Mixed Reality Speaker Identification as an Accessibility Tool for Deaf and Hard of Hearing Users

Abraham Glasser
Golisano College of Computing and Information Sciences
Rochester Institute of Technology,
Rochester, NY, USA,
atg2036@rit.edu

Edward Riley
Golisano College of Computing and Information Sciences
Rochester Institute of Technology,
Rochester, NY, USA,
emr9018@rit.edu

Kaitlyn Weeks
Department of Psychology
Gallaudet University,
Washington, DC, USA,
kaitlyn.weeks@gallaudet.edu

Raja Kushalnagar
Department of Science, Technology and Mathematics,
Gallaudet University Washington, DC, USA,
raja.kushalnagar@gallaudet.edu

ABSTRACT

People who are Deaf or Hard of Hearing (DHH) benefit from text captioning to understand audio, yet captions alone are often insufficient for the complex environment of a panel presentation, with rapid and unpredictable turn-taking among multiple speakers. It is challenging and tiring for DHH individuals to view captioned panel presentations, leading to feelings of misunderstanding and exclusion. In this work, we investigate the potential of Mixed Reality (MR) head-mounted displays for providing captioning with visual cues to indicate which person on the panel is speaking. For consistency in our experimental study, we simulate a panel presentation in virtual reality (VR) with various types of MR visual cues; in a study with 18 DHH participants, visual cues made it easier to identify speakers.

CCS CONCEPTS

• Human-centered computing~Empirical studies in accessibility

KEYWORDS

Deaf and Hard of Hearing, Speaker Identification, Mixed Reality

ACM Reference format:

Abraham Glasser, Edward Riley, Kaitlyn Weeks and Raja Kushalnagar. 2019. Mixed Reality Speaker Identification as an Accessibility Tool for Deaf and Hard of Hearing Users. In *Proceedings of VRST '19: 25th ACM Symposium on Virtual Reality Software and Technology (VRST '19), November 12-15, 2019, Parramatta, NSW, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3359996.3364720

1 INTRODUCTION AND RELATED WORK

DHH people benefit from text captioning to understand the audio component of video or live events. However, there are complex environments in which captions are insufficient for these users. During panel presentations, with multiple speakers having a live, unscripted discussion about some topic, it can be challenging for captioning (whether provided by a human transcriptionist or an automatic speech recognition (ASR) service) to clearly convey who is speaking. Unlike pre-recorded video of a spoken conversation, e.g. in a television program, the turn-taking in a live discussion may be rapid and unpredictable, and there are no camera view transitions to indicate which individual is currently speaking. DHH individuals report that it is tiring and distracting to view captioned panel presentations, which require them to look back and forth between speakers and captions [Kushalnagar et al., 2017].

With recent developments in Mixed Reality (MR) and sound recognition, it's becoming more feasible to use MR as a personalized accessibility tool for DHH individuals. It is possible to use sound detection technology to identify where the sound is coming from in a 3D audio setting. MR technology can be used to help a user identify who is speaking in a panel discussion. Also, developments in ASR could be used to display captions in a MR head-mounted display, or the text could be streamed from an external source, such as real-time stenography.

While some researchers have discussed guidelines for captioning in 360-degree video [Brown et al., 2017] for single speakers, there has been limited prior empirical research on captioning of content with multiple speakers, and there is a lack of such research that includes Deaf or Hard of Hearing (DHH) viewers. Prior work has investigated alternative methods of indicating who is speaking during captioning of live events, in Virtual Reality (VR) and MR:

In [Rothe et al., 2018], Rothe, Tran, and Hußmann experimented with static subtitles, captions that stayed in one place, and "dynamic" subtitles, subtitles that appear near the

VRST '19, November 12-15, 2019, Parramatta, NSW, Australia

ACM ISBN 978-1-4503-7001-1/19/11.
https://doi.org/10.1145/3359996.3364720

speaker, for cinematic VR. The authors included only one DHH participant in their study and indicated that additional research is needed for DHH users. Participants felt that dynamic subtitles forced them to look at the speaker, but they could not predict where subtitles would be next.

Similar to [Kurzhals et al., 2017], researchers in [Kushalnagar et al., 2017] investigated captions that moved with the speaker through a room. In [Kushalnagar et al., 2017], Kushalnagar et al. evaluated two different methods of displaying captions with speaker-identification and compared these to traditional captioning in a classroom that was instrumented with projectors that could display captions in various surfaces in the room, in an MR-like experimental setup. One was the "pointing" method, which puts the captions in a fixed spot centered above the speakers, and adds an indicator that points to the current speaker. The other method, "pop-up", puts the captions directly above the current speaker real-time. In their study, participants did not like how the captions would disappear and reappear in a different place in the pop-up method. Captions are a text representation of speech, and is designed to be static, so when it moved quickly between the speakers, it became difficult to follow.

2 METHODOLOGY

Given the limited prior research on MR captioning solutions for DHH users, especially for complex panel environments, we evaluate three different speaker-identifying visual cues added on top of traditional captions for a panel talk. Since there is not a working MR standalone system that can add speaker-identifying cues along with captions, we used computer-generated imagery to create a VR simulation of captions and speaker-identification cues being added to a panel talk. The use of VR to play a pre-recorded presentation enabled us to investigate MR display conditions while controlling for variations in the presentation itself. A Vuze XR Camera was used to record the live panel talk simulation. Then, Adobe After Effects was used to add captions and speaker-identification cues. A Google Pixel phone was used along with a Daydream VR headset for viewing the videos in VR.

Four different visual cue conditions were evaluated in this study. For the "caption standalone" condition, traditional captions were added to the video, appearing below the panelists (on the table), in white Arial font on an opaque black background. For the other three conditions, the captions are the same, but a speaker-identifying cue is added. In "lightbulb," a 2D image of a lightbulb appears above each person's head and is lighted if that person is speaking. In "glow," a yellow ellipsis with an orange circle in the center appears flat on top of the table, between the speakers and above the captions. In "pointing," a 2D hand appears above the speakers, pointing to their heads with its index finger.

Four different panel presentation scripts were simulated, and each visual cue condition was produced for each script for a total of sixteen combinations. Each participant watched four videos, one for each script and condition. The video order was counterbalanced throughout the participants to eliminate the possibility of the content having an impact. Also, all the videos were played without audio to eliminate the possibility of using the voices of the speakers as a speaker-identifying aid. A snapshot from each of the four conditions is shown in Figure 1.

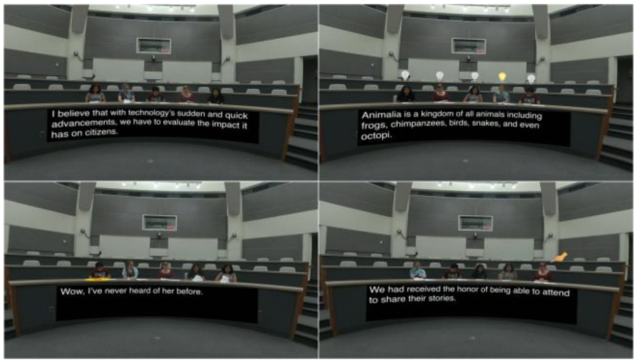
I believe that with technology's sudden and quick advancements, we have to evaluate the impact it has on citizens.
Animalia is a kingdom of all animals including frogs, chimpanzees, birds, snakes, and even octopi.
Wow, I've never heard of her before.
We had received the honor of being able to attend to share their stories.

Figure 1: Snapshots from the 4 different visual cue conditions

3 RESULTS AND DISCUSSION

18 DHH participants completed this study. 8 participants were 20-29 years old, 4 were 30-39, and 4 were 40 or above. 15 identified as deaf, and 3 identified as Hard-of-Hearing. During a pre-experiment questionnaire, participants were asked questions about their experience with closed captioning and technology usage. All participants said that they use closed captioning regularly, and 12 out of 18 said they had experience with VR technology. When asked if they "often have trouble with identifying the speaker in multi-person environments", only 2 out of 18 participants said no.

A Bonferroni correction was applied to correct for multiple comparisons. The lightbulb, glow, and pointing methods were significantly easier than the caption standalone (t-test, $p < .001$) in identifying the speaker. No visual cue was significantly better than the other in identifying the speaker. A data summary boxplot of responses for this question is shown on left side of Figure 2.

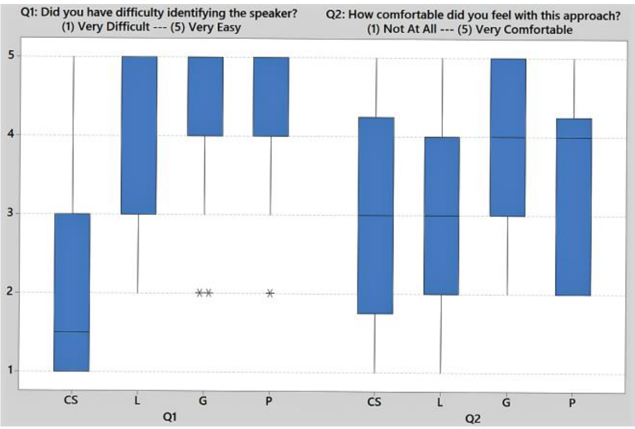

Figure 2: Boxplot data summary of participant responses to the two 5-point Likert-scaled questions

When asked how comfortable each visual cue was, participants preferred the glow method to the lightbulb method (t-test, p < .01). After a Bonferroni correction was applied, none of the other pairwise combinations yielded significant results. A data summary boxplot of responses for this question is on the right of Figure 2. In a post-experiment questionnaire, when participants were asked if they "hope to see closed captioning in virtual reality environments in the future", 16 participants said yes. When asked "do you hope to see speaker-identification in virtual reality environments in the future?" none of the participants said no.

REFERENCES

A. Brown, J. Turner, J. Patterson, A. Schmitz, M. Armstrong, M. Glancy. 2017. Subtitles in 360-degree Video. In 2017 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17 Adjunct). ACM, New York, NY, USA, 3-8. DOI: https://doi.org/10.1145/3084289.3089915

K. Kurzhals, E. Cetinkaya, Y. Hu, W. Wang, D. Weiskopf. 2017. Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles. In Proc. 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, NY, NY, USA, 6559-6568. DOI: https://doi.org/10.1145/3025453.3025772

R. Kushalnagar, G. Behm, K. Wolfe, P. Yeung, B. Dingman, S. Ali, A. Glasser, C. Ryan. 2017. RTTD-ID: Tracked Captions with Multiple Speakers for Deaf Students. In Proc. 2018 ASEE Annual Conference & Exposition (ASEE '18). Salt Lake City, UT, USA. https://peer.asee.org/30945

S. Rothe, K. Tran, H. Hußmann. 2018. Dynamic Subtitles in Cinematic Virtual Reality. In Proc. 2018 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '18). ACM, New York, NY, USA, 209-214. DOI: https://doi.org/10.1145/3210825.3213556