

Immersive Captions CG Final Draft

Last edited May 30, 2022

Introduction

The Need

Technology moves at an increasingly fast pace, yet accessibility has often been an afterthought—if it is one at all. Immersive video experiences, still in their infancy, give us an opportunity to ensure accessibility is built in from the start. It is our hope that this shift brings innovation to captioning in all relevant tech areas. After all, WHO (World Health Organization) predicts that by 2050 there will be [900M people with hearing](#) loss.

XR experiences are moving toward the mainstream, and there's a growing realization within the tech industry of the importance of accessibility in all emergent technology. "XR" is the abbreviation for extended reality and to represent both virtual reality (starting with the Sensorama in 1956) and augmented reality (first demonstrated by USAF Armstrong's Research Lab by Louis Rosenberg in 1992). XR is powering the widespread reach of immersive technologies, and with the support of the tech industry, accessibility is finally being discussed *before* the technology has reached real adoption.

Though there is a need for innovative captioning across all immersive experiences, this paper will focus on immersive and 360 video. In virtual reality (VR), interesting questions arise about how to effectively provide captioning because, for on example, the source of sound in the virtual experience may not be in the viewer's field of view. In that scenario, a deaf participant would not understand what is happening if there are no captions or if the captions have significant limitations.

Who We Are

In July 2019, Cornell Tech and Verizon Media sponsored the inaugural XR Access Symposium held in New York City. A collection of tech companies, educators, creators, and advocates attended breakout sessions and presentations on accessibility in XR. Several of the participants banded together in an effort to create immersive captioning.

Christopher Patnoe of Google volunteered to found and chair the community group and connected with Judy Brewer of World Wide Web Consortium (W3C). Together they formed an [Immersive Captions Community Group \(ICCG\)](#) under the Web Accessibility Initiative (WAI). The group is a broad coalition of people who identify as d/Deaf, hard of hearing, and hearing. We are made up of industry (Google, Meta Platform - formerly known as Facebook, etc.), advocates (National Association of the Deaf), educators (Brandeis University, Gallaudet University, Rochester Institute of Technology, University of Salford), and technologists.

We have spent the past two years exploring immersive technology and captioning opportunities. Our goal has been to break the boundaries of the little black box inherited from TV and find new ways to make virtual reality more inclusive. This paper represents our narrow but critical effort to

document our insights and ideas for accessible 360 video, and the seeds of this effort will impact future research.

Aligned efforts

No important work is ever done in a vacuum. Other organizations are also engaged in this discussion, including the XR Access initiative that started this effort. There is also the ImAc effort out of the EU whose research was core to our work.

At this time in the W3C, there are other initiatives working in the caption space such as:

- XAUR — XR Accessibility User Requirements
- RAUR — Real-time Communications (RTC) Accessibility User Requirements
- RTC specs like WebRTC [[webrtc](#)] and the IETF RFC 7478 'Web Real-Time Communication Use Cases and Requirements' document. [[ietf-rtc](#)]

Disclaimers

This group does not intend to declare *how* something is to look; we have taken great efforts to avoid this. Instead, we've tried to focus on what functionality will solve problems of situational awareness and to provide recommendations on functionality that will make the experience of the deaf or hard of hearing viewer more equal through technology. It is important to remember that there is no single solution that will work for everyone, so it's ideal to offer flexibility and tools that enable a user to customize the functionality to meet their own needs.

Recommended Features and Suggestions

Our group has identified a number of features as critical and advisory recommendations. We hope these considerations inspire best practices for content creators.

Scope

Our group agreed that captioning within VR must encompass all sounds and speech so that any deaf or hard of hearing user can perceive, understand, and fully participate in the experience to the same extent as everyone else. Therefore, the captioning must include localized speech from an identified speaker (human or otherwise) and non-localized speech, i.e. "Voice of God" audio content. Captioning must include all [spatial sounds](#) including inanimate objects sounds (e.g. weapons, appliances), animate objects sounds (e.g. footsteps, animal calls), point ambience sounds (spatialized ambient sounds that are diegetic such as a river near the user), surrounding ambience sounds (diegetic non-spatialized ambient sounds such as a crowd), and music sounds (non-diegetic background music).

Speaker Identification

In 360 video, the viewer defines their own viewing area and may not be looking at the source of sound(s). It is critical in 360 space to have indicators of speaker identification and/or audio location to help equalize the experiences of all users, regardless of disabilities. One example is to provide a speaker's identification as text ahead of each caption; users can have the option to only display the speaker identification when the source changes. In the spirit of keeping the experience similar to all users, care must also be taken to avoid caption "spoilers" such as pre-announcing the arrival of a character before they should be known. This may apply to keeping mystery around certain characters if it is appropriate in the storytelling. We noted an

example where a character arrived to the film with the label “The Snitch” in the captioning long before that information was revealed to other viewers.

Immersivebill

Identifying speakers is critical, but frequently using additional text in captioning affects the experience in other ways. We propose the ability for users to choose other methods for distinguishing speakers, such as options to use text labels, names, icons, and/or colors to represent each character as shown in figure 1. *Note that using color representation is not enough on its own as it requires users to remember the colors and this could also affect usability for someone with color blindness.*



Figure 1 - Examples of speaker Identification, include name, icon, color.

Our prototype presents these options prior to playing the video. Our group calls this the Immersivebill¹ feature. Inspired by Playbill's Cast List, this feature shows pictures of the characters (speakers) before the start of the 360 video. By default, each character in the 360 video has a text color assigned to them. If a viewer wants to assign certain colors to characters or background sounds, they can customize their preferences. Users can also customize the speaker ID to represent characters by text or image. See figure 2 for a sample of the Immersivebill from our working prototype. (W. Dannels, personal communication, coined the term “immersivebill” on April 1, 2021.)

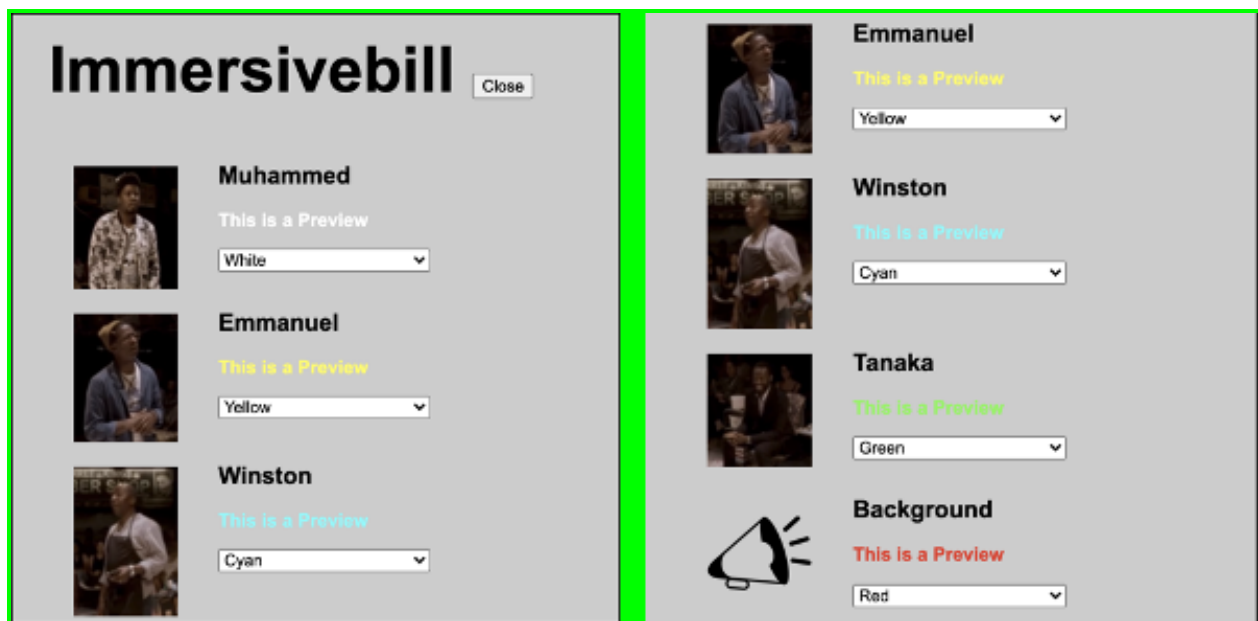


Figure 2 - List of where the audio comes from, preview captions and selected colors.

Additional Audio Source Identification

Identifying the source and location of audio besides speakers is critical to equalizing the experience for all content viewers. To assist in communicating where specific audio is coming from, one solution could be guide arrows. These helpful indicators should point towards the location of the audio, speaking character, or object. For consideration, different arrows could be used to identify different characters or sound sources, and the size and location of the arrow could be used to indicate the relative distance to the source. *Content creators should consider the relative prominence of specific audio in their assistive design to ensure these visuals do not overwhelm the primary content.* The key here is that the solution must provide visualization to indicate and locate sound coming from anywhere rather than just those in the user's current field of view. Covered spatial sounds should be enabled by default with an option to disable.

Interactive Captions and Transcript View

Our group proposes that users have options for different captioning modes that we refer to as interactive captions (IC) and transcript view. IC, as shown in figure 3, provides a few seconds of captioned conversation that can be used as dialogue-based navigation. The user can seek backwards and forwards by caption rather than by time. Although this is not in our "critical" selections, we find that interactive captions provide an aspirational framework for truly innovative captioning.

The interactive caption feature is an intermediate between captions and full transcripts. The goal of using interactive captioning is to give the user an option to follow the conversation in a way that is less rushed because captions generally only stay on the screen for 1-2 seconds. While having full transcripts is very useful for users such as Deafblind users, it is too long to display on screen for all platforms and lacks the ability to communicate spatial sound information.

Interactive captions can be enabled when a user wants to follow the conversation in more than one line of caption at a time. If a user chooses to preview 3-7 lines of captions at once, this makes it possible for each line to clearly demonstrate the unique speaker ID, speaker caption color, and the relative location of the speaker with the use of triangles pointing toward the person saying the line or other indicators. One interesting discovery was that slowing down playback makes it easier to follow captions. Currently our recommended playback speed is 75%, but a user could adjust the speed according to their needs.²

Seek-by Captions

A seek-by captions feature, as shown in figure 4, allows users to click on a caption to jump to the time in video or media where that specific caption occurred. Since viewers pay attention to what is said rather than time in video when that said line occurred, this feature makes it easier to find a specific part of the video. One can use the next/previous caption buttons to move between captions. Combined with interactive captions, this one-click navigation method allows users to more easily get to the section of media they wish to access. (Credit to Chris Huges.)

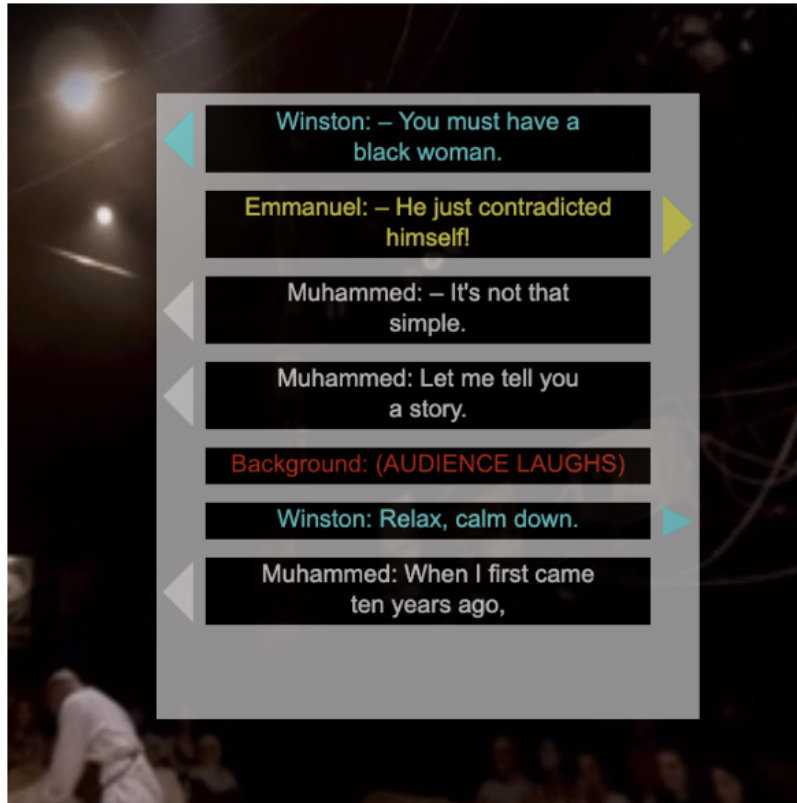


Figure 3 - A sample image of interactive captions

2. W. Dannels (personal communication) recommended to reduce playback rate to either .50 or .75 on April 1, 2021.

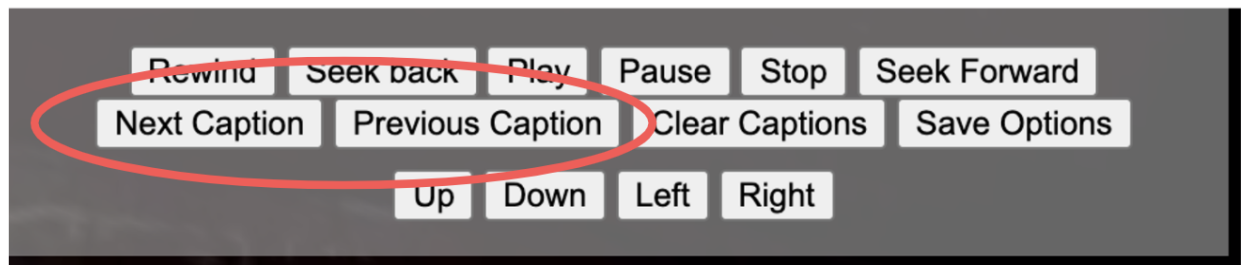


Figure 4 - A sample image of seek-by captions to navigate content

In contrast, the transcript view is a scrollable view of the entire dialogue and key sounds during the entire pre-recorded media session. It is not our intention to replace transcripts. Transcripts are useful for consuming the content via other technologies like Braille displays for DeafBlind users and screen readers for people with low vision. The transcript should always be available separately to aid in integration with a braille display, but may also be integrated in the UI of the video. The transcript must be well-formatted with clear speaker identification and broken into reasonable-sized paragraphs. The transcript view could be modified for interactivity on its own, yet the interface must allow for easy activation of the controls and settings.

Headlocked and Non-Headlocked Options

In terms of the visual display, our group generally preferred for the captions to persist within the user's field of view (FOV), which allows the deaf or hard of hearing user to understand via captioning while being able to look around. However, the user needs to have the ability to customize captions and adjust them for personal needs without complex barriers to configure those choices. It is also preferable to limit the visual movement of the captions.

There is no "right" answer for how captions should be presented in immersive video experiences. Platform and experience designers should allow for maximum flexibility. In our group, some people liked having the captions headlocked, while others liked having them anchored so they knew where to look or so they could avoid physical discomfort (see section on vestibular disorders). Thus, the optimal design allowed options for the user to customize.

Captions Anchoring

We experimented with a captions-anchoring feature that displayed captions next to their speaker and followed the speaker throughout the video. While adding too many moving components to a media could be overwhelming, some in our group wanted an option to "anchor" captions to the speaker rather than moving the caption with the user's field of view. The practicality of this feature could vary depending on the type of content and the amount of motion. It's also worth noting that most existing captioning formats do not support the type of metadata needed to include location information for captions. Still, this is an interesting opportunity to innovate existing captioning formats and add important metadata such as coordinates of characters to be able to anchor captions in immersive videos. This work could lead to other non-caption related features we haven't even considered. (Credit to ImAc)

Usability and User Interface

The user interface (UI) must allow for easy activation of each of the controls and settings, such as the ability to enable and disable both the captions and interactive captions mode. It must also provide customization for the UI including font, color, and playback speed (1.0x, 0.75x, and 0.5x). The UI for IC should also have customization options for speaker identification such as through the use of text, icons, or colors.

Key recommendations also include further customization for appearance (font, color, size, background, etc.), number of lines, and length of display time (time to live - TTL), which can be set to the caption end time, a fixed time, or a fixed extension time. This can also be managed through the option to set a number of captions which can remain on the screen at one time, with the oldest only being removed once the maximum number has been exceeded. There should also be an option for a pan and scan mode, but this mode should have the option to be disabled by users who may have motion sickness.

As with any long list of options, there should be easy methods for users to set their choices. There should be logical and intuitive groupings of features that belong together so it's easy to choose. A user experience professional could assist in these design considerations. Accessibility features should not be difficult to locate or adjust.

Caption Display and Formatting

Captions should always comply with the guidelines for the specific display type (e.g. TV, HMD). With respect to latency, it is critical that each segment of caption starts no more than 70ms from the start of the dialogue segment. The appearance of the captions should follow the regulations

of the Federal Communications Commission ([FCC](#)) which require appropriate usage of background and foreground color, text size, and positioning.

Content creators may wish to consider that reading captions when there are multiple speakers can be very challenging for caption readers because speakers can rapidly take turns in any order. This is especially true in immersive video, where the viewer has full control of their FOV but may not know where to focus their attention. The caption reader is usually focused on reading the captions and cannot anticipate who will speak next. As a result, the reader usually looks back and forth between captions and the speakers, and they become tired or distracted. Additionally, readers can feel left out of the conversation.

Physical Considerations

Vestibular Disorders

Immersive experiences can be challenging for those with vestibular disorders, and the additional movement of captions, including headlocked captions, can be difficult for some users. One study has found an estimated [70 percent](#) of deaf and hard of hearing children with sensorineural hearing loss have a vestibular disorder. Vestibular disorders affect many people, not just those who are deaf or hard of hearing. The Vestibular Disorders Association says more than [35 percent of US adults aged 40](#) and older experience vestibular dysfunction at some point in their lives. It's always a better user experience to let viewers have control over any motion.

Mental Processing

Meaningful understanding of audiovisual information requires mental processing for both aural and visual information and for understanding connections between them. Deaf and hard of hearing viewers often have to choose between watching visuals or aural-to-visual translation — the captions. Regardless of the choice, some information may be lost. They rarely have enough time to perceive and understand both visuals. Consequently, information needs to be presented in a way that gives sufficient time to process both effectively. In addition, viewers often encounter visual noise, such as line of sight interference, obstruction, or poor lighting. Visual noise tends to be a mere annoyance for hearing users, but it can significantly interfere with visual access for deaf and hard of hearing users.

Reading captions when there are multiple speakers can be very challenging for caption readers because speakers can rapidly take turns in any order. This is especially true in immersive video, where the viewer has full control of their FOV but may not know where to focus their attention. The caption reader is usually focused on reading the captions and cannot anticipate who will speak next. As a result, the reader usually looks back and forth between captions and the speakers, and they become tired or distracted. Additionally, readers can feel left out of the conversation.

While the traditional method of inserting the speaker's name in captions for speaker identification has been shown to be useful for viewers who are hard of hearing, studies have also found that viewers who were deaf did not view it as useful. One reason why deaf viewers may not view traditional captions as useful is that captions that stay in one location do not show the location of the speaker. Furthermore, if the speaker moves around, this creates distance between the visual information and captions, which forces the viewer to move his/her eyes or head constantly. As a result, the student will focus only on the captions or visuals to his/her detriment.

Viewers often use captions to confirm or correct what they heard because they can review spoken information by reviewing captions that are displayed for several seconds. Generally, they prefer to have several lines of captions. On the other hand, viewers who do not listen to speech prefer to have 2-3 lines of text to minimize scanning time for words as they read.

Where We Are Going

The features discussed in this paper are a small portion of the ideas that were discussed to address the many issues that arise from adding immersive captions to 360 media. Before we can find solutions to unexplored issues, we need to share our findings with other groups and get input from other W3C groups who were involved in producing this report.

As we continue our research, there are questions that we were unable to answer. This technology is new and we will continue to learn. Here are some of the questions that need to be answered before wide adoption is possible. There is a logical order of operations for the work to be done.

Regarding the technology, there are questions about the base spec for making changes. If the goal is to have a consistent interoperability for uses in VR, we would seek for a unified version of the spec so there aren't the [myriad of captioning specs](#) that we have to deal with today in 2D video across the internet. Some interesting opportunities could be the files such as SubRip (.srt), Timed Text Markup Language (.ttml), or WebVTT (.vtt).

Once we understand the specification, we should settle on the changes themselves. How do we describe the changes when dealing in a 360 environment? Our tool uses latitude/longitude and it works well, but this may not integrate nicely into all of the formats. And of course, as features are developed there will be further changes needed for the spec.

Once we have agreed on the spec and how to adapt it, we need to have an authoring environment that supports it. Thanks to the prototyping [tool created by Chris Hughes](#), we have an excellent place to start. But as anyone who builds prototypes understands, it takes a great deal of effort to go from proof of concept to shipping product. This Community Group seeks for the tool to be open sourced. and we welcome anyone who is willing and able to participate.

Similarly, we would like to share our findings with scholarly and academic publishing platforms, as well as technology magazines and publishers to get feedback and create opportunities for research collaborations. We would like to collaborate with our industry partners to explore integrating our features into current VR platforms, devices, and operating systems. This will allow us to fine-tune proposed solutions as well as explore new ideas.

Another interesting way of getting user feedback is planning Hackathons based around our findings. During these sprint-like events, experts can explore ways of improving our features and show new ways of using these findings to make 360 media accessible to all. The opportunity to innovate existing captioning formats shines brightly!

Surely, the more feedback we get, the more refined our solutions will be. Discussing our findings with partners and other passionate minds will invigorate this opportunity to innovate and close the gap between existing captions and then needs of emerging display technologies.

Special Thanks

These Immersive Captions Community Group findings were developed by the [W3C Immersive Captions Community Group](#) (ICCG). We would like to acknowledge the following leaders and participants.

Project Leadership

Lead editor

- Christopher Patnoe, Google LLC

Supporting editors

- Frances Baum
- Meryl K. Evans
- Howard A. Rosenblum, National Association of the Deaf

ICCG chair

- Christopher Patnoe, Google LLC

Contributing ICCG Participants

- Bill Curtis-Davidson, Partnership on Employment & Accessible Technology (Wheelhouse Group)
- Chris Hainsworth, Blind Burners
- Chris Hughes, University of Salford
- Delbert Whetter, Exodus Film Group
- Dylan Fox, XR Access
- Frances Baum
- George Karalis, Strivr
- Hannah Barrett
- Hazal (H) Uzunkaya, Brandeis University
- Howard A. Rosenblum, National Association of the Deaf
- Ian Hamilton
- Johan Verstraete
- Judy Brewer, W3C
- Ken Harrenstien, Google LLC
- Larry Goldberg, Yahoo Inc.
- Lonnie Ko, Meta Platforms, Inc. (formerly Facebook Technologies, LLC)
- Matthew Johnston, ThoughtWorks LTD
- Meryl K. Evans, Equal Entry and Diamond
- Myles de Bastion
- Raja Kushalnagar, Gallaudet University
- Samir Jain
- Wendy Dannels, Rochester Institute of Technology/National Technical Institute for the Deaf

Credits

We would like to thank **Facebook Technologies, LLC** for donating Oculus Go devices to the ICCG participants who did not have a head-mounted VR display. This gave the ICCG a common platform for prototyping and experimentation.

We would also like to thank **Chris Hughes, University of Salford**, for building [an immersive captions prototyping tool](#) to support the ICCG's exploration of immersive captioning approaches, formats and styles.

Finally, we would like to thank **DSPA (Deaf Services of Palo Alto)** for providing the ASL interpreters, and for **Google** for paying for their services every meeting. We also want to thank the talented interpreters themselves, who have been critical for our work.

- Leah Bartell
- Peter Bolesh
- Lisa Kalsbeek
- Debra Kajiyama