripted entertainment, as well as regional programming that is produced by the local station and transmitted only to the local geographic market. Some of this national-network programming and much of this local programming is broadcast live or with a brief time delay, rather than being pre-recorded.

Although streaming video entertainment services are increasingly popular, live television programming provides critical information, e.g. local news. This is particularly important during a public health emergency like COVID-19 where access to timely, local information is crucial for all. As discussed earlier, live television programming poses unique challenges in providing high-quality captioning for DHH viewers. In addition to the challenges in providing an accurate transcription of the spoken content, the real-time nature of the programming makes it more difficult to select appropriate 2D placement of the captions onscreen, without blocking other important visual information content. While human-powered captioning services for pre-recorded television programs can select optimum placement and timing for captions, due to the time pressure in captioning live programming, captions are not typically placed by a human in a carefully selected location [47]. Thus, the risk of captions occluding other salient visual content is elevated.

Popular genres of live television programming include: news, weather news, political debates, interviews or talk-shows, emergency announcements, and sports [31]. Each genre can be characterized according to the type of onscreen information content, as well as the screen-layouts that are typical within each. 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 [9]. In this paper, we refer to each major category of live programming as a **genre**, and we have divided some genres into what we refer to as layouts, which correspond to these typical camera views and information layouts appearing within a genre, e.g. Figure 1. The use of some standard graphics software packages in the television industry [21, 42, 48] also contributes to the common appearance of some standard layouts of onscreen text and graphics. For example, these packages are often used by local news broadcasters to display text or graphics onscreen, such as continuous crawling news tickers, text representing the current headline, a logo of the local station, the name of the presenter, and other details [42].

3 RELATED WORK

This section examines prior work on collecting preferences among DHH viewers about captioning, placing captions on a video, and evaluating captioned video quality. We first discuss how while there has been prior **user-based research on caption appearance among DHH users**, little prior work has considered the issue of placement and occlusion. We then describe **current methods of placing captions on the screen**, which are not currently based on empirical DHH users' preferences. Finally, we explain how **existing metrics for measuring the caption quality** do not consider captions blocking onscreen content.

Several studies have investigated **DHH users' preferences for the appearance of captions**, e.g. font size and color [15, 44], or the usability of captions in various contexts, e.g. classrooms or one-to-one meetings [4, 28, 36]. This work has revealed that DHH viewers prefer specific font sizes or color, depending upon their distance from a streaming device or the background of the caption. Some studies have examined adding color or highlighting to captions to convey additional information, e.g. the accuracy or importance of words [4, 24]. Prior work has also examined how a long text should be best segmented into multiple lines to improve its readability for DHH viewers [45]. Captions should also convey non-verbal auditory information, e.g. music, laughing, or background noises, and some studies have investigated how to best represent these sounds inside a caption, e.g. [29]. Most relevant to this study, a prior experimental study with 105 DHH participants found that users were concerned about captions occluding graphical content in online videos [4], but this prior study did not specifically focus on television content, nor live programming genres.

Studies have also examined how the gaze patterns of a DHH viewer are affected by the presence of various onscreen content. For instance, an eye-tracking study found that DHH viewers focus their gaze on onscreen news presenters' face and other textual information for 19% of the total TV program time, even when a sign language interpreter was present on the screen [46]. While that work had not examined captioning, it does suggest that DHH viewers are spending some time looking at various onscreen information content sources, beyond the linguistic content of the video itself. Although the presence of captions during video increases DHH users' access to auditory information, captions occluding onscreen content reduces the overall amount of visual information that DHH viewers perceive [16, 17]. This prior work suggests the importance of gathering preferences from DHH viewers as to which onscreen visual content should not be blocked by captions.

While this paper focuses on understanding what regions of the screen during television program should not be blocked by captions, there has also been related work on **automated methods for selecting where to place captions on the screen**. For instance, some researchers have proposed approaches for placing captions such that they follow the movement of a speaker on the screen, e.g. [19], [20], [39], under the premise that having captions closer to the speaker would benefit DHH viewers. While user studies revealed that such approaches enhanced users' experience [6], there can be challenges when the speaker is off-screen [27]. Other researchers have investigated content-sensitive dynamic caption placement methods, which must recognize the appearance of onscreen content regions that should not be occluded in a particular video, e.g. the face of the news presenter. In an evaluation, participants reported that this approach was beneficial [22]; however, other work has identified challenges in dynamic placement of captions, since DHH viewers may need to put extra effort into moving their gaze to changing caption locations [26, 27]. Some eye-tracking studies of dynamic captions revealed that DHH viewers spent less time reading captions, than when caption are placed in a static location [35]. Recent research found benefits from gaze-adaptive caption placement [27], in which eye-trackers identify the viewer's gaze location, which is considered when captions are placed onscreen.

In many of these automatic caption-placement approaches discussed above, the software must consider where visual information is on the screen, to avoid placing captions on those locations. While a heuristic rule can be used to select where captions should be located, this technology could be enhanced by the collection of additional empirical data – specifically, by gathering the preferences from DHH viewers as to how they would prioritize the various onscreen content regions – so that these preferences could guide such automatic caption-placement technology.

In addition to providing guidance for the developers of technology for selecting where to place captions onscreen, there is another key motivation for our collection of a dataset in this paper – namely, this dataset could inform the creation of metrics that could automatically evaluate the quality of how captions have been placed on a video. Several researchers have introduced **metrics for evaluating caption quality**, to produce a numerical score, which should ideally relate to DHH viewers' preferences. Among these metrics, many tend to focus on the accuracy of the text content within the captions, rather than the placement of captions on the screen. Many of these metrics are based on the classic Word Error Rate (WER) metric, which penalizes transcripts that incorrectly add, delete, or substitute spoken words [1]. In recent years, researchers have proposed alternative metrics for assessing the quality of caption transcription, e.g. the Number of Edition, and Recognition error (NER) metric [37]; Weighted Word Error Rate (WWER) [41]; or Automatic Caption Evaluation (ACE) [23]. These metrics perform a comparative analysis between the hypothesis text (a caption text which has been shown during broadcast) and a reference text (an accurate verbatim text of what was actually

spoken by the speaker), to generate a numerical quality score. As discussed in the introduction, there is a need for ongoing and regular assessment of the quality of captioning in television broadcasts. While the use of these various existing metrics have contributed to improving the quality of caption transcription in the television broadcast industry [34], these metrics listed above do not consider the extent to which captions block salient onscreen content, even though prior research with DHH viewers suggest that occlusion negatively affects their TV watching experience [4].

Progress in the field of automatic video analysis and recognition suggests that it is reasonable to consider automatic metrics that may be sensitive to where onscreen content appears. Existing image-recognition technology [18, 49] is capable of identifying or classifying human faces, text content in video, and other non-textual onscreen information. Similarly, video-genre-detection technology is capable of identifying various genres of video [2]. Given these advancements, it is reasonable to consider that future automatic caption evaluation metrics could penalize captions that block salient onscreen content, e.g. human faces or text information. However, in order to identify how severely a caption should be penalized, there is a need for a dataset of subjective preferences from DHH users, to prioritize what regions of the screen should not be blocked. For various genres of television content, such data would enable the weights within such metrics to be set empirically.

4 OVERVIEW OF STUDIES

Our analysis of prior work has revealed that existing caption-evaluation metrics do not consider whether captions block visual content, and empirical guidance is needed on how to prioritize what should not be blocked. In this study, we identified 6 **genres** of live television based on the viewership trends: News, Weather News, Interviews, Emergency Announcement, Political Debate, Sports [40]. For each genre, we identified a set of common screen layouts, i.e. arrangements of onscreen text or graphics during typical scene arrangements in that genre. We then enumerated various onscreen **content regions**, which are present in each layout, and we also identified some regions that were present across multiple genres, e.g. the mouth of the person speaking. To assist in collection of subjective judgments from DHH participants, we created videos of each layout and additional images containing content regions on each depicted and labeled. In a data-collection study, 19 DHH participants viewed these videos and provided subjective ordinal scores for each content region, to indicate how important it is that each region not be occluded by caption. Participants also shared some opinions about their choices. Section 5 describes this study, as well as the resulting **dataset of quantitative responses, which is a key contribution** of our study.

Next, to provide guidance for future users of this dataset, we conducted an analysis to address our first empirical research question: **RQ1. Does the severity an onscreen content region being blocked by a caption vary, depending upon the genre of video?** Section 6 describes the analysis of response data, which revealed that the relative importance of onscreen content regions, e.g. text displaying the the title of the person speaking, varied depending upon the genre of live television programming, e.g. weather news vs. emergency announcements. This finding suggests that future users of our dataset should consider response data for individual genres, rather than pooling data across genres.

Finally, to demonstrate how our dataset could be used, Section 7 describes a prototype metric to assign a quality score to a captioned video, based on the degree to which captions occlude onscreen content regions. For each genre, the weight for how occlusion of a specific content region affects the overall score was determined based on the importance-scores from participants in our dataset. For comparison, we also implemented a simple baseline metric that considered all content regions to be equally important. This enabled us to investigate: **RQ2. Does a genre-specific caption evaluation metric correlate better with the DHH viewers' judgment about caption placement during live TV programming than a baseline metric?** In a follow-up study, we recruited 23 DHH participants and asked them to rate the quality of caption-placement in videos on a ten-point scale. We then performed a comparative analysis between how well users' responses correlated with each of the two metrics: the baseline metric, and the severity-weighted genre-specific metric based on our dataset. This study suggests the dataset's value for future metric implementation.

5 CAPTION-OCCLUSION SEVERITY DATASET

This section describes the collection of our dataset of DHH viewers' perception of the importance that onscreen content regions in various layouts of live television genres not be blocked by captions. A study was conducted (remotely, using video-conferencing, due to COVID-19), in which DHH participants provided ordinal responses about the importance of each content region (listed in Table 1) that may appear during videos in each genre, as well as some open-ended comments about the rationale for their preferences. This section presents our methodology in three phases: (1) investigating each television genre to identify typical layouts and content regions to design stimuli and questions for phase 2, (2) collection of responses from participants, and (3) assembly and dissemination of the dataset.

5.1 Phase 1: Identifying layouts and onscreen content regions for each genre

Our first task was to identify typical content regions of the video where text, people, or visual information resides. We began by consulting various research and guidelines on live television visual standards as follows: *The Rise of Live and Interpretive Journalism* [9], *Quick Guide to NFL TV Graphics* [14], *Making Interactive TV Easier to Use* [8], *Verbal Turn-Taking and Picture Turn-Taking in TV interviews* [43], and *Visual Design Parameters on TV Weather Maps* [13]. Next, we examined 60 TV programs which had been broadcast live in 17 national and local TV channels: CNN, FOX, MSNBC, TODAY, PBS, KCTV5 News, WWLTV, News 8 WROC, 13WHAM ABC News, WPTV News, KPRC 2, CBS Los Angeles, WKYC Channel 3, ABC10, ABC7NY, ESPN, and CBS Miami. After evaluating these programs, a total of 14 different layouts were observed, spanning 6 different genres, and each layout included various content regions. For example, in TV news, most of the textual information, e.g. scrolling news headlines, the news presenters' name and title, the reporters' name and title, tends to reside in lower segment of the TV screen. Whereas, in weather news, most of textual information, e.g., city name, time zones, and temperatures tends to be located in

Genres	News			Interviews or Talk Shows		Emergency Announcement		Political Debate	Weather news			Sports		
Layouts	Presenter Only	Discussion	Remote Reporter	In-studio	Remote	No Interpreter	Interpreter	Several Candidates	Hourly Forecast	Map View	Weekly Forecast	NFL	NBA	MLB
Logo of the channel/network	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speaker's eyes	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speaker's mouth	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Listener's face		✓		✓	✓	✓	✓	✓						
Name of presenter/host	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Social-network handle of presenter/host	✓	✓	✓	✓	✓							✓	✓	✓
Topic of discussion or current news story	✓	✓	✓	✓	✓	✓	✓	✓						
Job title of presenter/host	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Geographic location of presenter/host		✓	✓	✓	✓									
Current time and temperature	✓	✓	✓						✓	✓	✓			
Name of the remote reporter/guest		✓		✓	✓									
Job title of remote reporter/guest		✓		✓	✓									
Title of TV program				✓	✓									
Text information behind presenter/host						✓	✓			✓				
Hand gesture of an ASL interpreter							✓							
Geographic location of remote reporter/host					✓									
Weather Map			✓						✓					
City name on weather map								✓	✓					
News ticker/Crawler	✓	✓	✓											
Hand gesture of the weather news reporter									✓	✓	✓			

Table 1: List of onscreen content regions which appeared on the wire-frame diagram for each genre and layout

the lower middle or upper area of the TV screen. We summarize the list of content regions across different layouts of genres in Table 1.

5.2 Phase 2: Collecting Importance Judgments

This collection of data was conducted during a one-hour appointment with each of 19 DHH participants. A researcher started the data collection by sending an IRB-approved informed consent form to our participants through email, which participants read and reviewed, prior to a video-conference meeting between the researcher and the participant. Participants responded to a demographic questionnaire which was presented as a Google Form. Before explaining details about experiment, researchers asked participants their preferred communication mode, American Sign Language (ASL) or Spoken English. (The researcher conducting the study was fluent in both ASL and Spoken English.) As per each participant's communication preference, the researcher briefed them about the aim of the data collection. The participants were told that our goal is to understand which onscreen content they do not want to be blocked by captions when watching a live video on TV or streaming devices. The researcher then sent the participants a link to the data collection instrument which was a Google Form.

The form was partitioned into several individual sections, one for each genre, the sequence of which was counterbalanced using a Latin square. To facilitate asking participants about the importance of each onscreen content region, we selected an example video stimulus for each genre, to display to participants, to help them visualize the type of videos within each genre; we ensured that the sample video stimulus contained the variety of layouts and content regions we had identified in phase 1 for that genre. For clarity and to support our participants as they answered questions about the importance of each content region, we also provided diagrams below the example video stimulus, example shown in Figure 1, for each layout for that genre. Arrows pointed to content regions on each diagram, which were labeled with a name of each content region, e.g. "Name of presenter/host," as listed in Table 2. The name in each label corresponded to the wording used in scalar questions: For each content region, participants indicated agreement with a statement "it is important that captions not block this region of the screen," on a five-point Likert-scale from "Strongly Disagree" to "Strongly Agree." This was followed by an open-ended question asking participants why they believed particular onscreen content regions were most or least important.

Participants were recruited by posting an advertisement on social media websites. The advertisement included two key criteria: (1) identifying as Deaf or Hard of Hearing and (2) regularly using captioning when viewing videos or television. Participants received $40 cash compensation for either the hour-long study conducted using a video-conferencing. A total of 19 people participated in the study including 10 women, 7 men, and 2 people who identify as non-binary, aged 18 to 37 (median = 23.5). Fifteen participants identified as deaf, and 4 identified as hard of hearing. Seven participants reported regularly using American Sign Language at home or work. Ten reported that they began learning ASL when they were 9 years old or younger. The remaining participants reported using ASL for at least 1 year and that they regularly used it at work or school.

5.3 Phase 3: Dataset Dissemination

Table 2 summarizes the responses of the 19 DHH participants in our data-collection study. This dataset (comma-separated file) available at http://latlab.ist.rit.edu/w4a2021occlusion. The dataset consists of 3,002 Likert-item responses, converted to 1 to 5 scale, based on participants' rating for each of the relevant information content regions, across 14 layouts, of 6 television genres.

6 REGION-IMPORTANCE ACROSS GENRES

To address empirical research question 1, we next analyzed the quantitative responses in our dataset to determine whether DHH users' importance-rating for onscreen content regions varies across different genres. We began by identifying the 13 information content regions which appeared in the layouts of more than one television genre in our dataset. In the case of an information content region that appeared in multiple layouts within a single genre, the

	Section 1 (Average of participants' importance-score for each layout)														Section 2 (Score for each genre)					
	News: Presenter Only	News: Presenter and Reporter Discussion	News: Remote Reporter	Interviews or Talk Shows: In-studio	Interviews or Talk Shows: Remote	Emergency Announcement: No Interpreter	Emergency Announcement: Interpreter	Political Debate: Several Candidates	Weather: Hourly Forecast Chart	Weather: Map View Forecast	Weather: Weekly Forecast Chart	Sports: Football (NFL)	Sports: Basketball (NBA)	Sports: Baseball (MLB)	News	Interviews or Talk Shows	Emergency Announcement	Political Debate	Weather News	Sports
Logo of the channel/network	3.158	3.000	2.895	2.842	2.842	3.105	2.947	2.947	3.053	2.737	2.684	2.737	2.842	2.737	3.018	2.842	3.026	2.947	2.825	2.772
Speaker's Eye	4.316	4.158	4.421	4.421	4.368	4.263	4.211	4.368	4.000	4.000	3.737	3.789	3.789	3.474	4.298	4.395	4.237	4.368	3.912	3.684
Speaker's Mouth	4.211	4.000	4.263	4.474	4.421	4.368	4.105	4.421	3.947	3.895	3.632	3.842	3.632	3.368	4.158	4.447	4.237	4.421	3.825	3.614
Listener's Face		3.895		4.211	4.421	3.368	3.263	4.105							3.895	4.316	3.316	4.105		
Name of presenter/host	3.579	3.684	3.526	3.526	3.737	4.368	4.211	4.368	2.842	2.737	2.842	3.684	3.158	3.105	3.596	3.632	4.289	4.368	2.807	3.316
Social-network handle of presenter/host	2.789	2.737	2.632	2.579	2.737						2.842	2.421	2.421		2.719	2.658				2.561
Topic of discussion or current news story	4.474	4.526	4.579	4.526	4.368	4.684	4.368	4.579							4.526	4.447	4.526	4.579		
News ticker or crawler	3.526	3.526	3.263												3.439					
Job title of presenter/host	3.421	3.368	3.263	3.263	3.737	4.211	4.684	3.842	2.632	2.526	2.526				3.351	3.500	4.447	3.842	2.561	
Geographic location of presenter/host		3.947	4.105	3.158	3.474										4.026	3.316				
Current time and temperature	2.842	2.526	2.421						3.947	3.895	3.947				2.596				3.930	
Name of remote reporter/guest		3.632		4.158	4.053										3.632	4.105				
Job title of remote reporter/guest		3.263		3.737	3.737										3.263	3.737				
Title of the program				3.632	3.947											3.789				
Text information behind presenter/host	4.000					4.579	4.421			4.579					4.000		4.500		4.579	
Hand gesture of ASL interpreter							3.894										3.894			
Geographic location of remote reporter/guest					3.421											3.421				
Graphical information behind remote reporter			4.421												4.421					
Hand gesture of the weather news presenter									4.421	4.368	4.211								4.333	
Weather map behind the presenter									4.789										4.789	
Day-wise weekly weather chart										4.842									4.842	
Name of the city on the weather map									4.684	4.579									4.632	
Logo of the sports league												3.211	3.211	3.053						3.158
Play/Game clock												4.368	4.579							4.474
Score												4.579	4.632	4.579						4.596
Players' individual statistics												3.737	3.684	3.737						3.719
Player(pitcher/ batter/ quarterback/ thrower)													4.421	4.368						4.395
Inning info/current quarter of the game												4.368	4.579	4.421						4.456
Timeout/shot clock												4.158	4.474							4.316

Table 2: Section 1 presents the average of participants' rating of the importance of 29 content regions across 14 layouts, and Section 2 presents the composite scores for each genre, in which data from all layouts of a genre are averaged.

importance-score a participant assigned to the content region in all layouts of that genre were first averaged, to produce a composite genre-specific importance score for each content region, for each genre, for each participant. Finally, for each of these 13 content regions that appeared in more than one genre, we performed a statistical analysis to identify whether there was a significant difference in important scores across genres.

6.1 Quantitative Analysis Results for RQ1

Table 3 displays the average of all participants' importance-score responses, for each of the 13 content regions that appeared in more than one genre. In this table, significant differences are indicated with asterisks as follows: *** if p<0.001, ** if p<0.01, or * if p<0.05. For each content region, a Friedman test was first conducted to determine whether participants' importance judgments varied significantly across genres. For those 7 content regions with significant differences, post-hoc pairwise comparisons were performed using a Wilcoxon Signed Rank test, with Bonferroni corrections, as follows:

(1) Speaker's eyes (χ^2=22.2, p<0.001***) with pairwise difference for: Interviews / Sports (p=0.003).

(2) Speaker's mouth (χ^2=22.9, p<0.001***) with no significant pairwise differences revealed during post-hoc testing.

(3) Presenter/host's name (χ^2=35.8, p<0.0001***) with pairwise differences for: Emergency Announcements / Sports (p<0.001***), Emergency Announcements / Weather (p<0.001***), Political Debate / Sports (p<0.001***), Political Debate / Weather (p=0.002**).

(4) Presenter/host's job title (χ^2=24.8, p<0.0001***) with pairwise differences for: Emergency Announcements / News (p=0.003**), Emergency Announcements / Weather (p<0.001***), Interviews / Weather (p=0.005**).

(5) Presenter/host's geographic location (χ^2=8.33, p=0.003**) with pairwise difference for: Interviews / News (p=0.003**).

(6) Current time and temperature (χ^2=8.07, p=0.005**) with pairwise difference for: News / Weather (p=0.001***).

(7) Text information behind presenter, e.g. city name on weather maps, (χ^2=6.4, p=0.04*), with no post-hoc pairwise differences.

Onscreen Content Regions	News	Interviews	Emergency Announcement	Political Debate	Weather News	Sports	Results of the Friedman Test
Topic of discussion or current news story	4.526	4.447	4.526	4.578			χ^2=2.33, p=0.506
Speaker's eyes	4.298	4.394	4.236	4.368	3.912	3.684	χ^2=22.2, p<0.001***
Speaker's mouth	4.157	4.447	4.236	4.421	3.824	3.614	χ^2=22.9, p<0.001***
Listener's face	3.894	4.315	3.315	4.105			χ^2=7.39, p=0.065
Name of presenter/host	3.596	3.631	4.289	4.368	2.807	3.315	χ^2=35.8, p<0.001***
Job title of presenter/host	3.350	3.5	4.447	3.842	2.561		χ^2=24.8, p<0.0001***
Geographic location of presenter/host	4.026	3.315					χ^2=8.33, p=0.003**
Name of the remote reporter/guest	3.632	4.105					χ^2=1.60, p=0.206
Job title of remote reporter/guest	3.263	3.737					χ^2=0.818, p=0.366
Text information behind presenter/host	4		4.5		4.579		χ^2=6.40, p=0.040*
Current time and temperature	2.596				3.929		χ^2=8.07, p=0.005**
Social-network handle of presenter/host	2.719	2.657				2.561	χ^2=0.174, p=0.917
Logo of the channel/network	3.017	2.842	3.026	2.947	2.824	2.771	χ^2=3.34, p=0.648

Table 3: Participants' average importance score (1 = lowest, 5 = highest) for those content regions that appeared in more than one genre, with Friedman test to reveal whether any significant difference across genres (* p<0.05, ** p<0.01, *** p<0.001)

6.2 Summary of Open-Ended Feedback

The focus of our work is primarily quantitative in nature, and we do not claim to have performed a formal qualitative analysis. However, we had collected text data with open-ended questions on (1) why participants prioritized the visibility of some information content regions and (2) why some content regions were most or least important to them. In this section, we briefly present this qualitative data, which may provide some insight on our quantitative results. Overall, we found participants' views to be context- and topic-dependent, as well as relating to how they receive information.

Our quantitative results revealed that content regions related to the eyes or mouth of the speaker (or of anyone else onscreen listening to the speaker) were generally rated as important. Our text data indicated that seeing the faces of people onscreen supported DHH viewers' speechreading and understanding of emotional aspects of a discussion, as well as the personality of the people onscreen. P12 said: *"I think it's important not to cover the eyes or mouth/facial region in general of the presenter as lipreading and reading body language is used to help understand the information."*

While our quantitative results revealed generally low-to-moderate scores for the name or the job title of the people appearing on the screen, analysis of text data revealed higher pairwise importance for these content regions during Emergency Announcements, with the source of information being important for the credibility of the content of the TV program. Some participants explained that it was important whether a trustworthy authority was providing information during emergencies:

> *"For health emergency announcements, the announcer's credibility and quality of information being given is of a higher priority because they are needed to ensure my safety and security."* -P8

The logo of the television channel appeared in all of the genres in our study, and quantitative data indicated it was generally a low-importance region of the screen. However, it was notable that a few participants mentioned that there were occasions in which it was important to know about the logo to determine if the origin and source of information was trustworthy. P4 explained:

> *"The logo of the news company is very important because some news companies are not so reliable so knowing which news company would help me to decide if I should believe this story or need to research further."*

When discussing onscreen text that revealed the current news headline or topic of discussion, several participants mentioned how this can be valuable in understanding what is being spoken, especially when captioning is inaccurate or the viewer is having a difficulty with speechreading. P11 shared their experience:

> *"It's important to be able to read the headline as well as the main points being presented on the screen because it helps clarify a lot. I find news captions to be inaccurate a lot of the time so being able to read the screen with the bullet points is one of the most important things."*

7 EVALUATING A DATASET-BASED METRIC

As discussed in sections 1 and 2, a key motivation for the collection of our dataset is to inform the design of a caption evaluation metric that could consider the placement of captions on the screen, specifically whether the captions are blocking any content regions. Without our dataset, it was already possible for someone to implement a simplistic **baseline metric** as follows: The metric could identify whenever a caption blocks any onscreen content region, under the assumption of uniform severity of occlusion, i.e. all occlusions would be penalized equally, regardless of which content region had been blocked. Alternatively, a metric could be **severity-weighted**, i.e. the penalty of an occlusion could vary, depending upon which onscreen content region had been blocked. Furthermore, the analysis in section 6 revealed that the importance DHH viewers assign to onscreen content regions varies across genres, which suggest that if future researchers wish to use our dataset to set the weights of such a metric (to determine how much to penalize a caption for blocking a particular content region), these weights should be **genre-specific**.

To demonstrate the utility of our dataset, we conducted a further user study to compare two prototype metrics: the **baseline metric** described above and a **severity-weighted, genre-specific metric**

that considers the importance-scores of each content region for specific genres. In this study, we collected subjective judgements from DHH participants about the quality of caption placement during short videos of live television programs, and we compared how well these two metrics correlated to users' judgements.

7.1 Prototype Metric Framework

Both the severity-weighted genre-specific metric and baseline metric are based on a common framework, which considers the: duration a caption occludes a content region, duration that content region is onscreen, and degree of occlusion (percentage of the region's area blocked by the caption). The output of the framework ranges from 0 to approximately 1, with higher scores indicating higher quality in how captions are placed, with fewer content regions occluded. The framework is based upon a weighted geometric mean of a set of terms, with each term representing the potential occlusion of each content region in a particular genre:

$$score = \left(\prod_{i=1}^{n}(1 - f_i a_i + \epsilon)^{w_i}\right)^{\frac{1}{N}} \quad (1)$$

In equation 1, f_i represents the percentage of **frames** in which the caption occludes the content region n (out of the total number of frames region i is onscreen), a_i is the average percentage of the **area** of this content region which is occluded, w_i is the importance-**weight** of content region i, and ϵ is a constant value 0.01 (explained below). For each content region, f_i and a_i is multiplied to generate an *occlusion score*, which has a value of 0 when the content region is never occluded, and a value of 1 if the content region is completely blocked for the entire time it is on the screen. This value is subtracted from 1 to compute a *visibility score* for that content region. Our study used pop-up style captions (which appear and disappear as blocks, with brief duration between each); so, no region was ever 100% occluded; however, to avoid the possibility of a single 0-value term dominating the geometric mean, we have inserted ϵ (constant value 0.01) for smoothing. The importance-weight w_i is applied as an exponent, to weight this term before taking the geometric mean (by taking the Nth root of the product of the N terms).

The difference between the two metrics is in how the w_i values are determined. The severity-weighted, genre-specific metric weights the importance of each content region in the overall geometric mean score by applying exponents based on our dataset. Specifically, the value of w_i is the average of participants' responses for content region i, for this television genre. On the other hand, in the baseline metric, all w_i exponents have a value of 1, under the premise that without our dataset, there would be no empirical basis for assigning the relative importance of each content region.

As an example, consider a news video in which a caption blocks 30% of the area occupied by the *topic of discussion or current news story* content region, and the caption occludes this region for 450 of the 900 video frames when that content region appeared onscreen. Thus, in our severity-weighted genre-specific metric, the total visibility score of that region would be $(1 - (0.3 * (450/900)) + 0.01)^{4.526} = 0.505$, where 4.526 is participants' average importance-score for the *topic of discussion or current news story* content region during the News genre (see Table 2).

To clarify, we do not claim a software implementation of this metric as a contribution. Optimization and automation of such metrics is left for future work, and instead we simply implemented them in a wizard-of-oz manner, with two human annotators who identified timing and degree of caption occlusion in the videos independently. They watched each video several times, to identify all the onscreen content regions present and all occlusion events, i.e. whenever a caption blocks a region. For each occlusion event, the individual video frames were examined to calculate the total duration a content region was onscreen and the percentage of time it was occluded (f_i). The percentage of the area of the content being occluded (a_i) was based on the average judgement of the annotators. The annotators had an Intraclass Correlation Coefficient of > 0.9, which is in the "excellent" range, as discussed in [25]. Notably, this human implementation was performed once for the entire framework, and then each of the two metrics were calculated using this same set of human judgements, simply by using different w_i values, as described above.

7.2 Experimental Study and Results for RQ2

To determine which metric would correlate better with the judgements of DHH viewers, we needed to collect subjective judgements from DHH participants about the quality of caption placement in a set of videos, which needed to span the genres included in our original dataset. To begin, we considered a set of 110 television videos from 15 different TV channels. To select particular stimuli to display in our study, we identified videos which contained several of the content regions for each genre, as listed in Table 1. Ultimately, we selected a set of 11 stimuli videos, from across 6 genres, from the following sources: News (CNBC, Good Morning Britain), Interviews or Talk Shows (American Medical Association's Youtube Channel), Emergency Announcements (ABC Wisconsin), Political Debates (Spectrum News NY), Weather News (CBS Florida, WDIV TV), and Sports (YouTube channels of the NFL, NBA, and MLB).

To begin, we examined the caption file (containing the caption-text and caption-placement information) to confirm that the text transcription was completely accurate. Next, we needed to create multiple versions of each stimulus video, with variations in where the captions were placed onscreen, so that we could collect a variety of subjective judgements about caption placement quality. It is important for readers to note that the placement of captions in the video stimuli in this follow-up study was not based on the judgements from DHH participants that had been collected in our dataset; we were simply interested in producing videos with a variety of caption placements, in which captions would block various onscreen regions, so that we could collect judgements from participants in this new study. Thus, it was important that the captions in these stimuli blocked various onscreen content regions, so that we would not miss an opportunity to gather judgements from participants about how problematic such occlusions would be. Rather than select random placements for the captions on the screen, we wanted the placement to more naturally appear in typical locations for television content. Thus, we examined guidelines discussed in [35], guidelines from EIA 608 [38], and observation of placement of captions in our original stimulus candidate set of videos. Ultimately the following three caption placements were selected:

- Upper segment of the lower third of the TV screen
- Lower segment of the lower third of the TV screen
- Upper third of the TV screen

We engineered caption files and embedded them in each video stimuli using FFMPEG [12], an open source video editing tool, to create three versions of each stimulus, in which captions are located in each of these three typical locations.

Our experiment was conducted during an one-hour appointment with DHH participants. Participants read an informed consent form for this IRB-approved study and confirmed by email, prior to a video-conference meeting between the researcher and the participant. Participants responded to a demographic questionnaire which was presented as a Google Form. The researcher then briefed the participants about the aim of the study: to obtain their feedback about various caption positions. Participants were shown the videos with various placements of captions across 11 different layouts. Subsequently, they were asked if they were happy with the location of the captions on a ten-point ordinal scale (frowny-face to smiley-face). At the end of the experiment, we asked our participants if they had any comments.

Participants were recruited from social-media advertisements, which included two key criteria: (1) identifying as Deaf or Hard of Hearing and (2) regularly using captioning when viewing videos or television. In addition, we allowed individuals who participated in data-collection study, since the video stimuli set we generated and the study set-up we planned for this study are significantly different than our preliminary study. Participants received $40 cash compensation for the study. A total of 23 people participated in the study including 11 females, 11 males, and one non-binary, aged 18 to 37 (median 24). Eighteen participants identified as deaf, and 5, as hard of hearing. Sixteen participants reported regularly using American Sign Language at home or work and learning ASL when they were 6 years old or younger. The remaining participants reported using ASL for at least 2 years and that they regularly used it at work or school.

Upon eliciting 759 ratings from 23 participants for 33 video stimuli across 6 genres, we computed two Spearman Rho correlation scores for (a) participants' score vs. score generated from our genre-specific prototype metric and (b) participants' score vs. score generated from the baseline metric. The correlation coefficient for (a) was rho_a = 0.462 with a p-value < 0.0001 and for (b) was rho_b = 0.374 with p-value < 0.0001. Then we performed a Fisher r-to-z transformation on correlations between rho_a and rho_b and observed a significant difference between these two coefficients (z-score = 2.08, 2-tail p-value = 0.0375), which indicated that the severity-weighted genre-specific metric was significantly more correlated to the participants' scores.

8 DISCUSSION

A key contribution of this paper is the creation, analysis, and demonstration of the use of a dataset of the judgements of DHH viewers as to how important various onscreen content regions are during several genres of live television. As discussed in section 2, while there exist some standards and guidelines for where to place captions on a video, e.g. [7], [3], there is a need for empirical data to support these guidelines for automatic placement approaches. Those guidelines do not provide importance-weights or prioritization of content regions, let alone genre-specific guidance for caption placement. Section 3 revealed that while there has been prior user-based empirical research on various aspects of captioning with DHH participants, relatively little work had examined how captions occlude other onscreen content. No prior study had provided a prioritization of how DHH viewers would rate the importance of various onscreen content regions being visible.

While there has been prior research on automatic methods for selecting where to place captions in video, e.g. [22], [19], [27], as discussed in section 3, there was a need for empirical evidence to serve as a basis for such approaches, which had previously used heuristic methods to determine which content regions onscreen should not be blocked. Further, while prior work in section 3 had identified some key content regions on the screen, e.g. the face of the speaker, our work has provided a large enumerated set of many onscreen content regions, across a variety of common layouts of several live television genres. Future researchers considering live television genres may benefit from the analysis of content regions which provided a basis for our wire-frame diagrams for each genre.

As discussed in section 1, compared to pre-recorded television or video content, the real-time nature of live television makes it less likely that a human places captions carefully to avoid blocking other important visual content, which motivates our focus on live television genres. As part of the oversight of television captioning in various local regions, regulators or advocacy groups periodically need to assess the caption quality for samples of televised content, which motivates advancements in automatic metrics, to enable more frequent evaluation. Existing automatic metrics do not consider whether captions occlude onscreen information content, and as discussed in section 3, empirical data from DHH viewers as to how they would prioritize the visibility of various content regions onscreen could be incorporated into such metrics. From this perspective, we have also considered two key empirical questions, which provide guidance for how future designers of such metrics could use our dataset for metric creation.

RQ1 focused on whether the severity of a caption blocking an onscreen content region varied depending upon the video genre. Section 6 discussed how we analyzed the quantitative responses from participants, with a focus on the importance-scores for a subset of content regions that appeared in more than one genre. Our analysis revealed that DHH participants' importance judgement about 7 of these 13 content regions varied, depending upon the genre of television program. This finding is important for future users of our dataset, since it suggests that it would be inappropriate to simply pool together all of the responses from participants; instead, importance-scores for content regions should be considered on a per-genre basis. Section 6.2 provided an informal summary of some open-ended comments from participants about why some content regions are more or less important to be visible. Participants mentioned how the importance of particular content regions vary according to genres based on the degree to which that content may affect users' trust in the information, e.g. knowing the TV network of a news program to determine potential bias, or knowing the identity of a speaker onscreen to determine authoritativeness during an emergency announcement. Participants also mentioned

how some onscreen content regions are used to provide context for DHH viewers when there are errors in the captions provided.

To provide a concrete illustration for future researchers as to how our dataset could be used to implement a caption-placement evaluation metric, section 7 presented a prototype metric with weights based upon our dataset. This study addressed RQ2, and it revealed that a genre-specific caption evaluation metric correlates better with the DHH viewers' judgments about caption placement in live TV programming than a baseline metric, with uniform importance-scores for all content regions. Since these two metrics were based on a common framework, discussed in section 7.1, the differences in predictions of each metric are based only on the importance-scores weights in each. Thus, this analysis revealed specifically how the importance-scores in our dataset enabled an improvement to a metric, which was otherwise identical in structure.

Finally, it is important to note that the methodology used to create our dataset was based upon an assumption: that the quality of caption-placement for a video can be estimated, in part, by asking DHH participants to *explicitly* give subjective judgments about how important various content regions are, across several live television genres. As an alternative, we could have asked participants to give overall ratings of caption-placement quality for a wide variety of videos, with captions in various locations and occluding various content regions (similar to the study in section 7.2 but at a larger scale). Through collection many such judgements, which may have *implicitly* measured how important the occluded regions were, a regression analysis may have revealed coefficients for how occlusion of specific content regions contribute to DHH viewer's overall judgement of captioned video quality. The downside of such an approach is that an extremely large number of judgments would need to be collected, across a large and diverse set of videos, with occlusion of diverse content regions, in order to obtain such coefficients through a regression analysis. Thus, in order to collect importance-score judgements with a manageable number of DHH participants, we instead collected explicit judgements. Despite this simplification, section 7.2 provides evidence of the validity of our approach: We found that a metric based on our dataset was correlated to DHH participants' overall subjective judgment of caption placement quality in real videos.

9 LIMITATIONS AND FUTURE WORK

As discussed above, our study relied upon explicit judgements about the importance of content regions, but future research could further examine how such judgements about the importance of content generalize to other forms of holistic evaluation, including, e.g. comprehension questions or other task-based measures.

In preparation for our original data-collection study, we selected a list of onscreen elements after sampling videos from 17 different TV channels; however, there are over 1700 TV channels in U.S. alone [30]. An analysis of video from more TV stations could reveal additional content regions or layouts for each genre. In future work, researchers could conduct an even larger analysis of a wider range of television content, to determine if there are additional onscreen regions that are important to DHH viewers. Furthermore, the focus of the dataset collection and analysis in this paper has been on genres of live television programming; however, future work could extend this focus to consider genres of pre-recorded television content, as well as genres of online video on social media platforms.

In our data-collection study (section 5) and in our later study evaluating metrics (section 7), we recruited a set of DHH participants in our study who represented a relatively young demographic and who were primarily from our geographic region. Future work would be necessary to determine whether the preferences and findings of our study would generalize to a wider range of DHH individuals, across various demographic characteristics, e.g. age, gender, or identity (Deaf, deaf or hard of hearing). For this reason, we have disseminated a table of the demographic attributes of our participants at http://latlab.ist.rit.edu/w4a2021occlusion, so that readers can better interpret our findings.

When demonstrating the use of our dataset for producing a metric of caption-placement quality (section 7.1), it is important to note that the metric presented is merely an example of how a framework could be structured, as well as how a set of importance-weights for individual content regions could be calculated naively by simply using the raw importance scores from our dataset. Future research could investigate the effectiveness of a variety of alternative frameworks or importance-weight calculations, which may use our dataset in other manners, to generate metrics that correlate even better to DHH viewers' judgements. The framework presented in section 7.1 assumes that the contribution to the overall score for a video is based on the degree to which content regions are occluded individually, whereas it may be possible that captions which occlude specific combinations of content regions may have a more severe effect on the overall captioned video quality.

10 CONCLUSION

The key contribution of this research is the creation of a dataset of DHH users' judgments of the importance of various content regions across live television genres, to inform current caption-placement guidelines and caption-evaluation methods. Beyond this main contribution, a subsequent analysis revealed the empirical finding that users' responses differed significantly across genres, which provides guidance as to how future researchers should make use of this dataset in a genre-specific manner. After demonstrating how to create genre-specific caption-evaluation metric using our dataset, a second user study was conducted to gather DHH viewer's overall judgments of caption-placement quality for a set of videos. An analysis revealed our second major empirical finding, that DHH viewers' judgements were better correlated to a metric with occlusion severity-weights based on our dataset, as compared to an analogous metric without severity-weights for each content region. This finding demonstrated the utility of the dataset, and it also provided evidence of the validity of the data-collection methodology used in this work, which relied upon collection of explicit judgements of the importance of content regions from DHH participants. The dataset, diagrams of content region layouts, and other materials are disseminated at http://latlab.ist.rit.edu/w4a2021occlusion, for use by future researchers or for replication of this work.

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] Ba Tu Truong and C. Dorai. 2000. Automatic genre identification for content-based video categorization. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Vol. 4. 230–233 vol.4.

[3] BBC. 2019. *BBC Subtitle Guidelines, 2018.* https://bbc.github.io/subtitle-guidelines

[4] 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

[5] 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

[6] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic subtitles: The user experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video. 103–112.*

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

[8] Leon Cruickshank, Emmanuel Tsekleves, Roger Whitham, Annette Hill, and Kaoruko Kondo. 2007. Making Interactive tv Easier to Use: Interface Design for a Second Screen Approach. *The Design Journal* 10, 3 (2007), 41–53. https://doi.org/10.2752/146069207789271920 arXiv:https://doi.org/10.2752/146069207789271920

[9] S. Cushion. 2015. *News and Politics: The Rise of Live and Interpretive Journalism.* Taylor & Francis. https://books.google.com/books?id=d8kqBwAAQBAJ

[10] Stephen Cushion, Rachel Lewis, and Hugh Roger. 2015. Adopting or resisting 24-hour news logic on evening bulletins? The mediatization of UK television news 19912012. *Journalism* 16, 7 (2015), 866–883. https://doi.org/10.1177/1464884914550975 arXiv:https://doi.org/10.1177/1464884914550975

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

[12] FFMPEG Developers. 2016. *ffmpeg tool (Version be1d324) [Software].* http://ffmpeg.org/

[13] 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

[14] NFL Football Operation. Retrieved from:. [n.d.]. *QUICK GUIDE TO NFL TV GRAPHICS.* https://operations.nfl.com/football-101/quick-guide-to-nfl-tv-graphics/

[15] Olivia Gerber-Morón, Agnieszka Szarkowska, and Bencie Woll. 2018. The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research 6* 5 (2018).

[16] 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.

[17] 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.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[19] 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

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

[21] Compix Media Inc. 2019. *Compix BROADCAST GRAPHICS.* https://www.compix.tv/

[22] 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 Posters. 15–16.*

[23] 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

[24] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the Benefit of Highlighting Key Words in Captions for People Who Are Deaf or Hard of Hearing. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 43–55. https://doi.org/10.1145/3308561.3353781

[25] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155 – 163. https://doi.org/10.1016/j.jcm.2016.02.012

[26] Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the action: Eye-tracking evaluation of speaker-following subtitles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 6559–6568.*

[27] 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

[28] 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

[29] DANIEL G. LEE, DEBORAH I. FELS, and JOHN PATRICK UDO. 2007. Emotive Captioning. *Comput. Entertain.* 5, 2, Article 11 (April 2007), 15 pages. https://doi.org/10.1145/1279540.1279551

[30] W. Luplow and J. Kutzner. 2013. Emergency alerts to people on-the-go via terrestrial broadcasting: The M-EAS system. In *2013 IEEE International Conference on Technologies for Homeland Security (HST)*. 779–783.

[31] 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.

[32] Konrad Maj and Stephan Lewandowsky. 2020. Is bad news on TV tickers good news? The effects of voiceover and visual elements in video on viewers' assessment. *PLOS ONE* 15, 4 (04 2020), 1–18. https://doi.org/10.1371/journal.pone.0231313

[33] S. Nam, D. I. Fels, and M. H. Chignell. 2020. Modeling Closed Captioning Subjective Quality Assessment by Deaf and Hard of Hearing Viewers. *IEEE Transactions on Computational Social Systems* 7, 3 (2020), 621–631.

[34] Ofcom. [n.d.]. *Measuring live subtitling quality, UK.* https://www.ofcom.org.uk/__data/assets/pdf_file/0019/45136/sampling-report.pdf

[35] 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

[36] 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

[37] Pablo Romero-Fresco and Juan Martínez Pérez. 2015. *Accuracy Rate in Live Subtitling: The NER Model.* Audiovisual Translation in a Global Context. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan, London.

[38] Society of Cable Telecommunications Engineers. SCTE. 2012. *STANDARD FOR CARRIAGE OF VBI DATA IN CABLE DIGITAL TRANSPORT STREAMS.* Technical Report.

[39] 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, 150–162* 88 (2019).

[40] The Nielsen Company (US), LLC. 2020. *THE NIELSEN TOTAL AUDIENCE REPORT: APRIL 2020.* Technical Report.

[41] Marcia Brooks Tom Apone, Brad Botkin and Larry Goldberg. 2011. *Caption Accuracy Metrics Project Research into Automated Error Ranking of Real-time Captions in Live Television News Programs.*

[42] NewscastStudio. The trade publication for TV production professionals. 2020. *TV News Graphics Package.* https://www.newscaststudio.com/tv-news-graphics-package/

[43] Friedrich Ungerer (Ed.). 2000. *English Media Texts – Past and Present: Language and textual structure.* John Benjamins. https://www.jbe-platform.com/content/books/9789027298959

[44] Toinon Vigier, Yoann Baveye, Josselin Rousseau, and Patrick Le Callet. 2016. Visual attention as a dimension of QoE: Subtitles in UHD videos. In *Proceedings of the Eighth International Conference on Quality of Multimedia Experience. 1–6.*

[45] 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

[46] Jennifer Wehrmeyer. 2014. *Eye-tracking Deaf and hearing viewing of sign language interpreted news broadcasts.* Journal of Eye Movement Research.

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

[48] wTVision. 2020. Broadcast Design. https://www.wtvision.com/en/solutions/broadcast-design/

[49] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. *EAST: An Efficient and Accurate Scene Text Detector.* In the Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).