# Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study

**Figure 1a.** Prototype 1*: AR Windows* displays captions in a HoloLens web browser window. Caption windows can be placed close to speakers or visual materials, such as lecture slides in 3D space.


**Figure 1b.** Prototype 2: *AR Subtitles* displays one caption window that is placed at a fixed distance in front of the user and moves with user's head.

**Dhruv Jain**
University of Washington
Seattle, WA 98195, USA
djain@cs.washington.edu

**Raja Kushalnagar**
Gallaudet University
Washington, DC 20002, USA
raja.kushalnagar@gallaudet.edu

**Bonnie Chinh**
University of Washington
Seattle, WA 98195, USA
bchinh@uw.edu

**Jon Froehlich**
University of Washington
Seattle, WA 98195, USA
jonf@cs.uw.edu

**Leah Findlater**
University of Washington
Seattle, WA 98195, USA
leahkf@uw.edu


*DIS'18 Companion, June 9–13, 2018, Hong Kong*

ACM ISBN 978-1-4503-5631-2/18/06.
https://doi.org/10.1145/3197391.3205404.

## Abstract
We explore an augmented reality (AR) approach to real-time captioning for people who are deaf and hard of hearing. In contrast to traditional captioning, which uses an external, fixed display (*e.g.,* laptop or large screen), our approach allows users to manipulate the shape, number and placement of captions in 3D space. We discuss design factors, describe two early prototypes, and report on an autoethnographic evaluation of the prototypes. Preliminary findings suggest that, compared to traditional laptop-based captions, HMD captioning may increase glanceability, improve visual contact with speakers, and support access to other visual information (*e.g.,* slides).

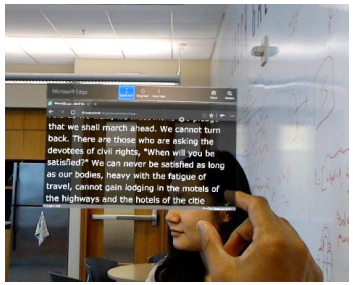
## Author Keywords
Deaf; augmented reality; real-time captioning.

## ACM Classification Keywords
Human-centered computing---Accessibility technologies

## Introduction
Real-time captioning converts aural speech to visual text for individuals who are deaf and hard of hearing (DHH), particularly in stationary contexts such as classrooms and meetings [9]. Captions are

**(a)**

**(b)**




**(c)**

**Figure 2.** Illustrative figures showing: (a) 1:1 meeting with AR Subtitles, (b) group meeting with AR Windows, (c) our lead author positioning a caption window close to a speaker with AR Windows.

generated via a human transcriptionist or an automatic speech recognition (ASR) engine, then are typically shown on a private laptop or a shared large screen.

While useful, this traditional setup has limitations. First, the user's visual attention is split between the screen, conversational partners, and other visual materials such as lecture slides [6]. Second, this attention split may be exacerbated when speakers move about the room but the caption display is fixed. Third, the user may not necessarily want the captions to be viewable beyond their immediate conversation partners for privacy (*e.g.,* intimate or financial information).

We are exploring an alternative approach: displaying real-time captions in 3D space on a head-mounted display (HMD). Users can privately view captions and can manipulate the shape, number, and placement of caption windows. Potential advantages include reduced visual dispersion, increased glanceability, and increased mobility. While prior research has proposed the idea of showing real-time captions on an HMD [5,7], no work has explored caption design specifically or performed user evaluations. Further, prior work has not examined positioning captions in 3D space using AR.

In this paper, we outline design factors for HMD-based captioning, present two early prototypes on a Microsoft HoloLens, and report on a preliminary autoethnographic study where the lead author, Jain, who is hard of hearing, used the prototypes in 10 classes and group meetings over 6 weeks. In comparison to laptop-based captions, Jain felt increased visual contact with other speakers and easier access to captions and other visual information. However, disadvantages included the size of the device and mid-air gestures needed to control it.

In summary, our contributions are: (1) a new 3D augmented reality approach to provide real-time captioning for DHH users, (2) empirical results from an autoethnographic evaluation, and (3) key design factors and recommendations for future work.

## Initial Prototypes

Informed by prior work [1,5,8] and our own experiences as persons with hearing loss, we synthesized 13 design considerations for HMD-based AR captioning related to caption rendering (Figure 3) and contexts of use (Figure 4). To begin exploring this design space, we developed two real-time prototypes with the Microsoft HoloLens. Both prototypes use *Streamtext* [10], a remote online captioning software, to receive captions from a professional transcriptionist.

**Prototype 1: AR Windows** displays captions in a HoloLens web browser window (Figure 1a). Using built-in HoloLens hand gestures, the user can duplicate, resize, and position this window in physical space (*e.g.,* one window above each speaker). Similar to traditional captioning, this prototype can display multiple lines of conversation. Captions scroll up and disappear at the top of the window as new captions are generated.

**Prototype 2: AR Subtitles** displays one caption line that is placed at a fixed distance in front of the user and moves with the user's head (Figure 1b). Similar to video captions, this prototype only shows the most recent generated captions (60 characters). With no option to resize or position the captions, AR Subtitles requires less user control than AR Windows.

## Evaluation

To gain preliminary insight into the benefits and limitations of AR captioning on HMDs and to help inform

**Caption placement:** how are the captions positioned in 3D space and do they automatically move (*e.g.,* to track the speaker).

**Caption length and size:** how many words and lines of text are presented and at what size?

**Transcription fidelity:** are the words transcribed verbatim or is summarization used (*e.g.,* topic summarization, nouns-only).

**Wearer's voice:** is the wearer's voice transcribed and, if so, how is it visually represented?

**Contextual information:** what level of contextual information is supplied about speaker (*e.g.,* speaker names, speaker tone, loudness).

**Non-speech information:** what non-speech sounds are important and how should they be represented (*e.g.,* dog barking, door opening).

**Error handling:** how are errors in transcriptions represented and potentially fixed?

**Figure 3.** UI design considerations for AR captioning on an HMD.

the design of our future prototypes, we conducted an autoethnography. Autoethnography includes a reflexive and analytic account of personal experiences, and connects those experiences to wider social and cultural groups [3,4]. The DHH lead author, Jain, adopted the role of researcher-participant and documented his experience with our prototypes over a 45-day period.

*Method*

Jain is a 26-year-old graduate student with severe-to-profound deafness. He uses bi-lateral hearing aids, can speak and speechread well, and relies on real-time captioning for classes and meetings. Jain used the two prototypes in 10 instances: 3 group meetings and 3 lectures with AR Windows, and 4 group meetings with AR Subtitles. The total usage time was ~7 hours. Jain also used laptop-based captions for ~25 instances in the same academic quarter before the study began.

Jain documented his experiences of each HMD session in the same day using a semi-structured approach, describing the context of use, his emotions, events that stood out, and his general experience. His notes contain a total of 7,053 words. We used a thematic approach [2] to analyze the data based on both inductive and deductive themes. The first author led this analysis, guided by multiple discussions with other team members as analysis iteration occurred.

*Findings*

For overall preference, Jain was initially split between the laptop and HMD captions, especially due to the discomfort of wearing the HoloLens. By the fourth session, however, he preferred the HMD. We highlight differences between our two designs and between the HMD and laptop. Quotes are from Jain's notes.

**Glanceability.** Jain felt he could switch more quickly between attending to captions and attending to the speaker when using the HMD than with laptop. For example, after the fourth session (3 hours of total use), he wrote: *"HoloLens was better than laptop since I could see [both] [speakers'] lips and, the captions…".* Consequently, Jain could "*make more face-to-face contact with [speakers].*"

Using HoloLens, captions were also closer to visual materials such as lecture slides, which increased access to information, particularly in cases where the speaker pointed to visual aids: *"[While using AR Windows,] I could see his hands pointing [at] various math equations on the screen, as he said 'derivative of this* [pointing at slides] *leads to this.', which is hard to follow with captions on a laptop."*

**Caption Placement.** Jain preferred to overlay captions on or below the speaker's face using hand gestures (AR Windows) or head orientation (AR Subtitles). Additionally, for AR Windows, Jain placed one caption window for each speaker in 3D space (Figure 1a) but used only one window for speakers who were close to each other. When looking at lecture slides, Jain positioned the captions directly below the slides to avoid visually obstructing the slide material.

Comparing our two prototypes, Jain preferred AR Windows during group meetings with multiple, seated speakers because he could set up a caption window for each speaker. He noted that "*though the captions were not always in my view [like AR Subtitles], I was able to speechread speakers while seeing their captions with [close to] them.*" Instead, with AR Subtitles, *"captions felt disconnected from speakers as [captions] appeared at a fixed distance [and angle] from me."* For multiple moving speakers, however, he preferred AR Subtitles, which allowed captions to remain in view when speakers moved. He speculated

**Visibility of information:** how visible are the captions? *E.g.,* private view (*e.g.,* viewable only by DHH user), or public view (*e.g.,* a large projected display in a classroom).

**Conversation group size:** how many people are involved in the conversation? *E.g.,* 1:1, medium-sized group (*e.g.,* around a dinner table), or a larger set (*e.g.,* a lecture).

**Physical activity:** what physical activity are the conversation partners involved in? *E.g.,* all people sitting, main speaker walking while others sitting (*e.g.,* a lecture) or is everyone moving (*e.g.,* walking).

**Topic sensitivity and interaction:** how may user needs change across conversation topics? *E.g.,* confidential information (*e.g.,* finances), or high-emotions (*e.g.,* intimate conversations),

**Expected length of interaction:** what is the expected length of the conversation?

**Relationship with conversational partners:** how may user needs change depending on their relationship and familiarity with conversation partners?

**Figure 4.** Context of use considerations for AR captioning on an HMD.

that he would prefer AR Windows overall though if *"captions could automatically move with the speakers."* For only a single speaker (*e.g.,* 1:1 meeting, non-interactive lecture), Jain preferred AR Subtitles, since he could position himself so that captions were close to the speaker and he was *"able to read captions when I moved my head"* (*e.g., "for taking notes",* Figure 2a).

**Social Aspects.** Jain reported feeling noticeable and socially awkward wearing the HoloLens, because it is an unusual device and because AR Windows required mid-air gestures to configure captions. For example, after a lecture with AR Windows, *"I think the hand gestures movements […] distracted students. [The instructor] deduced what was going on in HoloLens. But, he told me, if he didn't know, he would be like 'What are you doing in my class? Playing a game while I teach?'"*. One conversation partner also commented that the HoloLens partially obstructed Jain's face, making it difficult to have a natural face-to-face interaction.

**Design Limitations.** As the current version of HoloLens is heavy and conducts heat, Jain could sustain only 42 minutes of continuous use on average. Also, the display screen is not fully transparent, which impacted Jain's ability to see in a dimly lit room and in one instance made speechreading difficult. Jain also had to switch to laptop captions after 15 mins of one 1:1 meeting because the HMD made it difficult to read the other person's shared computer screen.

We also note some UI problems with AR Subtitles, where a single caption line was placed at a fixed distance from the wearer's head, and moved horizontally and vertically with the head. If the speaker was at a much greater distance from the wearer than the captions were, glanceability was reduced: *"it was hard to focus on both speaker and captions."* Conversely, speakers sometimes occluded the captions if they came too close to the wearer. Jain felt that these issues would be addressed if *"the [caption] window could automatically align [with speakers] in the depth space."* Finally, a single line of captioning was often not enough to understand well (*e.g.*, when captioning errors occurred), and in some cases Jain used laptop captions in parallel so he could review the caption history.

## Discussion and Conclusion

Our findings from this preliminary autoethnographic evaluation suggest that a real-time HMD-based captioning approach is promising, particularly in increasing visual contact with speakers and access to other visual material (*e.g.*, lecture slides). At the same time, we identified several improvements that could be made to our designs, such as auto-tracking speakers and providing less obtrusive control over the interface (*e.g.*, to resize and move captions). The HoloLens allowed us to quickly prototype our 3D caption interfaces, but an ideal form factor would be lighter, smaller, and would not obscure the wearer's eyes.

Our immediate plans include refining the designs as described above and conducting a larger user study to generalize our findings beyond one specific user while employing the prototypes as design probes. Our future work also includes creating more sophisticated prototypes that draw on the design considerations in Figure 3, such as providing information on speaker names or tones, and contexts of use from Figure 4, such as exploring HMD captioning for mobile contexts.

## Acknowledgment

We thank Liang He and Shuxu Tian. This research was funded by a Google Faculty Research Award.

## References


1. John W Du Bois and Stephan Schuetze-Coburn. 1993. Representing hierarchy: Constituent structure for discourse databases. *Talking data: Transcription and coding in discourse research*: 221–259.
2. Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2: 77–101.
3. Sally Jo Cunningham and Matt Jones. 2005. Autoethnography: a tool for practice and education. *Proceedings of the 6th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: making CHI natural*, 1–8.
4. Carolyn Ellis, Tony E Adams, and Arthur P Bochner. 2011. Autoethnography: an overview. *Historical Social Research/Historische Sozialforschung*: 273–290.
5. Dhruv LA, Leah Findlater, Christian Volger, Dmitry Zotkin, Ramani Duraiswami, and Jon Froehlich. 2015. Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 241--250.
6. Marc Marschark, Jeff B Pelz, Carol Convertino, Patricia Sapere, Mary Ellen Arndt, and Rosemarie Seewagen. 2005. Classroom interpreting and visual information processing in mainstream education for deaf students: Live or Memorex®? *American Educational Research Journal* 42, 4: 727–761.
7. Kazuki Suemitsu, Keiichi Zempo, Koichi Mizutani, and Naoto Wakatsuki. 2015. Caption support system for complementary dialogical information using see-through head mounted display. *Consumer Electronics (GCCE), 2015 IEEE 4th Global Conference on*, 368–371.
8. Mike Wald and Keith Bain. 2007. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society* 6, 4: 435–447.
9. What is real-time captioning? | UW DO-IT. https://www.washington.edu/doit/what-real-time-captioning.
10. StreamText.Net. http://www.streamtext.net/.