2023
The Journal on Technology and Persons with Disabilities

Live Captions in Virtual Reality (VR)

Pranav Pidathala, Dawson Franz, James Waller, Raja Kushalnagar, Christian Vogler
Gallaudet University
pranav.pidathala@gallaudet.edu, dawson.franz@gallaudet.edu, james.waller@gallaudet.edu, raja.kushalnagar@gallaudet.edu, christian.vogler@gallaudet.edu

Abstract

Few VR applications and games implement captioning of speech and audio cues, which either inhibits or prevents access of their application by deaf or hard of hearing (DHH) users, new language learners, and other caption users. Additionally, little to no guidelines exist on how to implement live captioning on VR headsets and how it may differ from traditional television captioning. To help fill the void of information behind user preferences of different VR captioning styles, we conducted a study with eight DHH participants to test three caption movement behaviors (head-locked, lag, and appear-locked) while watching live-captioned, single-speaker presentations in VR. Participants answered a series of Likert scale and open-ended questions about their experience. Participants' preferences were split, but most participants reported feeling comfortable with using live captions in VR and enjoyed the experience. When participants ranked the caption behaviors, there was almost an equal divide between the three types tested. IPQ results indicated each behavior had similar immersion ratings, however participants found head-locked and lag captions more user-friendly than appear-locked captions. We suggest that participants may vary in caption preference depending on how they use captions, and that providing opportunities for caption customization is best.

Keywords

Virtual reality caption access; deaf and hard of hearing; automatic speech recognition accessibility; virtual reality subtitle usability.

Introduction

Virtual Reality (VR) as a mainstream industry is an extremely recent development, with today's technology only becoming widely available around six years ago with the public release of VR headsets like the Oculus Rift, HTC Vive, and others. In virtual reality, users have a 360º view of their surroundings: they can freely view everything in the virtual world by moving either their head or a controller. However, with novel technology comes novel accessibility challenges.

VR technology that uses spoken audio must provide that information in an alternative format to be accessible to Deaf and Hard of hearing (DHH) users – such as through captions. Captions typically appear at the bottom of a screen during the duration of the spoken audio, with standard guidelines such as a black background box, sans-serif white font, and a maximum of 3 lines with 30-40 characters per line.

However, virtual reality adds several new dimensions to consider. For example, when the person moves, should the captions move with them? Or do they stay fixed in place, anchored to some landmark in the virtual environment? A crucial element of virtual reality is its sense of immersion, so it is import ant to design captions that are well integrated and do not break the immersive experience. Previous work has identified several different approaches to caption positioning in VR, which can be divided into two approaches – 1) 'head-locked,' and 'world-locked' (Brown et al., 2017). In the first approach, characterized as 'head-locked' captions, captions are consistently locked to a fixed position in the user's view. If the user shifts their head, the captions will follow. The captions stay in the same position within the user's view; they are always visible. One version of head-locked captions has a lag: the captions follow the head movements with a slight delay (allowing for smoother movement).

In contrast, in the ‘world-locked’ approach, the captions are not locked to a fixed position in the person’s *view*, but rather are ‘locked to the environment’. Just like other virtual objects, these captions stay fixed in the environment as the person moves around. If the user turns away from the captions, the captions disappear from view; they do not follow the user’s head movements. Some are ‘speaker/source-locked’ – meaning they appear near the source of the audio (usually below the speaker), or near an object associated with the caption content. Others, like ‘appear’ captions, use a hybrid approach between head-locked and world-locked captions. Appear-locked captions are locked to the user’s initial head position, so that when the captions appear, they appear where the user is looking. But, unlike head-locked captions, they stay in place in the world if the user looks away.

Recent research has compared different captioning styles with various users. Rothe et al. (2018) compared hearing people’s preference for speaker-locked or head-locked captions in a VR setting with foreign language audio, and found no significant difference. Agulló (2019) tested head-locked captions without lag and ‘fixed’ (world-locked) captions with hearing and DHH participants. In the world-locked caption condition there were three different caption set ups spaced evenly around the user 120 degrees apart, such that wherever the user was looking there was likely at least one set of captions visible. They found that 82.5% of participants (DHH and hearing) strongly preferred the head-locked captions over the multiple world-locked captions. A follow-up study confirmed these findings, and found that participants rated the head-locked captions as more immersive (Agulló et al., 2020).

Research so far primarily centers on pre-recorded, pre-captioned 360^{o} video content. These approaches could be expanded upon to include other types of interactive VR media as well, such as in applications with live user-to-user interaction, but we found little to no mention of live VR captioning in past research. Social media, education, and virtual meeting platforms that support VR and feature active user interaction should implement live audio captioning as an accessibility option to keep their platform accessible for DHH users, yet few of the platforms we tested had this as an option. In our research on VR meeting platforms, we found that only a few supported live captioning, and those that did support it seemed to provide captioning as a paid service with little to no customizability options. We found one web-based immersive meeting platform in particular, Mozilla Hubs, which does not currently support captioning but is open source and can thus be modified by anyone. Based on this information, we decided to implement some of the caption types listed above within this platform to get a better idea of how practical they are to use and see how they would impact a DHH user's experience within VR.

Our study tested three different live VR caption types with DHH users: 1) head-locked captions without lag, 2) head-locked captions with lag, and 3) appear-locked captions in which the captions always appear in the user’s visual field but stay locked to the world thereafter. We exclusively recruited DHH users, who may have different preferences for VR captions than other users: many DHH users are frequent caption users, and do not necessarily rely on audio information to know if someone is speaking at that moment.

Discussion

For the present study, we tested three types of captioning behaviors in VR with Deaf and hard of hearing participants in the U.S.

Implementation

We coded prototypes of the three different captioning behaviors in Mozilla Hubs, an open-source, web-based meeting platform that supports VR headsets through the WebVR application interface. We implemented three different caption conditions, as follows:

‘Head-locked captions’ (no lag)

- Always visible.
- Moves along x-axis only. Locked on the y-axis to avoid blocking content: it cannot move up or down.

‘Lag captions’

- Usually always visible, except when the user moves quickly.
- Do not move up to down. When moving on the x-axis, the captions usually take approximately one second to ‘catch up’ to the user’s movement.

‘Appear-locked captions’

- The caption box is shown when the user is speaking. Once the speaker starts a new sentence, the captions go to where the user is looking.
- If the user is looking in the same spot continuously, the captions go under each other.

For consistency, all aspects of the captioning types other than their movement behavior are kept the same and use the same textbox object. We emulate existing TV closed captioning guidelines for our caption textbox: large white sans-serif text on a non-transparent black box with up to three lines of text (max. 128 characters). At all times, the caption box is facing the user.

We used Mozilla Hubs as our meeting platform, and Javascript as our primary coding language. The website platform was hosted on AWS. With our custom client and cloud server, we add onto Hubs’ existing codebase using the A-frame framework for 3D Objects to support captioning. We populated the world, or scene, with various 3D background elements, such as stairs, elevated platforms, and images to give the user an environment to visually inspect and explore as they traveled through. Four separate world templates, all with the same layout and 3D objects but with different images, were created within Spoke, Mozilla Hubs’ default scene editor, to go along with a corresponding educational script and presentation: the first was about the history of the Meta Quest VR device, the second was about the history of captioning devices, the third was about the Gallaudet Eleven's contributions to motion sickness research, and the fourth was a practice scene with placeholder images. A fourth non-moving caption type (permanently world-locked, located next to where the user enters the scene) was used only in our practice scene for acclimating the user to the VR environment and was not formally tested.

Captions were provided by Microsoft Azure's speech-to-text service for the live voice transcription through a W3C Speech Recognition polyfill so it works on all browsers. In our implementation on the Meta Quest 2 VR headset, Mozilla Hubs runs on the device through the Quest's default browser, where the speech recognition polyfill sends the audio directly from the microphone of the headset to Azure Cognitive Services. Captions are then sent back to the browser and formatted for readability. The goal is to mimic the use of VR for a live educational presentation or virtual tour.

Methodology

Before the experiment, the participants fill out a questionnaire asking about demographic information and experience with VR and captions. At the start of the experiment, the experimenter explains the instructions and controls. The participant puts on the VR headset, calibrates the equipment, and familiarizes themselves with a practice environment. In our study, participants could move around the scene using either Hubs' teleport mechanic (by holding down the left or right trigger button on the oculus controller and then releasing), or by using the left-joystick to "walk" around. They were informed about both movement types, with the note that joystick movement could cause motion sickness. Sometimes the captions stopped displaying text due to a bug, so participants were also informed how to mute and unmute the microphone on the Quest to fix the problem.

During the experiment, the participant experienced a ‘live educational presentation’, which lasted approximately five minutes. The experimenter gave a scripted presentation about different topics which are displayed in the VR environment, such that the participant could wander around the displays as the experimenter discussed each in turn. Automatically generated captions would appear in the participant’s VR environment while the presenter was talking.

We did this three times, once for each caption condition. Each participant saw each scene once, and caption ordering was counterbalanced with the scene ordering to remove bias. After each caption condition, the participant took off the VR headset and answered a series of 7-point Likert scale questions regarding their immersion in the VR world (using the IPQ questionnaire, see Regenbrecht and Schubert, 2002), as well as how easy and user-friendly the captions were to use. Participants also had the opportunity to provide open-ended feedback, and at the end of the tests they were able to rank the three caption behaviors from favorite to least favorite.

Eight DHH participants participated in our study, with ages ranging from 19 years old to 63 years old. Four were women, three were men, and one was non-binary. Six of the participants identified as deaf, and two as hard of hearing. Most of the participants used ASL as their main form of communication. Some people preferred ASL along with written English.

All participants used captions regularly (every day or most days) – most commonly for watching TV or streaming videos, but also for education, webinars, videoconferencing, and phone calls. Six participants use captions as their “primary method to understand information” and two use them as their secondary method and “to catch a few words I missed.” All participants had little to no previous experience with VR.

When asked if they used hearing devices, six participants did not currently use any, one uses a cochlear implant, and one uses a hearing aid. Five of the participants had glasses and wore them during the study, and three did not. None of the participants identified as DeafBlind.

Results

Participants were split in their caption preferences: three preferred head-locked captions, two preferred lag captions, and two preferred the appear-locked captions. In their comments, several participants said they liked that the head-locked and lag captions were consistently in the same place for reading the captions (both locked to head position) – though for others the captions were too low or too close, which was a source of frustration. Conversely, those who favored appear-locked captions liked that they moved relative to the head and did not stay in one position. Both participants who preferred the appear-locked captions use hearing devices and used captions as their secondary method of understanding information. Some participants noted frustrations about the caption types. For head-locked and lag captions, many participants mentioned the captions were too low and some participants wished the captions followed up and down with their head as well as side to side.

There were some complaints about the position of the appear-locked captions – too low, or too close. Also, one participant notes the captions sometimes “obscured the picture [they were] looking at”. Another participant experienced “caption drift” with appear-locked captions, where the captions kept moving downward when they turned their head to read the text.

Positive comments were mostly directed at head-locked and lag captions. Many participants found them clearer than the appear-locked captions – the captions were always readable without having to look back at a specific place. One participant also commented the consistency in head-locked captions allows them to also view the environment better: “The consistency in place and movement made it easier to read and focus on the environment.”.

Overall, any issues that arose with the captions did not seem to turn participants off from the technology as a whole – post-study, all but one participant reported feeling extremely (n=4) or moderately (n=3) comfortable in VR, and several commented that they enjoyed the experience and felt immersed.

When asked to rate the user friendliness of the caption type (1 to 7 scale: 1 = Worst imaginable, 4 = OK, 7 = Best), appear-locked captions scored the lowest (M = 3.9) while the scores were similar for head-lock captions (M = 5.2) and lag captions (M = 5.0). The three caption types all scored between 4 and 5 on immersion (scores calculated from the IPQ, on a scale from 1 to 7).

General Discussion

DHH users are not uniform in their preferences, and it seems a flexible, customizable approach is needed. Not only did participants favor different caption types, but some suggested alternative approaches that allowed even more control over the captions, such as being able to drag and pin the captions to any position in the environment: “I would like to be able to either ‘pin’ the captions in a specific spatial location in my vision field (for head-lock approach). For the appear-locked captions, it needs to be [interact-able] so I can move it next to the pictures or the other [visual information] of interest.”

Caption preferences appear to vary based on how people use captions and the VR technology. For example, people who rely on captions as a secondary source of information may like to be able to ‘look away’ from captions, but others who rely primarily on captions prefer head-locked captions that are always visible.

Caption preferences may also depend on the way that users move around their virtual environment. For example, we observed that one participant complained about appear-locked captions being ‘too close’ – they also were one of the few used the teleport function to move around. It is possible that the teleport function offers less fine control of positioning than the joystick, so it is more frustrating to have to re-position relative to the appear-locked captions for those using teleport movement. Future research should account for the possibility that the ideal caption set-up depends on how users navigate the virtual environment – whether through joystick, teleportation, or physical movement (not tested in this study): i.e., walking around the physical environment to navigate the virtual one.

Conclusions

Compared to TV closed captions, live VR captions may require additional customization options and/or should exist as an interactable object, since the expanded view of VR may not allow for a one-size-fit-all approach. When developing for VR, it is a good idea to give users multiple accessibility options, along with the tools to customize them to what's best for them.

This study focused on the use of live captions in virtual reality rather than pre-recorded captions. While principles developed for live captions in VR may be largely applicable to pre-recorded captions, presentation of pre-recorded captions can also be tailored in ways that cannot be easily done with live captions – such sophisticated source-locked caption options that we did not test here. Additionally, pre-recorded captions could also include information such as emotion and tone of voice, or environmental sounds. How best to integrate this information into VR applications is still an open question.

Nonetheless, current immersive captioning work is promising, and shows that captions do not necessarily detract from the VR experience – many DHH participants are excited to participate in VR technology, and it is important that these technologies are accessible to all.

Acknowledgments

This work was supported by NSF #2150429 (REU AICT).

Works Cited

Agulló, Belén, and Anna Matamala. "Subtitling for the deaf and hard-of-hearing in immersive environments: results from a focus group." *The Journal of Specialised Translation* 32 (2019): 217-235.

Agulló, Belén, and Anna Matamala. "Subtitles in virtual reality: guidelines for the integration of subtitles in 360º content." Íkala, revista de lenguaje y cultura 25.3 (2020): 643-661.

Brown, Andy, et al. "Subtitles in 360-degree Video." *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*. 2017.

Regenbrecht, Holger, and Thomas Schubert. "Real and illusory interactions enhance presence in virtual environments." *Presence: Teleoperators & Virtual Environments* 11.4 (2002): 425-434.

Rothe, Sylvia, Tobias Höllerer, and Heinrich Hußmann. "CVR-analyzer: a tool for analyzing cinematic virtual reality viewing patterns." *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia* (2018): 127-137.