# Deaf and Hard of Hearing Users' Perspective for Genre-wise Content Informed Caption Placement

AKHTER AL AMIN[**], Rochester Institute of Technology
SAAD HASSAN[†*], Rochester Institute of Technology
MATT HUENERFAUTH, Rochester Institute of Technology

Abstract

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

Additional Key Words and Phrases: Dataset, Accessibility, Caption, Metric, Genre

**ACM Reference Format:**
Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2020. Deaf and Hard of Hearing Users' Perspective for Genre-wise Content Informed Caption Placement . In *Proceedings of CHI '21, October 29-November 1, 2021, Yokohama, Japan © 2021 Association for Computing Machinery. (CHI '21).* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Over 360 million people worldwide [12], who are Deaf and Hard of Hearing(DHH), use caption to access auditory information while watching live TV program. There are various kinds of live TV programs, e.g., emergency or breaking news about the local community, weather news, talk shows, and live sports, which are broadcast to local communities. These different types of TV programs are commonly referred to as **genres**.

The graphical interfaces used in these broadcasts, which include both textual and non-textual content, vary across different genres. However, the placement of captions is often standard across live TV shows of different genres which may result in the captions occluding with the onscreen content in some cases. There are certain classic locations for placement of captions on live television program as shown in Figure 1, e.g., the lower third or upper third parts of the screen. Since the captions will always be placed in these classic locations, the variable placements of onscreen content across different genres of live television programs poses a challenge for broadcasters while placing caption on the screen. For instance, classic caption location such as, on the lower third of the screen, may be a proper location for some programs, e.g. weather news or sports, but in other TV programs, e.g., breaking news or interviews, the caption may occlude some useful onscreen content such as scrolling news, the program title, the discussion topic, or the speakers' name or the title. As a result, DHH users may miss vital information and would dissatisfied with the captioning services [6, 19, 20, 31].

[*] Both authors contributed equally to this research.
[†] Both authors contributed equally to this research.

Manuscript submitted to ACM

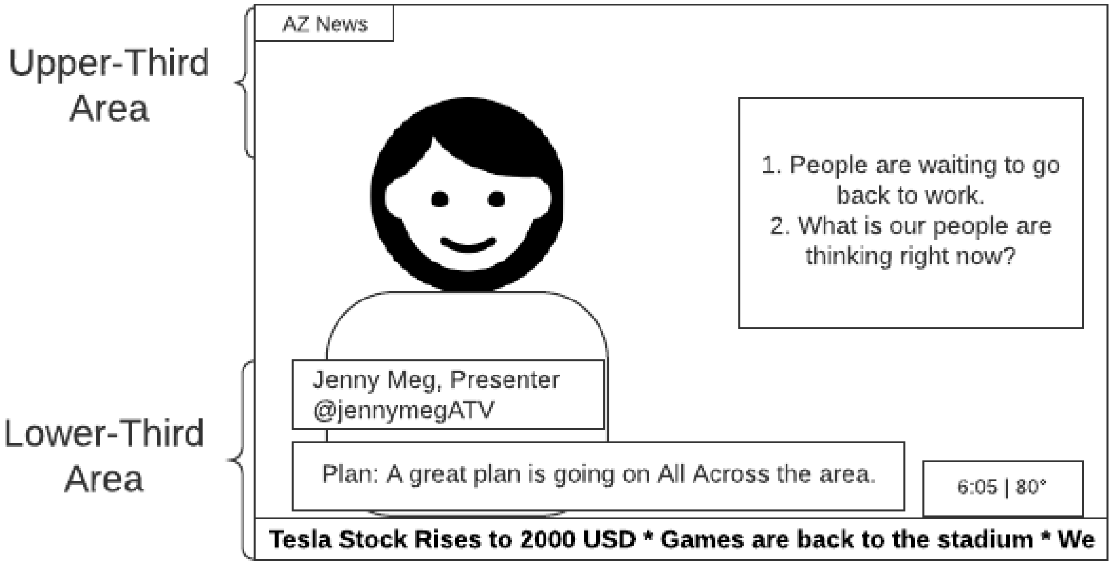
Upper-Third Area
AZ News
1. People are waiting to go back to work.
2. What is our people are thinking right now?
Jenny Meg, Presenter
@jennymegATV
Lower-Third Area
Plan: A great plan is going on All Across the area.
6:05 | 80°
Tesla Stock Rises to 2000 USD * Games are back to the stadium * We

Fig. 1. Common area of caption location on the TV screen. **Lower Third** is the lower rectangular segment and **Upper Third** is the upper rectangular segment of the screen

Like hearing viewers, DHH viewers also prefer different graphical contents on their TV screens while watching TV programs of different genres [16]. Considering DHH users' preferences towards onscreen content, several captioning regulatory bodies have mandated that captions should not occlude salient onscreen content [2, 3]. In past, the lower third area of the screen was considered a standard location for the placement of captions and was supported by regulatory authorities [8]. On the contrary, more recently regulatory authorities have provided guidelines which instead prevent the caption placement in the lower third area of the screen since a lot of onscreen content is present in that area [3]. However, standardizing a caption placement location that is suitable for all the genres may not be possible. This is because DHH users' preference for the graphical content may vary across different genres of live TV programs. For instance, while watching news the DHH viewers might find it necessary to view the news presenters' face, the reporters' face and other news related textual information. While watching sports they might prefer other onscreen content such as the game score and other game related textual information which are not placed at the same location as the onscreen contents in news genre [38].

Existing guidelines for caption placement are unaware of this variation of DHH users' content preference across different genre. Hence, although some caption placement in these TV programs may align with current guideline, there remains possibility that caption occludes such content which is important for the DHH viewers. This occlusion may, in turn, reduce DHH users' overall satisfaction with the TV program and they may lose vital information that is accessible to hearing viewers.

Caption evaluation metrics are used to understand the DHH viewers' perception of captioned live television. These include automatic caption evaluation metrics and semi-automatic caption evaluation metrics which enable regulators to monitor the quality of captions from DHH users' perspective by examining small segments of captioned live TV programs regularly. In recent years, researchers focusing on developing caption evaluation metric have broadly investigated the process of improving the transcription quality [5]. As a result, caption transcription quality has been improved significantly since the providers have minimized textual errors. However, these current metrics are agnostic towards the visual aspects of caption placement such as occlusions with different types of onscreen contents. They also fail to account for the genre of live television being analyzed. To measure the quality of captioned live TV programs more accurately, there is a need to inform the design of these caption evaluation metrics with DHH viewers visual content preferences across different genres.

To this end in this paper, we conduct an hour-long mixed method experiment involving 19 participants to identify DHH users' genre-wise preferences of captioned live TV programs. During this experiment, our participants were shown videos and wireframe diagrams representing placement of onscreen content from 6 different TV program genres: news, weather news, sports, interviews or talk shows, emergency announcements, and political debates. We divided each genre into a number of sub-genres to observe whether DHH users onscreen content preference changes when the onscreen placement of content varies due to camera location or a change of graphical layout across videos in same genre [15]. In our study, participants reported their preferences about the presentation of captions across different genres by responding to a Likert-scale question about how important is it for a particular onscreen content to not be occluded by caption. Based on the response data from our study, we then designed a prototype framework for evaluating the quality of captions by identifying the occlusion and calculating a numerical score based on the degree of occlusion and informed by the results from our user study.

Finally, in order to evaluate this proposed framework, in a follow-up study **25** with DHH participants we compared our framework against another framework that we designed which does not account for the genre of the live TV content being analyzed. In this study, participants were asked to report their subjective judgment of some video stimuli, from 6 different genres with various placements that occlude different combinations of onscreen content. Subsequently, we perform a correlation analysis between the caption quality judgment reported by these participants and the score generated by both genre-specific metric and genre-agnostic metric in order to find out if genre of the content being analyzed matters.

The contributions of this study can be broadly categorized into two types: non-empirical contributions and empirical contributions.

Our non-empirical contributions are as follows:

(1) We generate a dataset of DHH users' subjective preferences about how important it is that captions do not block various types of onscreen text and graphic information, for various genres of live TV programming. We release the dataset in a comma-separated file which will be publicly available. This dataset can be used to inform the guidelines about caption placement and the design of future caption evaluation metrics.
(2) We propose a caption evaluation metric framework that accounts for the genre of live TV show being analyzed. Research in the future can improve the framework to develop an automatic caption evaluation metric that captures DHH viewers perceptions of captioned live TV more accurately.

Our empirical contributions are as follows:

(1) We empirically investigate DHH users' preferences of how live TV content across different genres should be captioned to understand if users' desired placement resembles existing caption placement guidelines, and whether DHH users' preferences are unique to different genres.

| Genre | Sub-genre | Graphic Packages | Stimuli Video |
|---|---|---|---|
| News | News presenter is present on the screen | Vizrt, 3M Graphics, FEMA, Newscaststudio, MGN Online | CNBC, Good Morning Britain |
| | Discussion between news presenters or a reporter | | |
| | Reporter is reporting live from the place of incident | | |
| Interviews | In-studio interview | Newscaststudio, 3playmedia, compix | American Medical Association's Youtube Channel |
| | Remote interview | | |
| Emergency announcement | ASL interpreter present on screen | Newscaststudio, Sky gates studios | ABC Wisconsin |
| | ASL interpreter not present on screen | | |
| Political debate | N/A | Newscaststudio, Vizrt, Envato | Spectrum News NY |
| Weather news | Future weather news (hourly) | Newscaststudio, Metgraphics, Motionarray, Praedictix, Weathergraphics | CBS Florida, WDIV TV |
| | City-wise current temperature on the map | | |
| | Weekly weather forecast chart | | |
| Sports | National Football League | Vizrt, ESPN | NFL Youtube Channel, NBA Youtube Channel, MLB Youtube Channel |
| | National Basketball Association | | |
| | Major League Baseball | | |

Table 1. Choice of genres and sub-genres for preliminary study and source of stimuli across different genre for the study conducted to test the framework.

(2) We experimentally evaluate whether a prototype of genre-sensitive caption evaluation metric measure the quality of caption more accurately than a genre-neutral metric in regard to occlusion caused by captions during live TV programs.

## 2 BACKGROUND

TV broadcasters including local TV channels telecast live shows for a variety of reasons such as to inform the community about the emergency incidents, breaking news, or sports. There are several types of live TV programs which are broadcast in local TV channels which we refer to as genres in this paper. Some of the more popular TV genres in live television include: news, weather news, political debate, interviews, emergency announcement, sports [30]. Within each genre of TV shows, there are multiple categories of shows which are characterized by their own conventions of placement of onscreen content. For example in news genre, based on different camera angle and speakers' location the layout of screen may vary. During a news broadcast, sometime only news presenter might remain on the screen whereas on other occasions there might be a secondary news presenter or a remote reporter appearing on screen [14]. Similarly, in weather news background maps or temperature charts might appear on screen. Therefore, in this paper we have divided some genres into sub-genres representing the variety of graphical layouts that viewers might observe within TV programs of the same genre.

Each genre or sub-genre in live TV is also characterized by the types of content and their presentation on the screen. Recent advancements in television graphic packages have enabled live television broadcasters to display a variety content on screen using different stylistic presentations. In these graphics packages, information are placed in different location on the screen. For example, in news channels users are often presented with a continuous crawling news ticker, headline news, channel logo, time, current story title, and details of presenters simultaneously on the screen with the main story [7]. This diversity of content on screen and shows in live television poses unique challenges for the placement of captions. The ideal location for caption placement may vary depending on the content on screen (so that the caption does not occlude any useful content onscreen) which in turn depends on the genre of the show being broadcast.

Existing guideline instructs the broadcasters that caption should avoid salient graphical elements which might be important to DHH users[2, 3]. For example, Federal Communication Commission(FCC), a media regulatory commission in US, proposed a guideline for caption placement for digital TV. Also **BBC News** have included a guideline demonstrating the technology for moving caption vertically or horizontally across the TV screen to avoid occlusion

| Genres | News | | | Interviews/Talk Shows | | Emergency Announcement | | Political Debate | Weather news | | | Sports | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sub-genres | News Presenter Only | Discussion | Reporter | In-studio | Remote | No Interpreter | Interpreter | N/A | Hourly future news | City-wise map | Weekly chart | NFL | NBA | MLB |
| Logo of the channel | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| The Speaker's Eye | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| The Speaker's Mouth | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| The Listener's Face | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| Name of primary onscreen person | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| SN username of primary onscreen person | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | ✓ | ✓ | ✓ |
| Topic of discussion or current story text | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| Title of primary person | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| Location name of an onscreen primary person | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | |
| Time and Temperature | ✓ | ✓ | ✓ | | | | | | ✓ | ✓ | ✓ | | | |
| The name of the secondary person | | ✓ | | ✓ | ✓ | | | | | | | | | |
| The title of secondary person | | ✓ | | ✓ | ✓ | | | | | | | | | |
| Title of the TV program | | | | ✓ | ✓ | | | | | | | | | |
| Textual Information | | | | | | ✓ | ✓ | | | ✓ | | | | |
| Hand gesture of an ASL interpreter | | | | | | | ✓ | | | | | | | |
| Location name of onscreen secondary person | | | | | ✓ | | | | | | | | | |
| Changing Weather Map | | | ✓ | | | | | | ✓ | | | | | |
| Name of the city on the map | | | | | | | | ✓ | ✓ | | | | | |
| Headline/Current News | ✓ | ✓ | ✓ | | | | | | | | | | | |
| News ticker/Crawler | ✓ | ✓ | ✓ | | | | | | | | | | | |
| Hand gesture of the weather news reporter | | | | | | | | | ✓ | ✓ | ✓ | | | |

Table 2. List of onscreen elements in different genres and sub-genres

[1]. These guidelines clearly indicate that caption occluding a content, which is vital to DHH users, might affect DHH viewers' overall perception of a captioned live TV program [19, 20]. However, there is a need to better understand which onscreen content are more important for DHH users for different genres. Such information can help us understand where to place captions in screen in the presence of multiple onscreen elements. It can also inform the design of caption evaluation metrics and make sure they capture the DHH viewers perception of captioned videos more accurately by penalizing occlusions based on the importance of the content being blocked.

## 3 RELATED WORK

In this section, we will first take about user studies that have been conducted on various aspects of the appearance of captions. We will then discuss some of the current methods that are used to place captions on screen and the user studies that inform these caption placement preferences. Finally, we will discuss state-of-the-art caption evaluation metrics which are used to access the quality of captioned videos.

### 3.1 User studies on caption appearance

Over the recent years, numerous studies have investigated DHH users' preferred caption appearance style and evaluated the current caption usability in various application contexts such as classroom and impromptu one-to-one meeting [11, 28, 33]. Some of these studies investigated affect of changing font size and color [18, 36]. These investigation revealed that DHH viewers have preference for specific size of font and color depending on the viewers distance from the streaming device and background of the caption . Other researchers conducted user studies focusing on caption movement, and caption text background [11, 25] to gauge DHH users preferences about caption caption appearance and style while watching captioned videos. Research has also been conducted to investigate where should a caption text be segmented into two lines to improve its readability for DHH viewers. Another important aspect of captioning is the representation of auditory non-textual information e.g. music, non-verbal human sounds like laughing or crying, or background noises like sound of water dripping. Prior studies, e.g. [29], have investigated how to best represent these sounds inside a caption. The findings of [29] suggest that DHH viewers showed better perception and comprehension of the videos programs in which non-auditory information was present.

Studies have also looked at how the viewing patterns of DHH audience change depending on the presence of different onscreen contents. A prior study analyzing eye-tracking data had shown 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 sign language interpreter was present on the screen [37]. These findings indicate different types of onscreen content are also important for DHH viewers and they are not solely focusing on captions. While the presence of caption increases DHH users' access to auditory information, caption occluding onscreen content may reduce the overall amount of visual information that can be perceived by a DHH viewer [19, 20]. This makes placement of captions onscreen in live television a challenging task. In the following section, we elaborate on some of the prior work related to caption placement.

### 3.2 Existing Caption placement Methodologies

Several researchers have conducted user studies to investigate DHH users' preferred caption placement while watching captioned TV programming. Some researchers proposed a speaker following caption placement technique [21, 22, 35]. This technique assists DHH viewers in identifying current speakers by placing captions close to the speaker. While user studies show this caption placement method enhanced user experience [13], this method may pose other challenges when the current speaker is outside the TV screen [27] so its utility is only limited to a few settings in live television.

A recent study involving 105 participants investigated the DHH users' preferred caption appearance of captions generated using Automatic Speech Recognition(ASR) technology [11]. In this study, DHH participants' subjective responses revealed DHH users' concern about caption occluding graphical content. To place caption in a preferred location avoiding onscreen content, researchers introduced a content-sensitive dynamic caption placement technology and conducted a follow-up study with DHH participants to evaluate the usability of this dynamic caption placement [23]. This technology recognized the presence of onscreen content that should never be occluded in a particular TV show e.g., the face of the news presenter. DHH participants reported this dynamic caption positioning technique to be beneficial for them. However, since the captions are not present at the same location in dynamic captions may change location from time to time due to variations in camera angle or speaker position, DHH viewers might have to put extra effort into moving their vision focus from one part of the screen to another [26, 27]. Some studies that looked at DHH users' eye-tracking data revealed that while caption moves dynamically suggest that they spent a lesser amount of time reading caption as opposed to when the caption is placed in a static location [32]. Dynamic placement might explain the overall reduction in the time DHH viewers view captions since they also spend time moving their gaze from one location to another due to the dynamic placement of captions.

To address this concern with dynamic captioning, some recent research has proposed a gaze-adaptive caption placement technology [27]. This technology is an extended version of the dynamic caption placement method, which utilizes a head-mounted eye-tracker to place the caption following the viewers' gaze. In a user-study, DHH viewers expressed that this gaze-adaptive caption placement method increases readability of the caption. While this method resolves the earlier concern regarding caption location uncertainty on the screen, the reliance of this method on head-mounted eye-tracker raises usability concerns.

Despite innovations in captioning technologies, caption placement remains a challenging task. There is a need to continuously monitor captioned television content in order to make sure that DHH viewers do not lose any auditory or visual information that is available to hearing viewers. Caption evaluation metrics are used to assist with the task of monitoring the quality of captioned television content. In the next subsection, we will discuss the state of current caption evaluation metrics.

### 3.3 Existing Metric of Evaluating Caption Quality for DHH Users

Researcher have started to look at evaluating caption quality from DHH viewers' perspective in recent years. To this end, caption evaluation metrics have been introduced to quantitatively determine the quality of caption. A few studies have been conducted on how to improve these caption evaluation metrics so that they capture DHH viewers perception of captioned videos more accurately. Among the metrics that have been proposed in the past, Word Error Rate(WER) is a popular one and has been used by regulatory commissions to evaluate the caption transcription [10]. The metric performs a comparative analysis between hypothesis text (caption text which has been shown during broadcast) and reference text (verbatim text which is actually spoken by the speaker) to generates a caption quality score by counting number of discrepancies (number of deletion, substitution, and inclusion of new words) between these the two texts. However, a concern that caption providers raise regarding the performance of WER as caption quality measurement metric is that it evaluates both major and minor transcription errors to the same extent (e.g. deletion of article and deletion of verb has equal effect on caption quality). Therefore, WER may erroneously evaluate the quality of a high-quality caption transcript as lower and vice versa. To resolve this issue of penalizing both major and minor errors equally, Named-entity Recognition(NER) was introduced which is a semi-automatic caption evaluation employing a human annotator [34]. However, due to the dependency on a human annotator, this metric might be too costly and time consuming to be used in a real setting. Recent research reveals that the caption evaluation process can be automated by evaluating each transcription error level by a probabilistic model-based algorithm named Weighted Word Error Rate(WWER) [6]. Although this caption evaluation metric has the potential to automate the caption quality evaluation process used by regulators, the effectiveness of this metric as a replacement of human regulator has not been evaluated so far and it is still under review for adoption by both broadcasters and regulators.

In recent years, use of Automated Speech Recognition(ASR) is observed broadly to generate caption in real time broadcast settings. To evaluate the quality of caption provided by ASR, Kafle et al. introduced 'Word Importance Model'(WIM) to measure the weight of each word in the text which is utilized during penalizing the script. This WIM posit weight to each word using machine learning based Word Prediction Model [24]. While this model automates the evaluation of caption transcription quality, to the best of our knowledge no prior research has been conducted to quantitatively investigate the degree of the effect of caption occlusion with onscreen salient content on DHH viewer's TV watching experience. Consequently none of the current caption evaluation metrics holistically evaluate the caption quality by identifying different types of occlusions that can occur across different TV genres and generate a comprehensive score to reveal DHH users' judgment.

## 4 LIST OF STUDIES AND RESEARCH QUESTION

Above mentioned prior work indicated that existing caption evaluation metrics are unable to address the caption occluding graphical content scenario and measure the variation of impact of this occlusion on quality of caption due to different TV genres.

We identified 6 different genres of live television based on the viewership trends: News, Weather News, Interviews, Emergency Announcement, Political Debate, Sports [9]. We divided some of these genres which were too broad into sub-genres. We then enlisted different types of onscreen contents that are presented in these genres. We selected the contents which were present in more than one genre. We then created labelled wire-frames which represent how TV screen looks like for different genres. We added all possible onscreen content that can be present in a TV show for the particular genre and labelled them. This was followed by a user study in which we recruited 22 DHH participants who looked at these wire-frames and provided quantitative feedback about how important it is for an onscreen content to not be occluded by caption and some qualitative feedback about why do they think certain contents are more or less important.

The quantitative feedback forms the basis of the dataset of DHH viewers preferences about onscreen content that we are releasing with this paper. We further analyzed the quantitative feedback data to investigate which types of occlusions are more problematic for specific genres. We address the following research question with our analysis of this data:

**1. Does the impact of captions occlusion with different content may vary due to the TV program genre?** Therefore, in this paper, we intend to experimentally investigate DHH users' genre-wise graphical content preference and propose a framework to measure the caption quality employing this data with a follow-up framework evaluation experiment by answering the following research questions:

- Does the impact of captions occlusion with different content may vary due to the TV program genre? To address this question we conducted a user study with DHH participants. In this study we collected participants' numerical response to various graphical contents across different TV program genre. This response data would guide us to identify the DHH users' content preferences variation across different genre and propose a prototype for a caption evaluation metric.
- How well do a genre-sensitive caption evaluation metric evaluate a captioned video than a genre-neutral caption evaluation metric? To answer this question, we developed two version of prototype metric, one is genre-sensitive and the other one is genre-neutral. First one will evaluate a captioned video stimuli employing the data genre-specific data collected from our first study and second one will evaluate a captioned stimuli without considering DHH users' genre-specific content preference. Then we conducted a follow-up study to collect DHH users' response to the same set of video stimuli which has been assessed by the prototype of the metric. A comparative analysis between the data collected from the follow-up experiment and the score generated by the genre-sensitive metric and genre-neutral metric might reveal the performance of these metrics.
- Why does the impact of caption occlusion with different content vary across different TV program genre? Since we have collected subjective response while our first user study, a qualitative analysis might rhetorically answer this research question.

## 5 EFFECT OF CAPTION PRESENTATION ON USERS' JUDGEMENTS OF CAPTION QUALITY ACROSS GENRES

### 5.1 Study Design

The objective of our study was to identify how DHH users' preference for onscreen contents varies across different genres while watching captioned live TV programming. To elicit their preference, we constructed a study website to conduct a remote study employing video conferencing software and asked provide their objective and subjective preference for different onscreen graphical elements which were depicted in wireframe diagram within each genre section. These diagrams assisted participants to envision the real-life scenario and to report their preference score in a likert scale for each of the element listed in table ??. We will describe our design of the experiment in two steps:

*5.1.1 Step 1:* While designing this experiment our first challenge was to identify the information zone of the TV screen where graphical contents resides for a certain amount of time. While identifying these information zone, we examined 60 different TV programs which were broadcast live in 15 national and local TV channels(CNN, FOX,

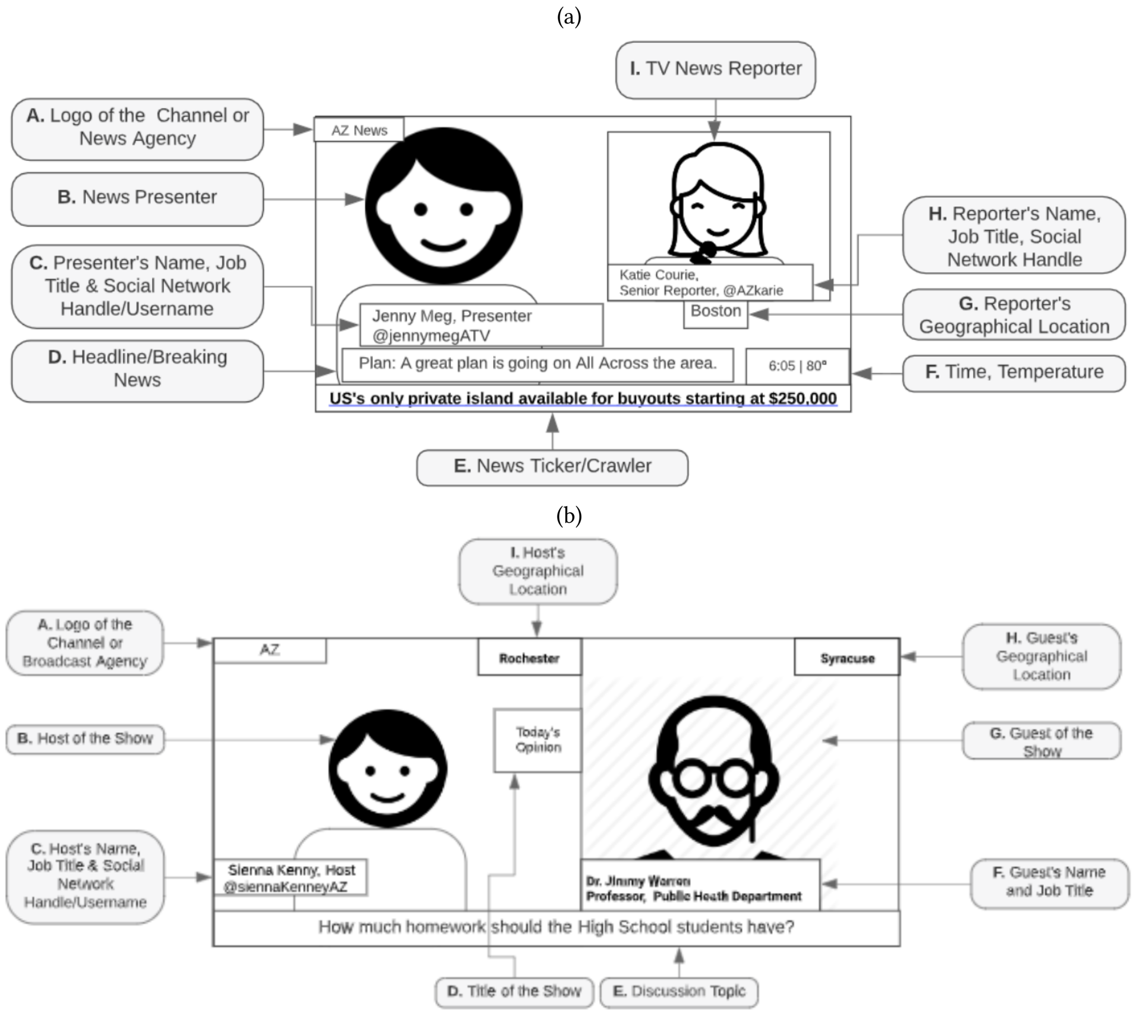
(a)
I. TV News Reporter
A. Logo of the Channel or News Agency
AZ News
B. News Presenter
C. Presenter's Name, Job Title & Social Network Handle/Username
D. Headline/Breaking News
H. Reporter's Name, Job Title, Social Network Handle
G. Reporter's Geographical Location
F. Time, Temperature
Katie Courie, Senior Reporter, @AZkarie
Boston
Jenny Meg, Presenter @jennymegATV
Plan: A great plan is going on All Across the area.
6:05 | 80°
US's only private island available for buyouts starting at $250,000
E. News Ticker/Crawler
(b)
I. Host's Geographical Location
A. Logo of the Channel or Broadcast Agency
AZ
Rochester
Syracuse
H. Guest's Geographical Location
B. Host of the Show
Today's Opinion
G. Guest of the Show
C. Host's Name, Job Title & Social Network Handle/Username
Sienna Kenny, Host @siennaKenneyAZ
Dr. Jimmy Warren Professor, Public Heath Department
F. Guest's Name and Job Title
How much homework should the High School students have?
D. Title of the Show
E. Discussion Topic

Fig. 2. Sample wireframe Diagram shown to participants for News and Interviews Genre.

MSNBC, TODAY, PBS, KCTV5 News, WWLTV, News 8 WROC, 13WHAM ABC News, WPTV News, KPRC 2, CBS Los Angeles, WKYC Channel 3, ABC10, ABC7NY, CBS Miami). After evaluating these TV programs from 6 genres, we observed a variety of content location. For example, in TV news, most of the textual information, e.g., scrolling news, news presenters' name and title, reporter's name and title tends to reside in lower segment of the TV screen, but in Weather news, most of textual information, e.g., city name, time, temperature tends to be located in the upper segment of the TV screen. In this way, for our experiment, we identified a number of content location for different TV program genres and sub-genres. We summarized the list of contents across different genres and sub-genres in table ??

*5.1.2 Step 2:* After identification of content location for each genre, we created wireframe diagram indicating these graphical elements. To construct these diagrams, we used ludichart, a widely used wireframe diagram builder [4]. While creating these diagram, we replicated the content location in this diagram and used pseudo name for these contents such as, across the news and weather news wireframe diagram we used a sample news broadcaster name *AZ News*. Then, we maintained the order of content from left to right. For instance, in figure 2(a), the left most content is "Logo of the Channel or News Agency" which has been addressed by alphabet "A" and the right most content is "TV news Reporter" which has been addressed by alphabet "I". We followed these alphanumeric order in the matrix Likert scale questions which was asked for each genre.

### 5.2 Data Collection

This experiment was conducted during a one-hour appointment with a set of DHH participant. A researcher started the experiment with sending an 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. The researcher then briefed the participants about the aim of the study. The participants were told that our goal is to understand which onscreen content they do not want to be blocked by captions during watching a live video on TV or streaming devices. The researcher then sent the participants a link to the experiment which was adapted in a Google form.

The google form was partitioned into several individual sections based on genres. Each genre section consisted of a few sub-genre sections. For a given sub-genre, participants were first shown a GIF image of how a TV screen typically looks like for that sub-genre. Afterward, the participants were shown a wireframe diagram representation of a typical screen for that sub-genre with different parts of the screen labelled. Examples of these wireframe diagrams can be seen in figure 2. The participants were asked how important was it for different segments of the screen to not be occluded by captions on a five-point Likert-scale from "Strongly Disagree" to "Strongly Agree". This was followed by three open-ended questions in which the participants were asked to explain why some onscreen content are more important, while some content are less important to them, and if they had any other comments to share.

### 5.3 Participants

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

### 5.4 Experimental Results

*RQ1: Does the impact of captions occlusion with different content may vary due to the TV program genre?* To answer this question, we showed participants one sample videos from each genre and one wireframe diagram for each sub-genre to display the content placement on the screen. For each sub-genre, participants reported their preferred onscreen content. We would like to discuss genre-wise participants' response more elaborately below: From our participants' numerical preference score, we derived a heatmap shown in figure 3. We observe from this figure 3 that, our participants have greater preference for the mouth and eyes of the speaker while watching news, interviews, political debate and emergency announcement, the face of listeners while watching interviews and political debate, the name of the current

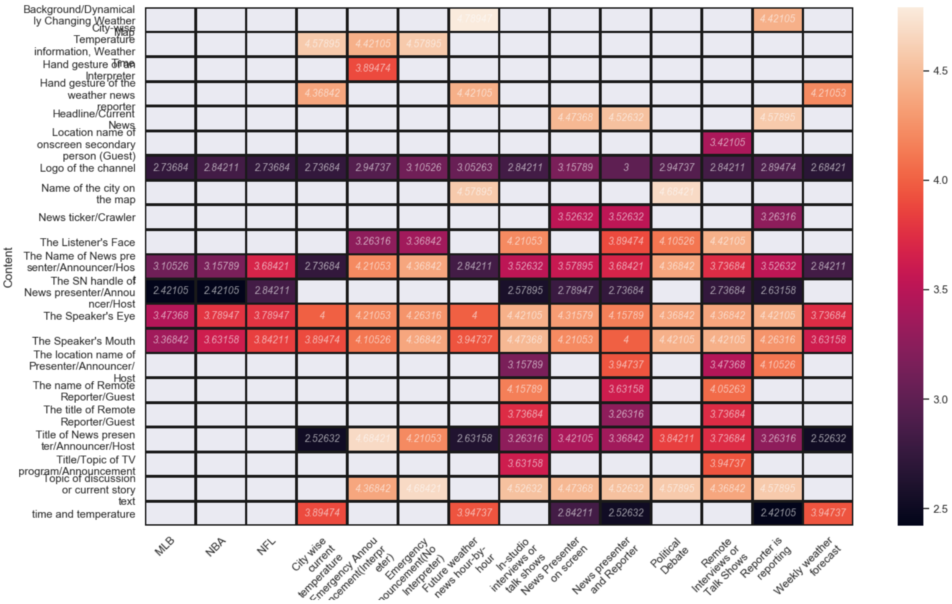

Fig. 3. DHH users' preference score Sub-genre wise Contents.

speaker while watching emergency announcement and political debate, topic of the discussion or breaking news while watching news, interviews, emergency announcement and political debate, the title of the speaker who is currently speaking while watching emergency announcement, location name of the reporter while watching news, temperature on the map or textual information on the chart while watching weather news and emergency announcement, hand gesture of the weather news reporter while watching weather news and game score, play clock, game quarter, players(pither, batter, quarterback)), timeout while watching sports.

Our objective of this study was to identify whether DHH users' onscreen content preference vary across the listed genres and sub-genres. Since, we designed our experiment in nested format more specifically each genre was divided in one or multiple sub-genre, we applied fixed effect Nested-Anova statistical method for each content across all the genres and sub-genres. Although we did not observe any significant effect of sub-genre within each genre, the fixed effect analysis revealed DHH users' significantly different preference for some contents across different genre. We summarize our nested anova results in table ??. We describe our result more elaborately below:

*5.4.1 Eyes of the onscreen Speaker.* A nested Anova was conducted to compare fixed effect of genre and sub-genre within each genre on DHH users' preference for the eyes of the onscreen person who is speaking. The fixed effect for genre yielded an F ratio of F(5, 19) = 4.064, p = .001, indicating a significant difference between news (M = 4.16, SD = 0.94), interviews (M = 4.45, SD = 0.76), emergency announcement (M = 4.24, SD = 0.88), political debate (M = 4.42, SD = 0.84), weather news (M = 3.82, SD = 1.34) and sports (M = 3.61, SD = 1.29), but no effect of sub-genre was observed within each genre. However, a followup pairwise Posthoc comparisons using the Tukey HSD test did not reveal any significant difference.

| Content Name | Genre | Genre Effect |
|---|---|---|
| Location Name of the speaker | News.<br>Interviews. | F(1,19) = 9.639, p = 0.00272 |
| Onscreen Listeners' Face | Emergency Announcement.<br>Interviews.<br>News.<br>Political Debate. | F(3,19)=4.757, p = 0.00373 |
| Time and Temperature | News.<br>Weather News. | F(1,19)=26.618, p=1.13e-06 |
| Title of the Speaker | News.<br>Interviews.<br>Emergency Announcement.<br>Political Debate.<br>Weather News. | F(4,19)=11.829, p=1.23e-08 |
| Name of the person who is speaking | News.<br>Interviews.<br>Emergency Announcement.<br>Political Debate.<br>Weather News.<br>Sports | F(5,19)=12.799, p=4.27e-11 |
| Onscreen Speakers' Eyes | News.<br>Interviews.<br>Emergency Announcement.<br>Political Debate.<br>Weather News.<br>Sports. | F(5,19)=4.064, p=0.00145 |

Table 3. List of content and their corresponding Genre in which participants' preferences for these content vary significantly.

*5.4.2 Name of the person who is speaking.* A nested Anova was conducted to compare fixed effect of genre and sub-genre within each genre on DHH users' preference for the name of the person who is speaking. The fixed effect for genre yielded an F ratio of F(5, 19) = 12.799, p < .0001, indicating a significant difference between news (M = 4.16, SD = 0.94), interviews (M = 4.45, SD = 0.76), emergency announcement (M = 4.24, SD = 0.88), political debate (M = 4.42, SD = 0.84), weather news (M = 3.82, SD = 1.34) and sports (M = 3.61, SD = 1.29), but no effect of sub-genre was observed within each genre. Then, a followup pairwise Posthoc comparisons using the Tukey HSD test indicated that DHH users' mean preference score for name of the onscreen speaker was significantly different between while watching weather news (M = 3.82, SD = 1.34) and while watching interviews (M = 4.45, SD = 0.76), while watching weather news (M = 3.82, SD = 1.34) and while watching emergency announcement (M = 4.24, SD = 0.88) and while watching weather news (M = 3.82, SD = 1.34) and while watching political debate (M = 4.42, SD = 0.84).

*5.4.3 Title of the speaker.* A nested Anova was conducted to compare fixed effect of genre and sub-genre within each genre on DHH users' preference for the title of the person who is speaking. The fixed effect for genre yielded an F ratio of F(4, 19) = 11.829, p < .0001, indicating a significant difference between news (M = 4.16, SD = 0.94), interviews (M = 4.45, SD = 0.76), emergency announcement (M = 4.24, SD = 0.88), political debate (M = 4.42, SD = 0.84) and weather news (M = 3.82, SD = 1.34), but no effect of sub-genre was observed within each genre. Then, a followup pairwise Posthoc comparisons using the Tukey HSD test indicated that DHH users' mean preference score for name of the

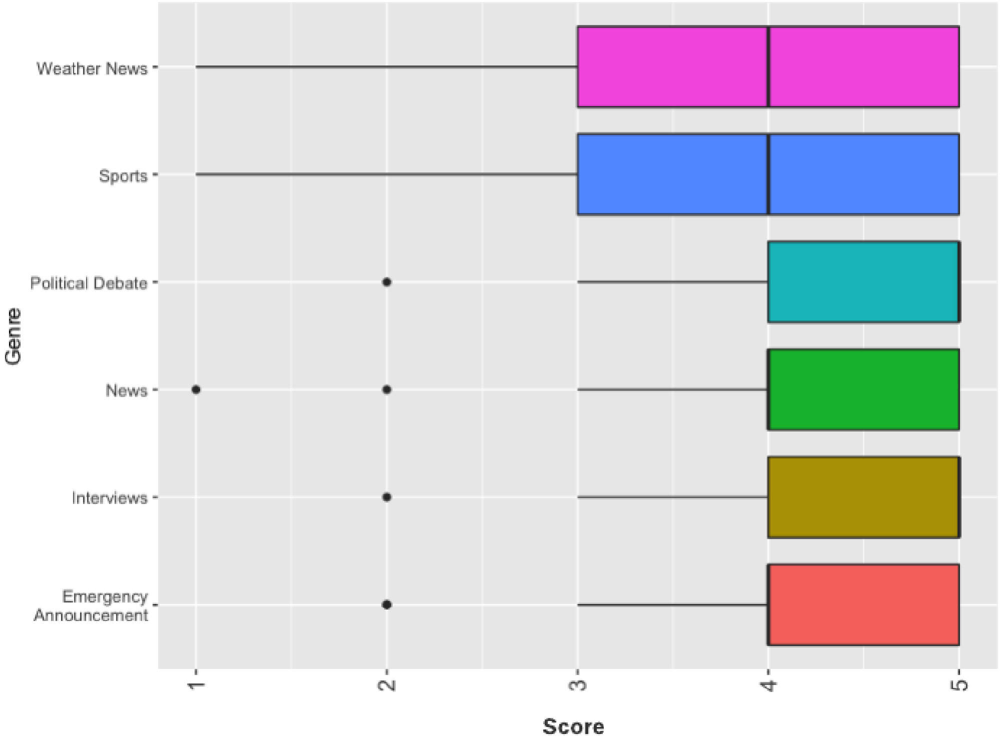

Fig. 4. Participants' subjective scalar response for Onscreen Speakers' Eyes across different genre.

onscreen speaker was significantly different between while watching weather news (M = 3.82, SD = 1.34) and while watching interviews (M = 4.45, SD = 0.76), while watching weather news (M = 3.82, SD = 1.34) and while watching emergency announcement (M = 4.24, SD = 0.88) and while watching weather news (M = 3.82, SD = 1.34) and while watching political debate (M = 4.42, SD = 0.84).

*5.4.4 Location Name of the speaker.* A nested Anova was conducted to compare fixed effect of genre and sub-genre within each genre on DHH users' preference for the location name of the person who is speaking. The fixed effect for genre yielded an F ratio of F(1, 19) = 9.639, p = .002, indicating a significant difference between news (M = 4.16, SD = 0.94) and interviews (M = 4.45, SD = 0.76). However, a followup pairwise Posthoc comparisons using the Tukey HSD test did not reveal any significant difference.

*5.4.5 Time and Temperature.* A nested Anova was conducted to compare fixed effect of genre and sub-genre within each genre on DHH users' preference for the time and temperature. The fixed effect for genre yielded an F ratio of F(1, 19) = 26.618, p < .01, indicating a significant difference between news (M = 4.16, SD = 0.94) and weather news (M = 3.82, SD = 1.34), but no effect of sub-genre was observed within each genre. Then, a followup pairwise Posthoc comparisons using the Tukey HSD test indicated that DHH users' mean preference score for name of the onscreen speaker was significantly different between while watching news (M = 4.16, SD = 0.94) and while watching weather news (M = 3.82, SD = 1.34).

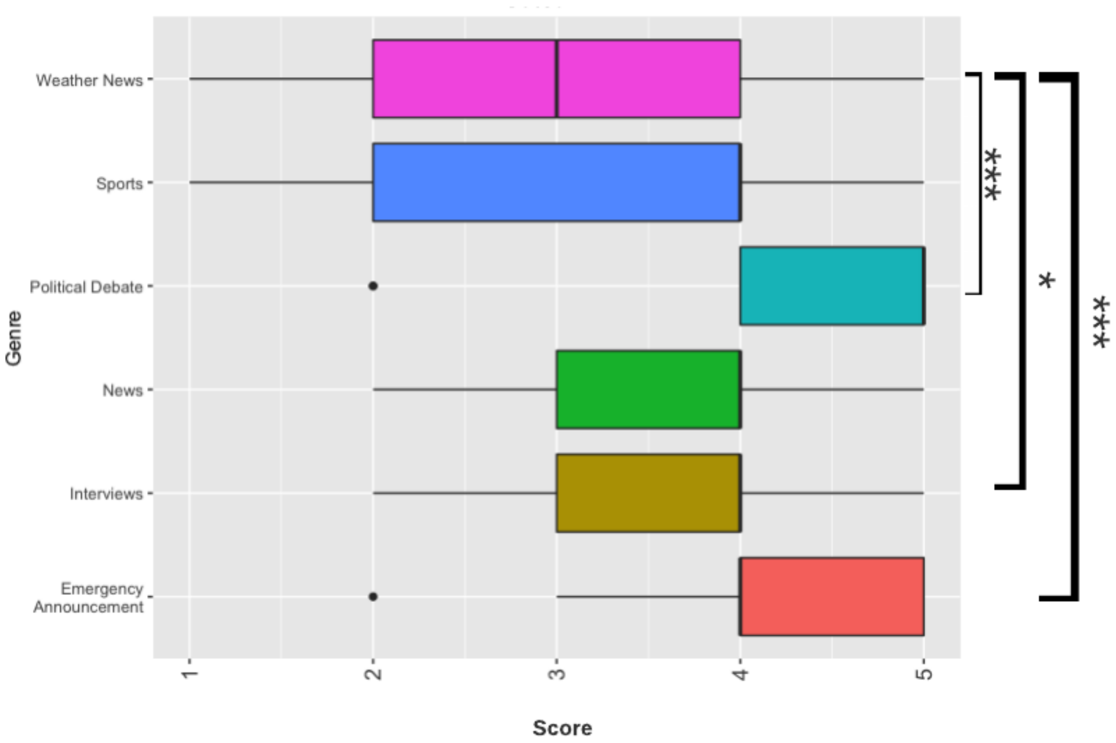

Fig. 5. Participants' subjective scalar response for name of the onscreen person who is speaking across different genre.

*5.4.6 Onscreen Listeners' Face.* A nested Anova was conducted to compare fixed effect of genre and sub-genre within each genre on DHH users' preference for the face of the onscreen person who is listening. The fixed effect for genre yielded an F ratio of F(3, 19) = 4.757, p < .01, indicating a significant difference between news (M = 4.16, SD = 0.94), interviews (M = 4.45, SD = 0.76), emergency announcement (M = 4.24, SD = 0.88), political debate (M = 4.42, SD = 0.84), but no effect of sub-genre was observed within each genre. Then, a followup pairwise Posthoc comparisons using the Tukey HSD test indicated that DHH users' mean preference score for the face of the onscreen person who is listening was significantly different between while watching interviews (M = 4.45, SD = 0.76) and while watching emergency announcement (M = 4.24, SD = 0.88).

## 6 BUILDING PROTOTYPE OF THE CAPTION EVALUATION METRIC

We wanted to understand whether the effect of content occlusion varies across different genres and thus warranting a need of including a genre-based framework in existing caption evaluation metrics. To this end, we designed two frameworks, one genre-sensitive and the other genre-neutral, described below.

### 6.1 Genre-sensitive prototype metric

This metric evaluates each stimuli employing their genre specific property. For instance, if the genre of a video stimuli is news, while evaluating the stimuli, this metric would consider "News Presenters' Face" as most important content, at

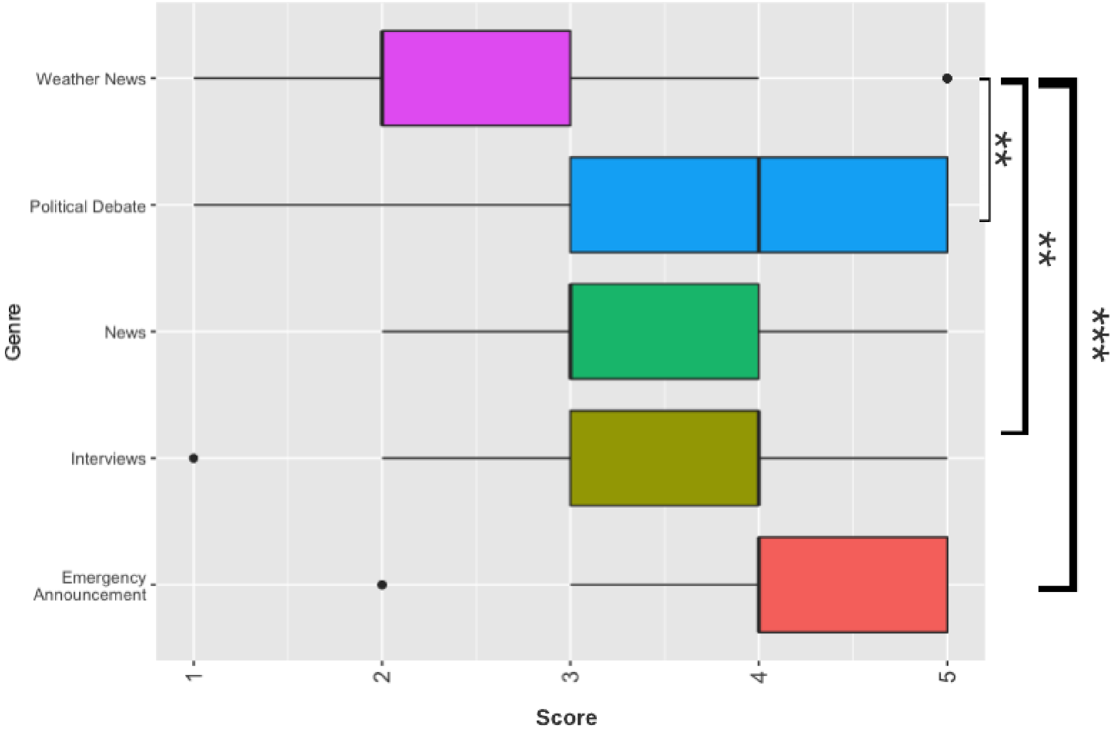

Fig. 6. Participants' subjective scalar response for title of the onscreen person who is speaking across different genre.

the same time if the genre of the video stimuli is "Weather News", while evaluating the stimuli, this metric would consider onscreen textual information such as "Temperature", "Weather Time" as most important content. Which means while calculating the cumulative caption quality score, this framework applies numerical co-efficients for each content and these co-efficients were determined by the results of the genre-wise experiment. If the caption does not occlude any salient onscreen content, the metric returns a value of 0. More particularly, the more the score the metric returns for a captioned video, the lower is the quality of caption. In other words, our framework tends to penalize a video stimuli by evaluating whether the caption location is occluding any salient underlying content. Here is the equation, we used, for calculating the penalized score of the captioned video:

$$\frac{(t_1E_1w_1 + t_2E_2w_2 + t_3E_3w_3 + .... + t_NE_N.w_N)}{T} \tag{1}$$

In equation 1, $t_1$ is the number number of seconds in which the caption occludes content 1, $E_1$ is the average area of content that has been occluded by caption, $w_1$ is the weight of content 1 (the co-efficient determined based on the genre-wise experiment) and T is the length of the stimuli in second. This is how we calculate this score for each of the content 1,2,3...N. The score of individual content are added to produce the summative score, and normalization is applied by dividing the total score by the length of the stimuli in second.

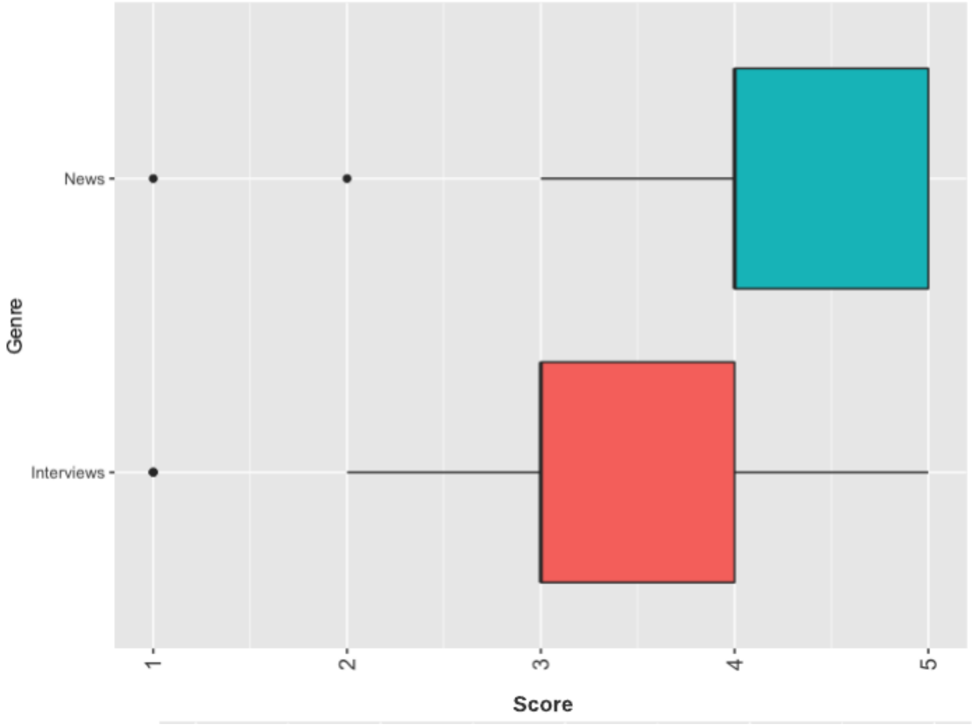

Fig. 7. Participants' subjective scalar response for location name of the onscreen person who is speaking across different genre.

### 6.2 Genre-neutral prototype metric

This version of the metric evaluates each stimuli disregarding the genre-specific property. For instance, this metric considers the importance of "Presenters' Face" in a stimuli from news genre is same as a stimuli from weather news genre. Therefore, while calculating the score of a captioned video, this metric does not apply different weights for the same content across various genres. In other words, this metric applies unique weight for each of the content across different genres. The equation for calculating the score of the captioned video becomes:

$$\frac{(t_1E_1\bar{w_1} + t_2E_2\bar{w_2} + t_3E_3\bar{w_3} + .... + t_NE_N.\bar{w_N})}{T} \tag{2}$$

In equation 2, $t_1$ is the number of seconds in which the caption occludes content 1, $E_1$ is the average area of content that has been occluded by caption,  refers to the weight of content 1 which remains same across different genre and T is the length of the stimuli in second. This score is calculated following the same procedure as genre-sensitive metric.

## 7 COMPARISON BETWEEN GENRE-SENSITIVE AND GENRE-NEUTRAL CAPTION EVALUATION METRIC

The goal of this study is to evaluate the prototype of the two types of metric developed in previous step of this research. In this study, we showed participants 11 different videos each with three different caption locations. The stimuli video

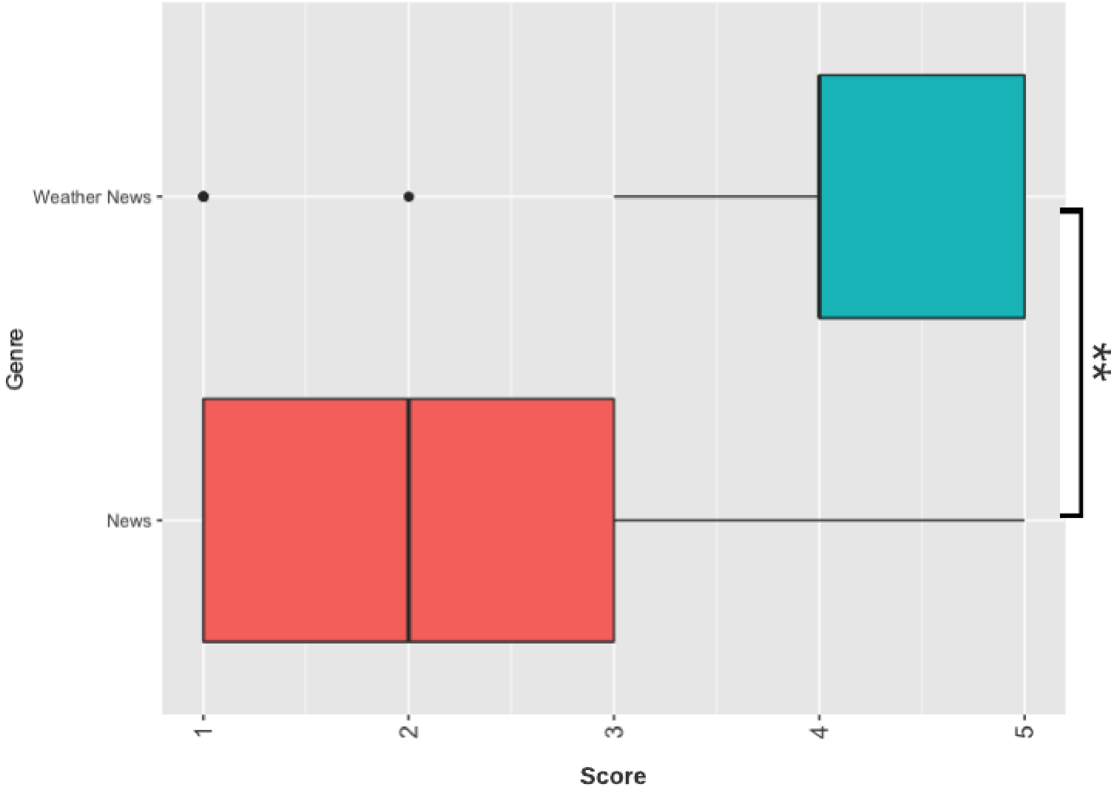

Fig. 8. Participants' subjective scalar response for time and temperature across different genre.

column in table 1 summarizes where these videos were taken from. For each of the source identified in the table we took one video except NFL Youtube Channel for which we took two videos.

### 7.1 Experimental Design

While designing the experiment, we had to select at least one video stimuli from each genre that we examined in our preliminary study. The main challenge we encountered while selecting these stimuli was the presence of most of the contents that we have listed in table ??. After examining 110 videos from 15 different TV channels, we selected 11 stimuli across 6 different genres. We then manually modified the caption file of each stimuli by examining the audio information so that no speech and non-speech information remains excluded. Our next challenge was to select the placement of the caption. The experiment design described in [32], previous caption position guideline EIA 608 [8] and our manual observation of a captioned video revealed three location might be the common location of the caption. Here is the list of caption location we determined for our follow-up experiment: *Upper segment of the lower third of the TV screen, Lower segment of the lower third of the TV screen and Upper third of the TV screen.* After determining the caption location, we engineered caption file and embedded them with each video stimuli employing FFMPEG [17], an

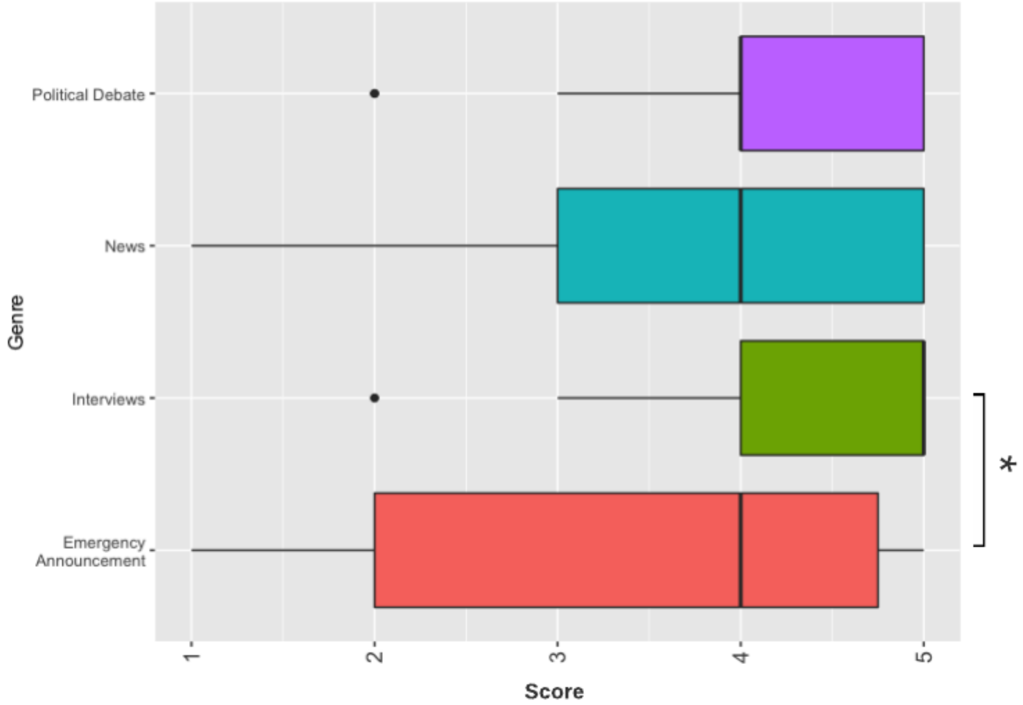

Fig. 9. Participants' subjective scalar response for onscreen person's face who is listening across different genre.

open source video editing tool, to create three version of each stimuli in which captions are located in the three common location.

### 7.2 Data Collection

Our experiment was conducted during an one-hour appointment with a set of DHH participant. A researcher started the experiment with sending an 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. The researcher then briefed the participants about the aim of the study which is to obtain their feedback about various caption positions. Participants were shown videos with three different placements of captions across 11 different sub-genres. Subsequently, they were asked if they were happy with the location of the captions on a ten-point smiley-face scale. At the end of the experiment, we asked our participants if they had any comments.

### 7.3 Participants

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

### 7.4 Experimental Results

## 8 DISCUSSION

## 9 FUTURE WORK AND CONCLUSION

## REFERENCES

[1] [n.d.]. *BBC Subtitle Guidelines, 2018.* https://bbc.github.io/subtitle-guidelines

[2] [n.d.]. *The Described and Captioned Media Program, 2010, Captioning Key for Educational Media, Retrieved from: http://access-ed.r2d2.uwm.edu/resources/captioning-key.pdf.* https://www.fcc.gov/document/closed-captioning-quality-report-and-order-declaratory-ruling-fnprm

[3] [n.d.]. *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.* https://www.fcc.gov/document/closed-captioning-quality-report-and-order-declaratory-ruling-fnprm

[4] [n.d.]. *https://app.lucidchart.com/.*

[5] [n.d.]. *Measuring live subtitling quality, UK. Retrieved from: https://www.nidcd.nih.gov/health/captions-deaf-and-hard-hearing-viewers.* https://www.nidcd.nih.gov/health/captions-deaf-and-hard-hearing-viewers

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

[7] [n.d.]. *TV News Graphics Package.* NewscastStudio, The trade publication for TV production professionals. Retrieved from: https://www.newscaststudio.com/tv-news-graphics-package/.

[8] 2012. SCTE 21 2012 - STANDARD FOR CARRIAGE OF VBI DATA IN CABLE DIGITAL TRANSPORT STREAMS. *(PDF). Society of Cable Telecommunications Engineers. SCTE* 21 (Oct. 2012), 13.

[9] 2020. *THE NIELSEN TOTAL AUDIENCE REPORT: APRIL 2020.* The Nielsen Company (US), LLC.

[10] 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

[11] 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

[12] 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

[13] 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.*

[14] Stephen Cushion. 2015. *News and Politics: The Rise of Live and Interpretive Journalism.* https://books.google.com/books?id=5MgqBwAAQBAJ&pg=PA44&lpg=PA44&dq=does+time+for+live+reporting+increased+than+studio+news&source=bl&ots=CW4o9_uNCP&sig=ACfU3U3DSLAUDNkoehrImj47AhuuM3WMbw&hl=en&sa=X&ved=2ahUKEwjlobDC6t7pAhWaVs0KHZKsCi4Q6AEwAHoECA0QAQ#v=onepage&q=does%20time%20for%20live%20reporting%20increased%20than%20studio%20news&f=false

[15] 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

[16] Zoé de Linde and Neil Kay. 1999. Processing Subtitles and Film Images. *The Translator* 5, 1 (1999), 45–60. https://doi.org/10.1080/13556509.1999.10799033 arXiv:https://doi.org/10.1080/13556509.1999.10799033

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

[18] 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).

[19] 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.

[20] 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.

[21] 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

[22] 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).

[23] 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*.

[24] 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

[25] 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

[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] 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.

[31] S. Nam, D. I. Fels, and M. H. Chignell. 2020. *Modeling Closed Captioning Subjective Quality Assessment by Deaf and Hard of Hearing Viewers.* In Proceedings of IEEE Transactions on Computational Social Systems, DOI: https. http://doi.org/10.1109/TCSS.2020.2972399

[32] 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

[33] 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

[34] 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.

[35] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia. [n.d.]. DEEP-HEAR: A multimodal subtitle positioning system dedicated to deaf and hearing-impaired people. *IEEE Access 7, 150–162* 88 ([n. d.]).

[36] 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*.

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

[38] X. Zhu, J. Guo, S. Li, and T. Hao. 2020. Facing Cold-Start: A Live TV Recommender System Based on Neural Networks. *IEEE Access* 8 (2020), 131286–131298.