ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #35170

ScribeAR: A New Take on Augmented-Reality Captioning for Inclusive Education Access

Prof. Lawrence Angrave, University of Illinois at Urbana - Champaign

Lawrence Angrave is an award winning Fellow and Teaching Professor at the department of computer science at the University of Illinois at Urbana-Champaign (UIUC). His interests include (but are not limited to) joyful teaching, empirically-sound educational research, campus and online courses, computer science, engaging underrepresented students, improving accessibility and creating novel methods that encourage new learning opportunities and foster vibrant learning communities.

Mr. Colin P. Lualdi, University of Illinois at Urbana-Champaign
Mona Jawad
Timur Javid

ScribeAR: Design and Use of Augmented-Reality Captioning for Inclusive Education Access

Mona Jawad[1], Timur Javid[2], Colin P. Lualdi[3], Lawrence Angrave[2]
Department of Bioengineering, University of Illinois at Urbana-Champaign[1]
Department of Computer Science, University of Illinois at Urbana-Champaign[2]
Department of Physics, University of Illinois at Urbana-Champaign[3]

Abstract

Existing real-time captioning technology has greatly improved educational access for deaf and hard-of-hearing (DHH) students. Yet delivering captions on a computer display or slides causes a delay in information processing, as the students must switch their attention between the lecturer and captions. To address this challenge, we report on the iterative design process of ScribeAR, a lightweight real-time captioning platform capable of streaming captions on an augmented-reality (AR) headset or laptop display. This system could decrease the information gap and improve accessibility for DHH students both classroom and spontaneous educational settings. We discuss plans for evaluating the viability of our system in actual educational settings as well as expansion possibilities.

Introduction

Real-time captioning is often an effective communication access tool for the deaf and hard of hearing (DHH) by converting inaccessible speech into accessible text. Consequently, captioning is prevalent in many settings, especially in education. However, traditional captioning delivery mechanisms (e.g., streaming captions on a computer display) requires a DHH student to split their attention between, for example, the lecturer, slides, and captions. This causes delayed information processing and increased cognitive load for DHH students (Kushalnagar, 2014).

DHH students relying on real-time captions have significantly lower performance in the classroom in comparison to their hearing peers (Marschark, 2006). Numerous efforts to address this challenge have been made over the past two decades. Of particular interest is the use of augmented-reality (AR) headsets to deliver the captions directly into the user's line-of-sight as opposed to a separate display set to the side (e.g., Jain, 2018). Along these lines, AR headsets that project American Sign Language (ASL) interpreters onto the lens have also been explored with promising results (e.g., Miller, 2017), including commercialization (SignGlasses, www.signglasses.com). However, these systems are designed for use in controlled environments (e.g., the classroom) with the captioning or interpreting service paid for by an institutional accommodations office. However, in post-secondary settings a significant portion of a student's educational experience takes place outside of the classroom via interactions with instructors and peers (e.g., office hours or study groups). The effective transplanting of traditional captioning approaches to these spontaneous scenarios is difficult: the arduous task of scheduling and setting up a dedicated captioning display makes it impractical. Similarly, requiring a constantly on-call human captioner or ASL interpreter is often financially unscalable, both for the institution and student.

Though the motivation for this paper is to report on the design and implementation of an AR captioning system, we note that the system has benefits that extend beyond STEM education as well as the DHH population. For example, performing arts centers could provide ScribeAR's augmented reality experience to provide an inclusive experience for DHH patrons who attend concerts, plays, and community events. Similarly, captioning benefits students with learning challenges such as attention deficit disorder (ADD) or attention deficit hyperactivity disorder (ADHD), because the current and recent caption lines provide a short-term context of the educational content being discussed. As is the case for DHH users, this facilities re-engagement with the current educational context after a temporary distraction.

Design Objectives

The design team, led by a DHH graduate student, identified six initial design objectives to address the recognized need. They are summarized here.

1. **Practical:** Compatible with portable AR hardware that is also sufficiently comfortable to wear for long periods (multiple hours).
2. **Independent:** Usable when AR glasses are unavailable.
3. **Flexible:** Adaptable to student's changing captioning needs and priorities in different educational contexts; allow the user to choose between desired transcription cost and accuracy. For example, a student would utilize high-quality human captioning during a lecture (where tolerance for transcription errors is low). After class, the student could switch to less-accurate automated captioning to discuss lecture details with a fellow student (where the tolerance for errors is higher due to opportunities for clarifications inherent in a back-and-forth dialogue).
4. **Multi-Context:** Able to adapt the caption presentation to different visual contexts. The presentation should be configurable to allow reduction of interference between the live scene and overlayed caption text.
5. **Archival:** Allow transcription to be archived or saved for future review.
6. **Depth:** Provide additional non-textual visual cues of auditory information.

Development Work

ScribeAR has been under development since 2019 by College of Engineering undergraduate students at the University of Illinois at Urbana-Champaign under the guidance of a Computer Science faculty member and a Deaf Physics graduate student. The implementation work has been guided by the initial design objectives described above. Naturally, reevaluation stages of the development cycle revealed additional design objectives, described below.

Figure 1 presents the overall conceptual system that follows the data and information processing from the original human speaker to the multimodal representations available to the user.

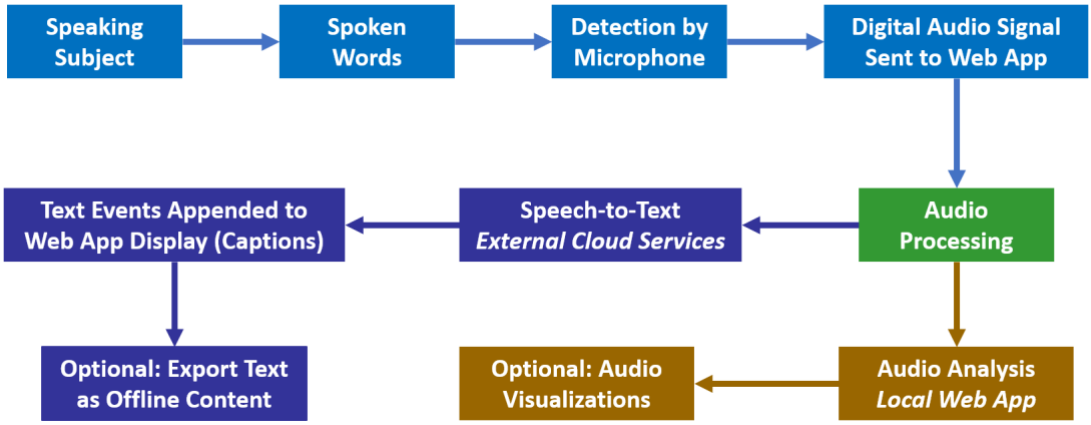

Figure 1: System diagram showing information flow in ScribeAR.

The codebase is publicly available on GitHub[1] to encourage community adoption and contributions to accessibility technologies.

Figure 2a presents the default ScribeAR main page as it appears when first loaded in a web browser. Selecting the hamburger icon (≡) on the top left corner of the screen opens the Options menu (Figure 2b), which allows additional customization of ScribeAR's appearance and functionality.

(a) (b)

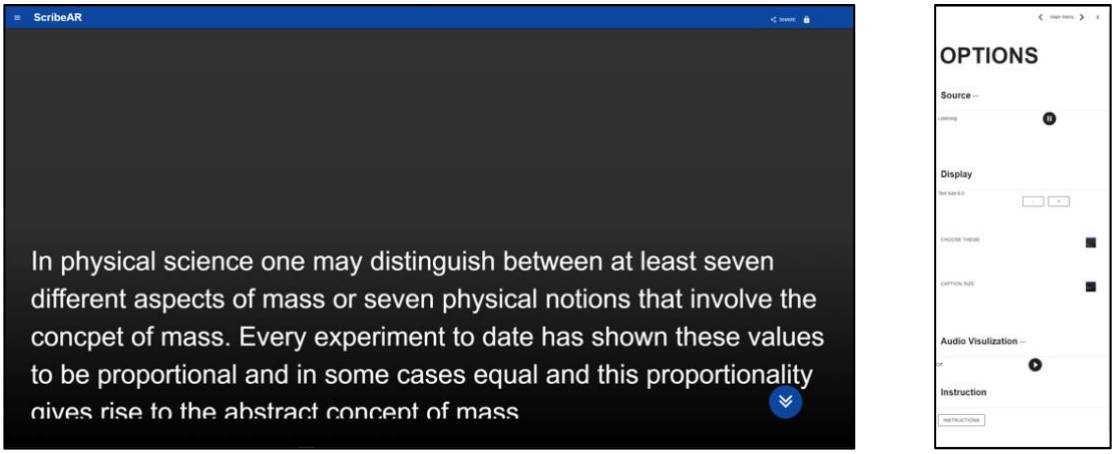

Figure 2: (a) Main screen of ScribeAR with default options. The text shown was spoken by a speaker reading from the Wikipedia article on Mass (en.wikipedia.org/wiki/Mass) and automatically transcribed by the Microsoft Azure Speech-to-Text service. (b) ScribeAR options menu.

[1] The ScribeAR repository is located at www.github.com/scribear/ScribeAR.github.io.

Design Choices:

Criteria 1. Compatible with portable AR hardware that is also sufficiently comfortable to wear for long periods (multiple hours).

The ScribeAR platform is intended to be compatible with a wide variety of computer hardware and augmented reality headsets. It is straightforward to develop ScribeAR as a traditional captioning delivery mechanism on a computer display due to the prevalence of such systems in modern-day society. However, AR development is more challenging due to the nascent nature of AR hardware.

The market currently offers a variety of sophisticated off-the-shelf AR headsets, such as the Microsoft HoloLens 2 (released 2019) featuring built-in 3D tracking capable of projecting virtual 3D objects embedded into the visual real-world experience. The HoloLens 2 was not selected because it was expensive (>$1000) and cumbersome to wear for long periods, especially when weighing 0.6 kg (1.3 lbs).

A better alternative was the repurposing of drone first-person-view AR Glasses offered by Epson as a part of the Moverio product line[2] (Figure 3). They are relatively lightweight at approximately 0.1 kg (3.5 oz) and are as comfortable as wearing a normal pair of glasses, albeit with a cord attached. Furthermore, they function as simple external monitors, meaning they can act as a secondary display or mirror the display of attached computers or mobile devices. All power and display information can be delivered via a single USB-C cord. These headsets thus distinguish themselves from systems such as the HoloLens 2 by having no onboard computing hardware, resulting in significantly reduced their size, weight, and power (SWaP) characteristics. The product lineup is also more affordable with entry-level models costing approximately $500.

Consequently, all development relating to the AR aspect of ScribeAR was performed on Moverio hardware. However, every effort was made to avoid hardware-specific design decisions to maintain ScribeAR's compatibility with general AR hardware as a part of our commitment to delivering a flexible platform.

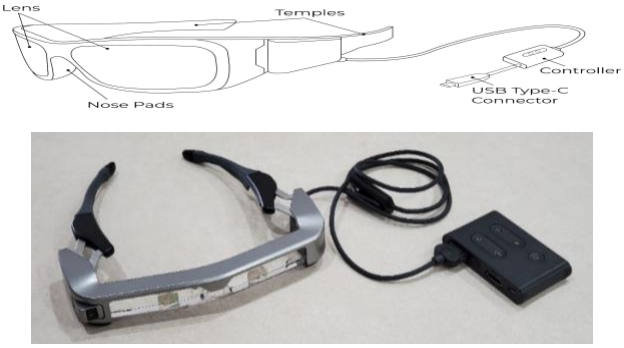

Figure 3. An illustration of the Epson Moverio BT-30C (USB-C input) and picture of the BT-35E (USB-C and HDMI) models. Nose pads allow the headset to be worn in addition to lens-correction glasses.

[2] Full details regarding Epson's Moverio products may be found at www.epson.com/For-Work/Wearables/Smart-Glasses/c/w420.

Criteria 2. Usable when AR glasses are unavailable.

Using modern HTML5 application programming interfaces (APIs), today's web browser environment can support computational requirements that traditionally would have necessitated writing custom applications targeted at multiple operating systems running on desktop or mobile environments (Windows, OSX, Linux, Android, IOS). By designing ScribeAR as a web application, ScribeAR was able to run in multiple computing and display environments including Windows and OSX laptops, mobile phones, and tablets, and present the captions on the device's display or in an AR headset if attached.

React and Redux JavaScript libraries were chosen as the main JavaScript components to build the application. These popular libraries supported development of state-of-the-art interactive web applications with sophisticated user interfaces and enabled a maintainable event-based component software design.

Another advantage of designing ScribeAR as a web application was that the system could be loaded immediately without any installation process or special administrative permissions. Upgrades to the platform were easily deployed via publishing the latest version online.

Criteria 3. Adaptable to student's changing captioning needs and priorities.

Automatic speech-to-text transcription is the central feature of ScribeAR. Two options were implemented:

- **Web Speech API[3]:** This HTML5 API provided browser-based automatic speech recognition. Support among popular browser platforms (Firefox, Chrome, Safari, Edge) was limited to the Google Chrome web browser[4] and the transcribing accuracy varied greatly. However, the no-cost nature of this service ensured it was a valuable entry-level option for those with limited resources and thus contributed to lowering the technology adoption barrier.
- **Microsoft Azure Speech-to-Text[5]:** This service offered cloud-based automatic speech recognition. The transcription accuracy was markedly improved compared to that of the Web Speech API, especially when provided with clear input audio (see Figure 2 for example). The cost was approximately $2 per hour. While not free, this was an order of magnitude cheaper than human captioning services (typically costing approximately $1 per minute), and did not require prior scheduling.

The WebSpeech API was the default speech-to-text engine for ScribeAR due to its no-cost nature. It allowed us to provide a zero-cost "demonstration" version, where the only technology requirements were a Chrome browser, Internet connection and a microphone.

To switch to the Azure service, the user accesses the Azure settings menu and enters an Azure key loaded with cloud credits (Figure 4). Once the key is verified, the speech-to-text engine

[3] For more information, see wicg.github.io/speech-api/ and developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API.
[4] Available for download at www.google.com/chrome/.
[5] Service details are available at azure.microsoft.com/en-us/services/cognitive-services/speech-to-text.

switches from WebSpeech to Azure-based recognition. In addition to providing more accurate captions, the Azure service can provide transcription for many other common languages, as well as real-time translation of speech input for multiple language pairs.

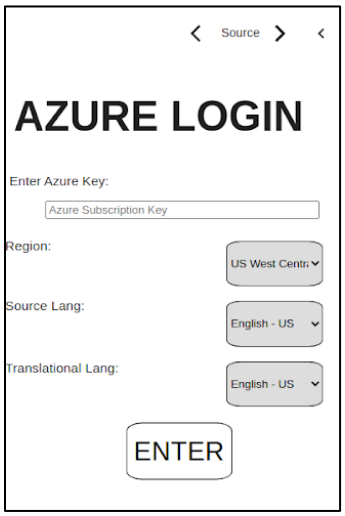

Figure 4: Azure settings menu. When the Azure key is verified, the speech recognition switches from Web Speech API to Microsoft Azure.

When the WebSpeech and Azure solutions were tested by the student development team they discovered that a given captioning source was not always the appropriate choice due to cost or accuracy constraints. They concluded that expanding the number of captioning source options would benefit the user. Consequently, the team is currently implementing integration with other automatic speech recognition engines and online human-based captioning infrastructure, e.g., Communication Access Realtime Translation (CART). The Azure transcription service also supports trained audio models and domain words which improve accuracy. We plan to explore and evaluate these features in the future.

Criteria 4. Able to adapt the caption presentation to different visual contexts.

The development team designed four features to allow the ScribeAR platform to be adaptable to a wide array of visual contexts. These are discussed below.

Firstly, the ability to switch the theme from dark (Figure 2a) to light (Figure 5). Different themes were better suited to different live scenarios. For instance, when the dark theme was paired with an AR display, all black-colored areas became transparent as the Epson AR glasses use an additive display technology; i.e., they do not project the color black onto the viewing surface. The viewer is thus able to see both the captions and the surrounding environment. On the other hand, the light theme was more comparable to the format offered by existing online caption-streaming platforms designed for use with a traditional computer display such as StreamText[6].

[6] Platform details are available at www.streamtext.net.

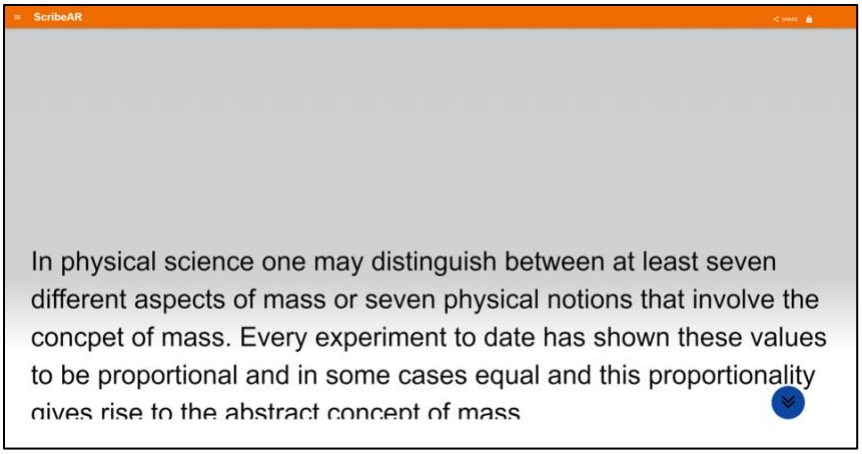

Figure 5: ScribeAR secondary (light) theme.

Secondly, users could adjust the font size to enable clear reading for students with low vision and for contexts with multiple users. For example, the captioning font did not need to be large when ScribeAR was being used by a single person sitting in front of a laptop. However, if multiple users were utilizing a single instance of ScribeAR on a laptop, then the laptop would need to be placed further away so that everyone could view the display, necessitating a larger font.

Thirdly, the location and size of the captioning text area were adjustable. Three options were implemented: lower bottom only, half screen, and full screen. The full-screen option was designed for use on a traditional laptop display, whereas the bottom-only option was designed for use with an AR display. By confining the captions to the lower area of the user's visual field, the interference between the live scene and the caption text was reduced.

Fourthly, it was distracting to have the ScribeAR navigation bar always visible at the top of the screen, especially when using an AR display. As a solution, we added a feature to unlock the navigation bar by pressing the lock icon in the top right corner of the screen, after which the bar would automatically hide. Touching or hovering the cursor over the bar area would cause it to reappear (Figure 6).

Figure 6: Locking and unlocking the navigation bar. The bar becomes visible when the cursor hovers on the top of the screen where the bar is usually visible.

Criteria 5. Allow captions to be archived or saved for future review.

To support scenarios where the user desired a transcript, such as a student saving a transcript of a class lecture as a study aid, we implemented an exporting feature where the transcript could be downloaded as a plain text file.

Criteria 6. Provide additional non-textual visual cues of auditory information.

While a transcription of speech provided significant access to auditory information, we also recognized the value of non-speech auditory cues in conveying secondary information such as the speaker's emotional state. As a step towards making these cues accessible, we provided a tool to visually indicate the input audio volume. A suite of visualization cues was created which were controlled by the preferences settings (Figures 7a and 7b). They varied in size and style, and the smaller ones could be placed at the user's desired location on the screen via an intuitive drag-and-drop method.

During testing we discovered that the volume indicators served as a useful diagnostic tool for the speech-to-text feature, especially for DHH users. For example, if the transcription accuracy degraded noticeably, the user may suspect that there may be loud background noises distorting the audio feed and verify via the audio visualization. Upon confirmation, the user could then take the appropriate actions to restore audio quality, such as by moving to a quieter location or addressing the source of the noise.

(a) (b)

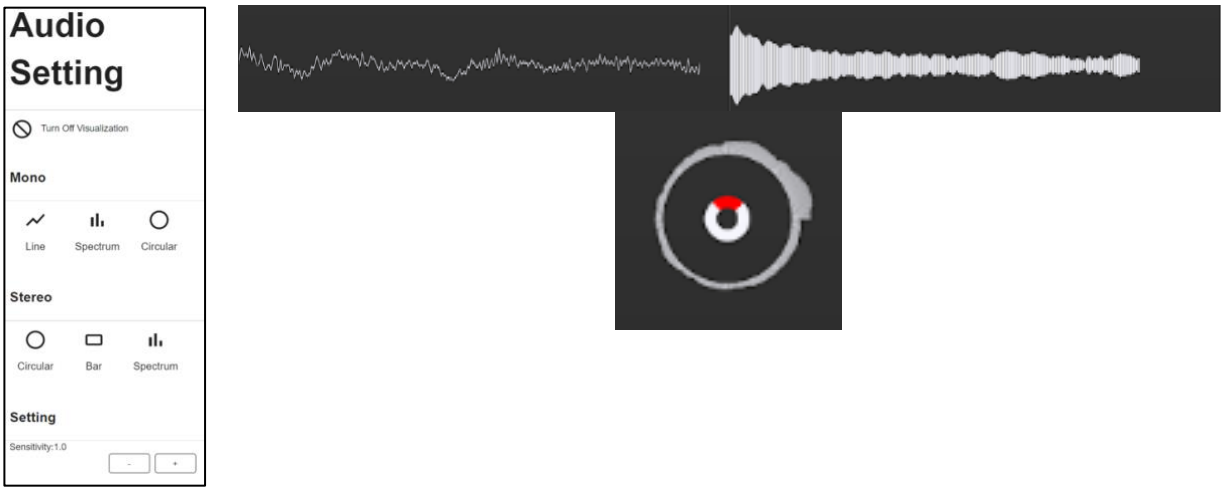

Figure 7: (a) Audio Settings. (b) Examples of 3 Audio visualizations: line, waveform, and circular. To support AR contexts the visualization can be dragged into different positions onscreen.

Future Work

To ensure that the six ScribeAR design criteria have been addressed from DHH users' perspectives in authentic educational settings, the development team will seek feedback once COVID-19 restrictions are lifted and in-person education activities resume. We briefly outline two new avenues of research that this work has created.

Use and evaluation of ScribeAR by DHH individuals for inclusive educational experiences

Participants will be provided with AR hardware for use with ScribeAR. Participants will be requested to utilize ScribeAR in multiple educational contexts (e.g., lectures, study groups, and

office hours) in both AR and non-AR modes. In seeking feedback, hardware practicality will be explored. Ease of use, subtitle comprehension, and social comfort level with respect to the hardware will also be evaluated.

Impact and tradeoffs of audio input mechanisms

An open question is the impact on transcription quality from using different audio sources. A variety of options exist, including a remote lapel microphone, laptop microphone, or classroom microphone. Each option will be likely to impact the quality of the transcription as well as the practicality of the overall system. It is difficult to evaluate the feasibility of specific audio input mechanisms in the development laboratory since we lack the ability to reliably simulate real-world conditions (such as background noise). Data gathered from real-world testing will be significantly more informative. We will work with test subjects to ensure that a variety of audio input mechanisms are trialed in a variety of cases and gather users' feedback. We believe this will lead to the identification of an additional design criteria with regard to ensuring a practical, high-quality audio input.

Conclusions

We identified and presented six major design objectives for an augmented reality captioning system for use in educational settings. The ScribeAR system was designed and built to satisfy these objectives. Using web and cloud technologies, ScribeAR has enabled the presentation of augmented-reality real-time transcriptions with minimal visual interference between the live scene and the transcription text. Audio visualizations provided additional contextual information to DHH users. Our design has identified additional education and research activities and we look forward to evaluating and improving the system in authentic, in-person university education settings once COVID-19 restrictions are relaxed.

Acknowledgements

This project would have not started without financial support from Lance Cooper and the Department of Physics Graduate Office, for which we are grateful. We also thank the VR@Illinois program at the University of Illinois at Urbana Champaign for providing additional seed funding for this project. We appreciate the Azure cloud credits provided by Microsoft Corporation for speech-to-text transcription, and the support of the Microsoft University Relations Director, Harold Javid. We thank illustrator,William Ryan, for his help in preparing this manuscript. Lastly, we thank current and former ScribeAR team members for their contributions to this project in addition to those of the authors: Alex Ackerman, Blair Wang, Jiaming Zhang, Nikhil Richard, Sicong Zhang, Will Foster, and Xinyu Liu.

References:

Kushalnagar, R. & Kushalnagar, P. Collaborative Gaze Cues and Replay for Deaf and Hard of Hearing Students. in *Computers Helping People with Special Needs* (eds. Miesenberger, K., Fels, D., Archambault, D., Peňáz, P. & Zagler, W.) 415–422 (Springer International Publishing, 2014).

Marschark, M. *et al.* Benefits of Sign Language Interpreting and Text Alternatives for Deaf Students' Classroom Learning. *The Journal of Deaf Studies and Deaf Education* **11**, 421–437 (2006).

Jain, D., Chinh, B., Findlater, L., Kushalnagar, R. & Froehlich, J. Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study. in *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems* 7–11 (Association for Computing Machinery, 2018).

Miller, A. *et al.* The Use of Smart Glasses for Lecture Comprehension by Deaf and Hard of Hearing Students. in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* 1909–1915 (Association for Computing Machinery, 2017).