Teleconference Captioning Accessibility

Emelia Beldon, Michael Tota, Norman Williams, Christian Vogler, and Raja Kushalnagar

Abstract

To have access to any type of information or conversation is vital for every single human, regardless of their disabilities or the situation. Now, 36 million Americans who are deaf and hard of hearing desire for equal accessibility and inclusion to information. Having equal access means the DHH community would benefit and be able to thrive from receiving information at the same time as everyone else. The studies and push for the improvement in accessibility in all situations has been in the making for a long time, however this year, the focus has been shifted to teleconferencing. The sole reason being COVID-19, because when it hit, the balance of physically working and learning shifted abruptly to millions requiring to work and learn online - through teleconferences. This had a domino effect on the effort of improving accessibility with real-time captioning, especially with automatic captioning. This leads us to the purpose for this paper and our product, finding the best solution/standard to provide a product that improves the accessibility to information in teleconferencing platforms. For this reason, we created a product to display real-time captioning and transcript with features that were lacking in the commercial teleconferencing apps based on research. The product is created with WebRTC and contains two appearances of captioning, one with captioning on each participant's video window and another with a transcript on the side. We then presented our product and its features to a group of participants to gather constructive user feedback and criticism to improve our product for better user accessibility and experience. The goal of this study and our product is to eliminate the barriers that have been up for too long and provide the accessibility that the DHH people need.

Introduction

The COVID-19 pandemic era has made an impact on almost everyone, little or big. And with that, the community of 36 million Americans who are deaf and hard of hearing took a direct impact on access to information in teleconferencing. To expand, captioning for teleconferencing is essential because in the past few months there has been an increase in the usage of teleconference [2], as illustrated in Figure 1, the chart shows that video chat apps usage have increased by 500%. Thus with the information, access to captioning in teleconferencing makes for a much more significant impact on participation in society. Especially now for those working or studying from home during the era. In addition to the increase of teleconferencing, we are also currently in the middle of a disruptive transition from human-generated captions to computer-generated captions (Automated Speech Recognition) that can be viewed anywhere, anytime, including through teleconferencing [6]. These new technologies provide a wider platform but also create different types of caption errors in comparison to human-generated

captioning techniques that have evolved in the past 40 years. Errors that prevent full accessibility to information. As a result, there has been much more consumer frustration and problems like appearance, accuracy, and location of the captions have now popped up with the computer-generated captions [4].

With this project, we aim to research, develop and provide a better quality and efficient technological product for automatic (computer-generated) captions to create an environment that improves the accessibility to information in teleconferencing platforms.

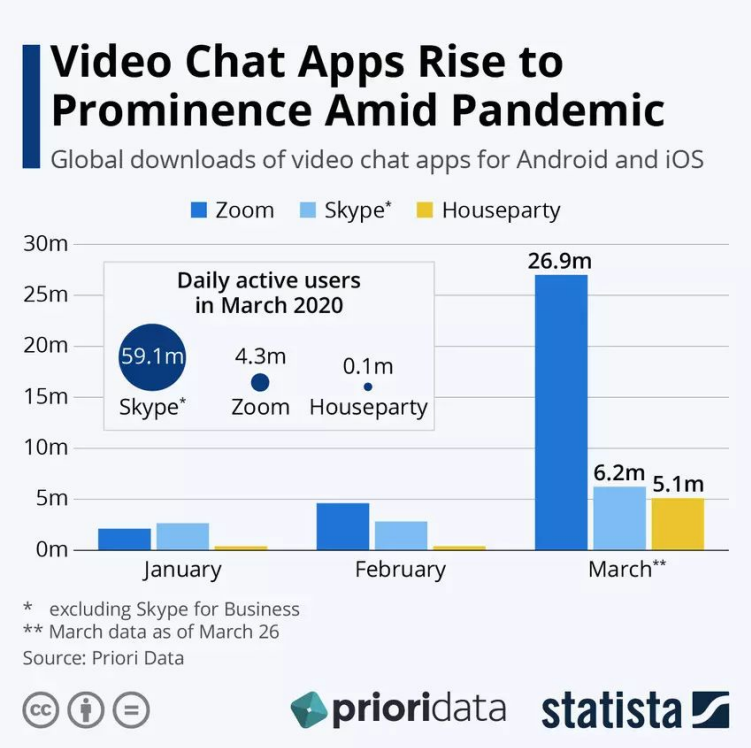


Figure 1

Related Work

Prior studies show the benefits of using an automated real-time captioning or transcription have been shown to enhance learning and teaching in meetings or lectures [9] as well as having cost and preparation overhead reduced and accessibility increased [7]. Thus it is significant to include transcription or captioning in teleconferences, especially if essential information is being shared.

To discuss the difference between real-time captioning and real-time transcription, work has been done about the potential switch of traditional captioning to the transcript on the side [7]. The difference between captions and transcripts lies in the text lines shown on screen. Captions typically have 2 to 3 lines in an overlay, transcripts on the other hand, have many more lines and are shown on the side or a separate window.

Results show that traditional captions are preferred by users in typical use but transcripts are preferred in technical content [7]. Results also support a longer caption history that makes it easier for viewers to absorb information from video sources regardless of the positioning of the captions. To support this statement, there are studies on the eye-tracking of users and how much time they spend looking at captions and transcripts versus the screen. One study found that readers tend to look at the captions about 84% of the time [4] with another study that found that readers spend a less amount of time on transcripts while viewing the screen, at about 68% of the time [1]. So the results of prior studies show that no matter of positioning, the history of captioning makes a difference in the users' usability and perception of information.

Although, prior studies also show that there is insufficient work on the intersection of ASR live captioning or transcription of teleconferencing or small group meetings and the preferences in appearance of the DHH community [3], hence the purpose of our study. However, a study based on the survey of DHH users shows that incorporating ASR captioning into commercial software is supportive and receptive based on a survey of DHH users to support their conversations in the workplace or in small group meetings.

The results show that users prefer to have ASR to transcribe their audio of meetings as well as having the ability to customize the appearance of the captions in order to reduce the users' cognitive burden [3]. In addition to that, there is no single standard for displaying visual transcription (captions or transcript) on the web [7]. Most are shown through browser plugins with default features and limitations so with this study, we aim to find a standard that maximizes user accessibility, usability, and experience. A standard that consists of features that addresses the problems of accessibility such as giving the users an option of having captions directly overlaying the video or in the transcript on the side, color coding and naming speakers to identify who is speaking, an option of switching between ASR services, and indicators to communicate if the participant is speaking or typing.

Methodology

To create a product with accessible and qualified computer-generated captions, we used WebRTC [11], an open sourced video and real-time-text (RTT) chat playground to display captions. This is a Real Time Communication platform, similar to Zoom or Google Meets, that uses Node.js for the backend and the standard JavaScript, HTML5, and CSS for the frontend.

Before this study, a reconstructed playground with live video streaming and real-time-text chat in each users' video window was created using the basic demos of WebRTC by Norman Williams and Gallaudet University Technology Access Program [10]. However, the reconstructed playground did not have any ASR services or real-time captioning, and only contained the RTT feature. So we adopted the scripts and added coding to create the accessible real-time captioning and transcription.

We first opened our study with exploring the computer-generated captions services provided by different commercial services such as Google, IBM, and Microsoft and comparing their usage and usability. We implemented the sample services in our coding of WebRTC by using documentations from the commercial services and adding the text in the original textarea

for RTT. We implemented the services to test and filter out the inefficient ones and then chose the two most qualified services in our finalizing WebRTC product. The two ASR services we chose were Microsoft Azure's Speech-To-Text service[12] and Web Speech API under Google[13] based on the flexibility of the service, accuracy, and punctuation. After selecting our services, we moved to the next phase which was to develop an accessible WebRTC with real-time captioning/transcription and the RTT feature.

Building Accessible WebRTC

Creating the accessible WebRTC, we developed two different ways to display the two ASR streaming services and its captioning. We developed in hopes of creating the most accessible platform.

Teleconference Participant Captions

The first way was captioning directly on each participants' video window along with the RTT feature shown in Figure 2. We did this by streaming the text in the textarea of the RTT feature as mentioned earlier. Coding was required to stream the text live by having the code stream the ASR service's automatic recognizing code and replacing it with the ASR service's final and accurate text. Users would be able to get live computer-generated captioning as well as communicate via text/chat.

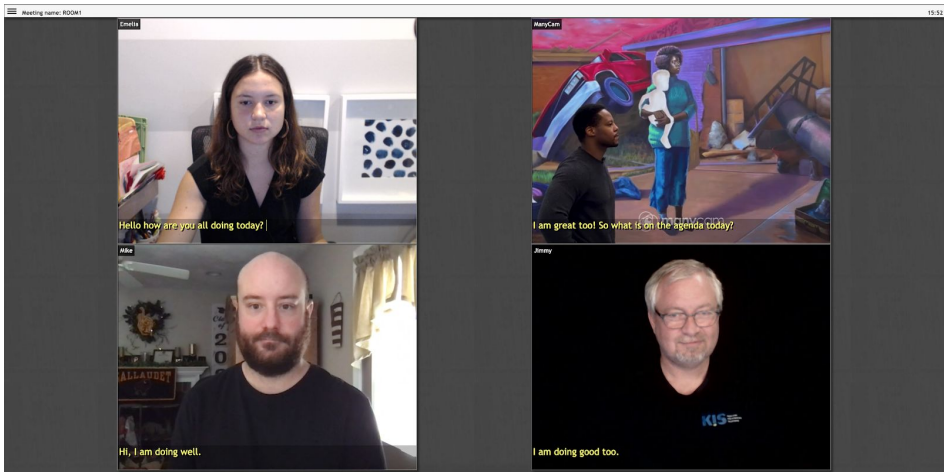


Figure 2

Teleconference Transcript

The second way was in a transcript form in a separate window on the side with speaker identification and the RTT feature shown in Figure 3. We developed this appearance by creating a new window and having the ASR service's automatic computer-generated captions stream not only to the textarea but to the new window for content history and references. Extra coding was

required for this because this appearance had more features like difference in color for each user in the conference and the ability to save the transcript in a text file for future purposes.

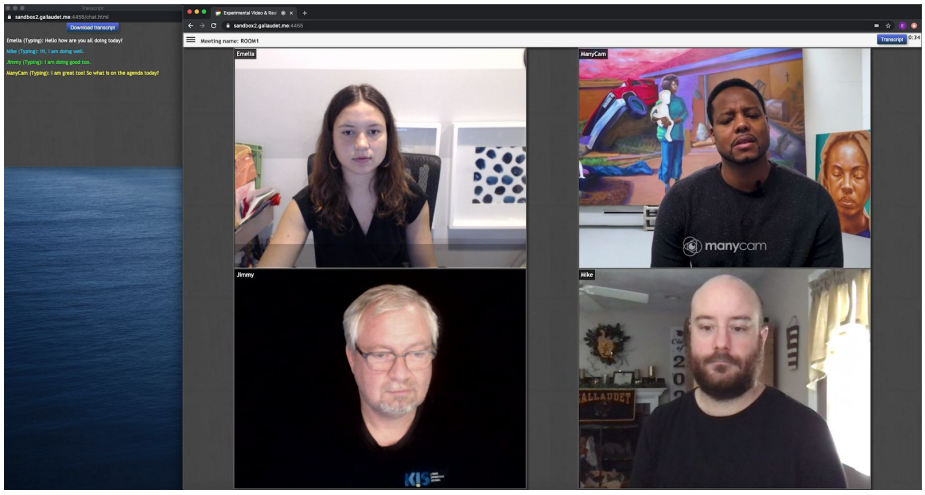


Figure 3

Typing Option

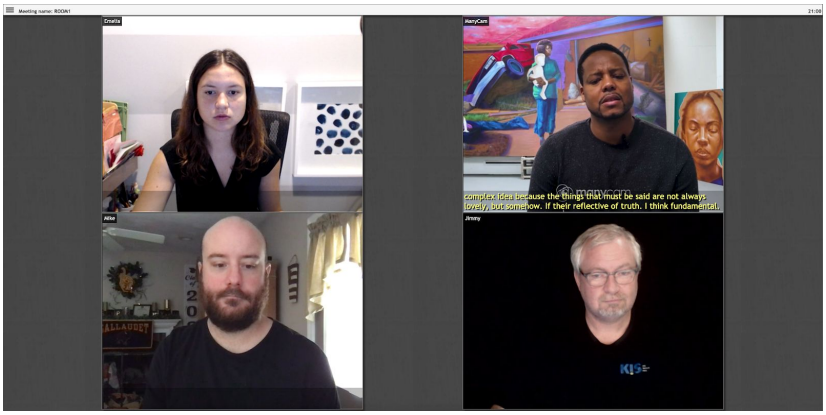
Some more features were also added to maximize the user interface experience. Features, as mentioned in previous paragraphs, were the real-time text streaming, indicators, (Speaking) and (Typing), to communicate if the participant is speaking or using RTT chat, and pairing names with the users to identify who is speaking/typing. Users of the product would also be able to the automatic appearance of captioning on the screen or click a button for a transcript window on the side, as well as toggle between the two ASR services.

Evaluating Accessible WebRTC

Now to evaluate user experience with computer-generated captions and the appearance of our product, we created two different environments/versions to test. The versions were the two different ways to display real-time captioning and transcript with Version A with the captioning directly in the participants' window and Version B with the transcript window on the side. The two versions were created in the goal of gathering constructive feedback about each feature and finding the best interface for users to implement in the final product post study.

We tested the two versions by conducting A/B testing with 2 hearing and 4 DHH participants to get user feedback. Each participant would navigate both versions and answer questions post testing. The navigation consisted of 4 scenarios for each version. Each of the scenarios mimicked a situation to present and test our product. The scenarios were a video playback of a TedTalk to mimic a presentation, a television playing to mimic another speaker, audio from Text-To-Speech app through the hosts' audio to mimic a conversation between two speakers, and all participants using the RTT feature. These scenarios are shown in Figures 4 to 7. After the testing occurred participants were to answer a survey about two sections; their

experiences with general video conferencing apps for general feedback and their experience with our product for specific appearance and usability feedback.



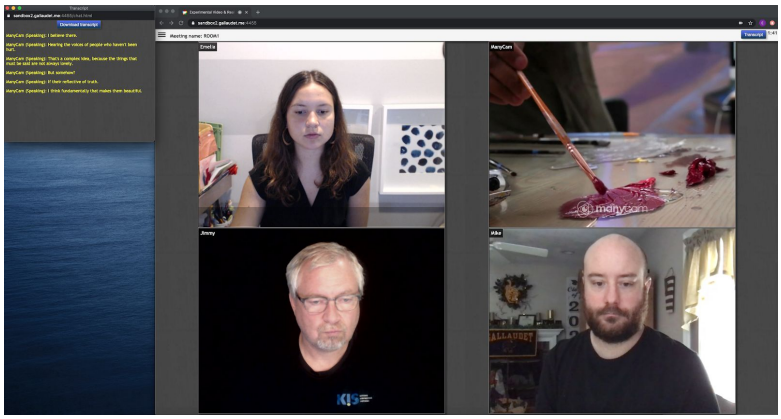
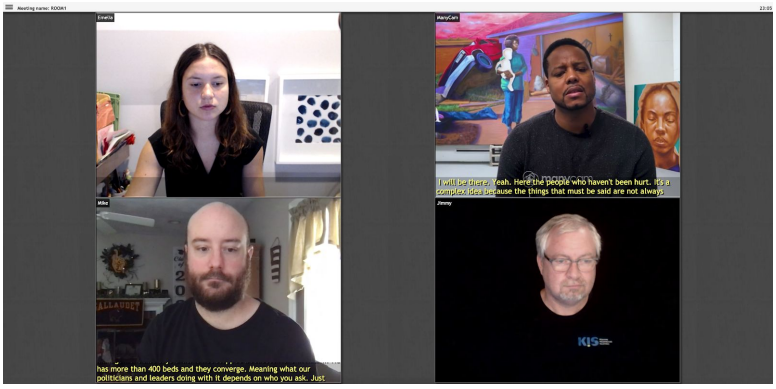


Figure 4



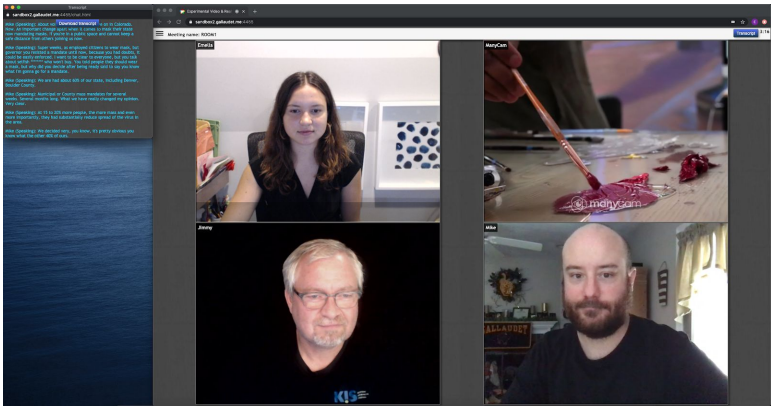
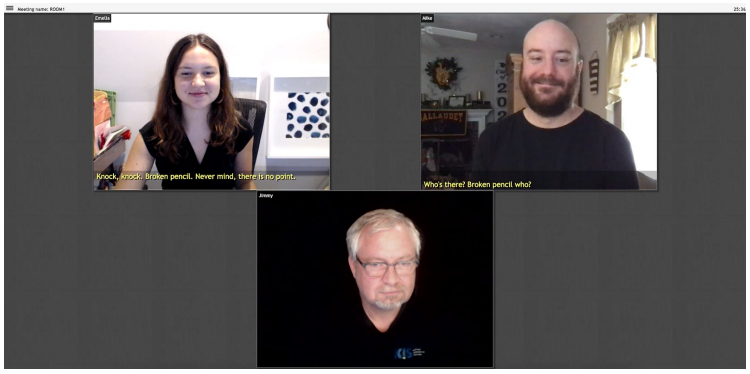


Figure 5



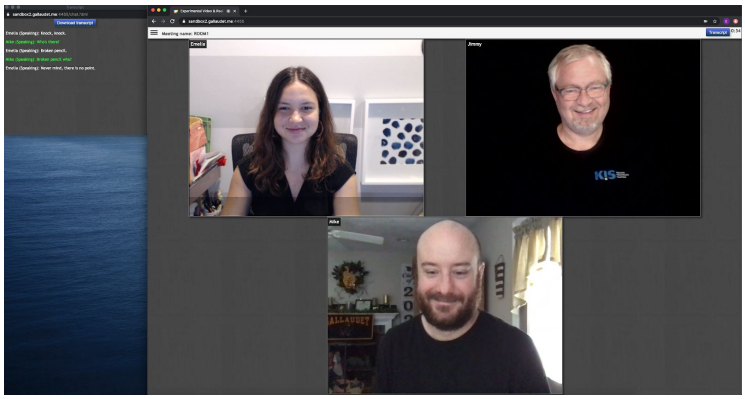
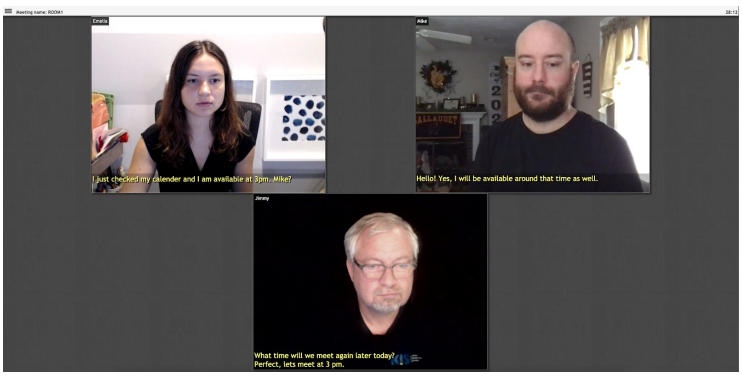


Figure 6



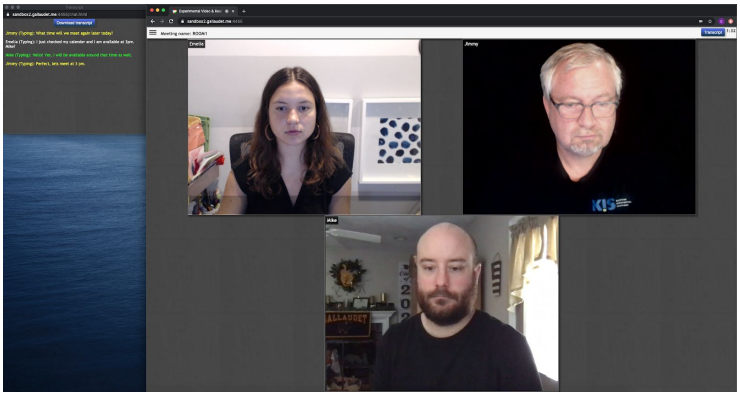


Figure 7

Results

The results of the A/B testing had shown a number of things in regards to their general experience with accessibility of commercial teleconference services, its captions, or anything else related to accessibility, and their experience with the versions of our product.

Current Teleconference Accessibility Experiences

For the first section, questions were asked about the navigation and experience of captioning in their “go-to” commercial video conferencing app. Questions that were asked are listed below. The answers showed that more than half of the participants have had awkward or poor experiences. P2 mentioned;

“*I was in a Zoom meeting recently with some outside parties, who were hearing. We had interpreters in the meeting. However, somebody found the Gallaudet Zoom feature that provides RTC, and turned it on. I then had to figure out how to disable it for myself while trying to keep track of the conversation.*”

and P3, also in agreement, commented;

"*It was awkward at first, I did not know how to navigate the video conference app, Zoom, to turn on the captions. The instructions wasn't very clear.*".

The next question asked about the accessibility of the captions and only half were able to say "Yes, it was accessible and easy to use" - refer to Figure 8. Questions were also asked about how they captioned their calls, what method they used or their third-party plug in. Answers again showed that more than half don't have a preference or a "go-to" third party for captioning or they rely on the option provided by the video conferencing app. This shows that most users don't have access to a better standard or real-time captioning for their video conference calls.

Lastly, in the first section, a question asked how would the participants improve their captioning experiences or what would they change in their video conferencing app. P1 mentioned the quality of the real-time captioning (the ASR service it uses), and adding color of fonts for who is talking. P2 talks about the system not automatically setting the RTC for all users regardless of their preference, and P4 mentions instructions for DHH or Hearing users to use the captioning for the first time. These feedbacks were beneficial and contributed to the adjustment and improvement of our product.

Q1:	If you can remember, what was your experience like for the first time using captions in video conference calls?
Q2:	How do you caption your video conference calls?
Q3:	Which video conferencing app is your go-to?
Q4:	For the #1 one app you chose - why do you use that one?
Q5:	In r.e., to the #1 app you chose - did you find it easy to navigate the app?
Q6:	In r.e., to the #1 app you chose - were the captions easy to find? Were they easy to turn on and off?
Q7:	In r.e., to the #1 app you chose - how would you improve the captioning experience?
Q8:	In r.e., to the #1 app you chose - if you could change anything in the app, what would you change?
Q9:	For the other apps on the list - Why do you not use x app?
Q10:	In r.e., to the apps you did not choose - how would you improve the captioning experience?

Table 1

Figure 8

Participant Feedback

The second section had questions directly related to our product and the two versions the participants tested. Questions that were asked are listed below. In the survey, we had the participants first enter their experiences with both versions, shown in Figure 9. This gave us an insight of their experience.

We then had them list the things they liked and disliked about each version so that we could take those criticisms in and make adjustments for our final product as well as record what they liked for future references.

Version A had positive feedback on being able to see exactly who is talking with captions streaming in each participants' own video window and the ability of being able to see both the captions and the speaker to stay connected and interactive. The positive feedbacks for version B were the ability to see everything in one place because that was easier on the eyes, the scrollable feature for content history, the use of colors to identify speakers, the (Speaking) and (Typing) indicators, and the ability to save/download the transcript for future use.

Now for the criticism in Version A, we received feedback about the captions being unscrollable (not being able to see content history), reduce shadowing in the font style, the confusion about being able to type in the caption area, and the difficulty to type in a noisy room with the ASR service possibly overlapping the RTT feature. And the criticism for Version B was the inconvenient way of switching between the transcript window and the video conference window when using the RTT feature, having the transcript window also closing when closing the video conference window, and the disconnection between the speaker and reader because the readers had to advert their attention to another window.

This feedback and criticism part of the survey was very crucial in the improvement of our product. With these answers, we are able to know areas of strength and weaknesses within our product.

Finally, the last question asks which display of captions/transcript they would recommend to the commercial video conferencing apps, it was a split right in the middle vote with 3 voting Version A and another 3 voting Version B - refer to Figure 10. With these answers, we were able to make the necessary adjustments in our product to improve the usability and quality of our real-time captioning and transcription service.

Q1:	How was your experience with the captioning directly on the participants' windows in WebRTC?
Q2:	How was your experience with the captioning in the transcript that overlays the windows?
Q3:	What are the features you liked the most with the captioning directly on the participants' windows?
Q4:	What are the features you liked the most with the captioning in the transcript section that overlays the windows?
Q5:	What are the features you disliked the most with the captioning directly on the participants' windows?
Q6:	What are the features you disliked the most with the captioning in the transcript section that overlays the windows?
Q7:	Which method would you recommend for all teleconferencing apps? I.e. Zoom, Google Meets, Microsoft Hangouts

Table 2

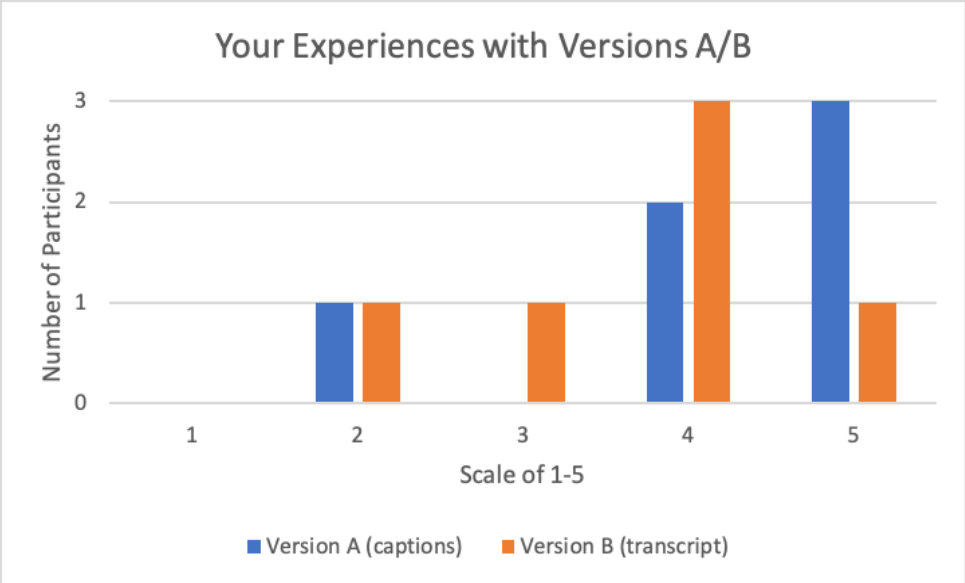


Figure 9

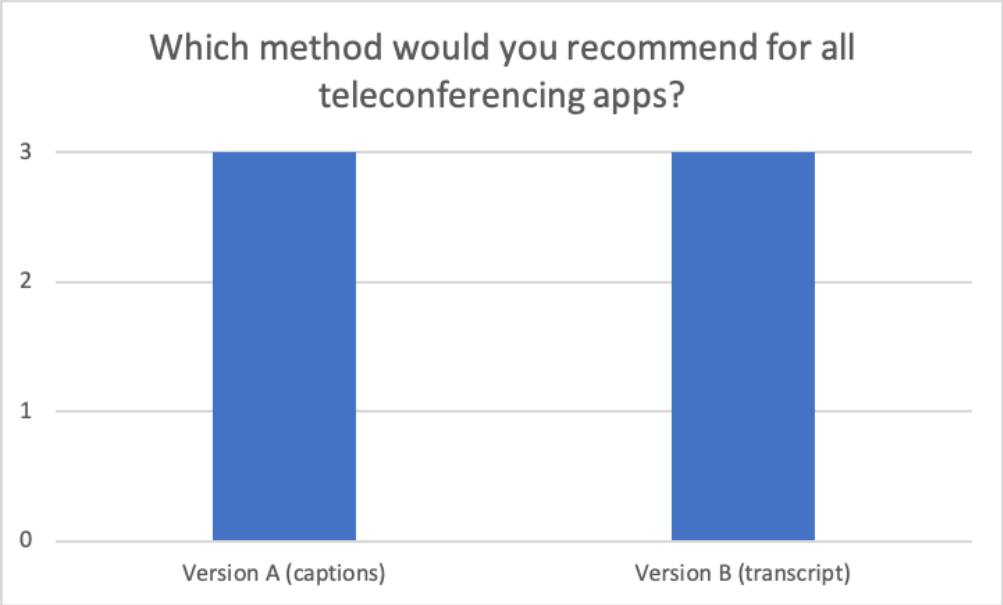


Figure 10

Iterative Development

Based on the user feedback and the criticism section of this paper, we were able to make changes to our product on the go and hopefully improve the experience for each test participant . These are some of the adjustments we made.

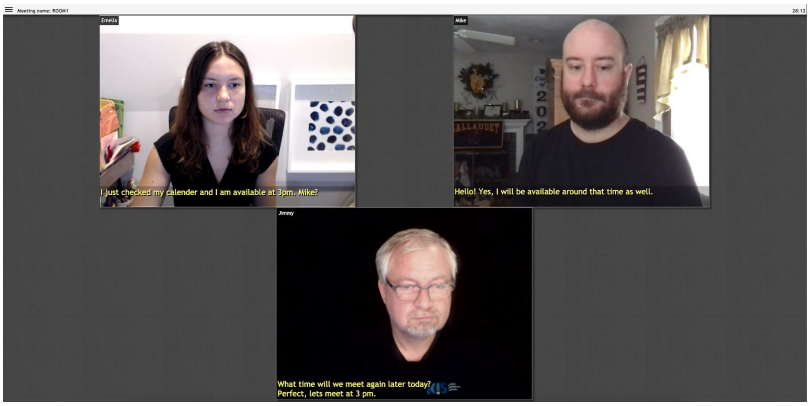
The first participant had constructive feedback about improving the captions' font style - the color and shadowing to make it more clear so we made changes in selecting better colors

and reducing the shadow effect. P1 also gave feedback on the wording for our indicators (Speaking) and (Typing). We originally had the wording: (Speaking) when the speaker was talking and then (Spoke) when the speaker stopped talking but that was a visual distraction because when the word would change from Speaking to Spoke, readers would refer back and be misguided that a change occurred in the sentence or what was being said. So we took that in and had the indicators stick with (Speaking) for whenever the speaker was talking and (Typing) for when the user uses the RTT feature. This way, we could keep the wording consistent in our product.

The other addition that we made in our product after receiving feedback, was the ability to review the caption history of the caption overlay in each participant's video window. We first depended our history on the transcript window on the side but after some tests, we realized that users also favor the capability to review the captions in the original window in case they missed context or if it scrolled up too fast. So we decided to add a feature where you could double click on the participant video's caption overlay and it would expand the text area to show the complete caption history. This way, the users would be able to review the history in both appearances. Refer to Figure 11 for example.

Lastly, the final adjustment we made was based on a feedback that we received from P2 about the transcript window. She mentioned that when closing the main video conference window, the transcript window did not close automatically. This is considered to be annoying to the users because it requires more user interaction and loses the concept of making the navigation easier. So with that, we added some code and was able to have the transcript window close automatically when the main video conferencing window closes. This was a constructive feedback and hopefully an adjustment to make our product accessible and convenient to use.

We were not able to face all of the feedback with the time we had, but we made the adjustments to the best of our abilities to improve user experience as much as possible. These feedbacks will be marked and confronted when we finalize our product.



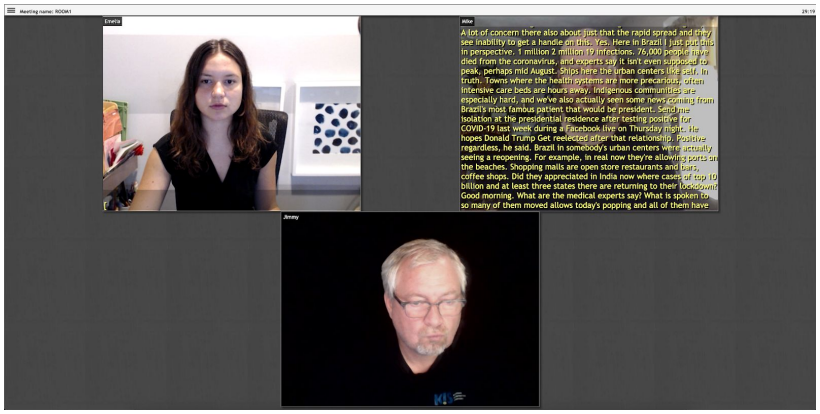


Figure 11

Discussion

Based on the results of the A/B testing, participants seem to value the ability to refer to captioning history, the RTT feature for deaf users to communicate with hearing users, identifying different speakers with names and color, the interaction and connection to the speaker, and to save/download transcripts. These features are consistent in both Versions: captioning and transcript. With being able to interact and connect with users as well as clearly indicating who is talking in the captioning in each participants' video window appearance. And with being able to color code speakers and save/download transcripts in the transcript in a new window appearance. These results tell us that users are in favor of features provided by both of our product's two ways of displaying real-time computer-generated captioning. These features seem to give an extra boost of accessibility that is not present in most current commercial video conferencing apps. The takeaway of this is that the users would prefer or be more open to having both appearance and having the capability to choose either or both to be able to benefit from all features.

Now to refer back to previous studies on this concept, the results of this study shows support in the result of study on Captions vs Transcript in Online Content [8]. Users of real-time captioning would prefer to be able to choose their preferred style of captioning, whether it be captioning directly on the video windows or a transcript window on the side or both.

Future Work

Now, because this study was conducted in limited time, we were unable to complete the final product for publication. We want to conduct more A/B testing to receive more feedback along with the feedback we have not yet referred to. Some of the adjustments we hope to make are having instructions or an explanation of our features on the webpage for first time users, the ability to keep the transcript window in the front of the video conference window at all times, regardless of the interaction of either window, and automatically muting or turning off ASR

whenever the user starts typing so there would be no overlay in services. Hopefully when we finalize our product, it is able to provide an experience with features that would maximize accessibility and reduce any barriers of the DHH community.

Conclusion

Throughout this paper, we exhibited our product's two ways to display real-time computer-generated captions with numerous features included. Features help support a positive user accessible experience to information and conversations in teleconferencing apps. Features such as the ability to conversate with speaking or real-time text, review content history, identify speakers with names and color coding, position and location of the captions (in each participants' windows instead of in the middle of everything), turning on captioning for only the user rather than automatically for everyone, and toggling between the appearances and ASR services. The features listed support a new standard to satisfy users' preferences based on our study and previous studies. Also note that post study, we will continue to add and adjust features to provide the most accessible environment. In all, our product allows users to have access to information as well as communicate information with little to almost no barriers. With our product and its features as well as the developments and continuous future work, we hope to finalize our product with a better accessible standard before releasing it to Gallaudet University and then hopefully for all video conferencing apps to adopt as well.

Acknowledgements

We would like to thank our participants for all the constructive feedback and criticism that supported the development of our product. Funding for this paper was generously provided by NSF 1757836, NSF 1763219 and NIDILRR 90DPC0002.

References

[1] Anna C. Cavender, Jeffrey P. Bigham, and Richard E. Ladner. 2009. ClassInFocus: enabling improved visual attention strategies for deaf and hard of hearing students. In Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility (Assets '09). Association for Computing Machinery, New York, NY, USA, 67–74. DOI :https://doi.org/10.1145/1639642.1639656

[2] Beauford, Moshe, (2020, March 5). With COVID-19 Spreading, Video Conferencing is Booming. *UC Today.* *https://www.uctoday.com/collaboration/video-conferencing/with-covid-19-spreading-video-conferencing-is-booming/*

[3] Berke, L., Albusays, K., Seita, M., & Huenerfauth, M. (2019, May). Preferred Appearance of Captions Generated by Automatic Speech Recognition for Deaf and Hard-of-Hearing Viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-6). http://library.usc.edu.ph/ACM/CHI2019/2exabs/LBW1713.pdf

[4] Glasser, Abraham, Kesavan Kushalnagar, and Raja Kushalnagar. "Deaf, hard of hearing, and hearing perspectives on using automatic speech recognition in conversation." *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 2017. https://arxiv.org/pdf/1909.01176.pdf

[5] Jensema, Carl J., Ramalinga Sarma Danturthi, and Robert Burch. "Time spent viewing captions on television programs." *American annals of the deaf* (2000): 464-468. https://search.proquest.com/docview/214477300?fromopenview=true&pq-origsite=gscholar

[6] Juang, Biing-Hwang, and Lawrence R. Rabiner. "Automatic speech recognition–a brief history of the technology development." *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1 (2005): 67. https://my.fit.edu/~vkepuska/ece5527/ASRHistory-Juang+Rabiner.pdf

[7] Kheir, Richard, and Thomas Way. "Inclusion of deaf students in computer science classes using real-time speech transcription." *Proceedings of the 12th annual SIGCSE conference on Innovation and technology in computer science education*. 2007. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.297.3545&rep=rep1&type=pdf

[8] Kushalnagar, Raja S., Walter S. Lasecki, and Jeffrey P. Bigham. "Captions versus transcripts for online video content." *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. 2013. http://www.cs.cmu.edu/~jbigham/pubs/pdfs/2014/captionvstranscripts.pdf

[9] Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P., & Duerstock, B. S. (2013). Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies*, *6*(4), 299-311. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6529071

[10] https://webrtc.gallaudet.edu/

[11] https://webrtc.org/

[12] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

[13] https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API/Using_the_Web_Speech_API

[14] "The Video Apps We're Downloading amid the Coronavirus Pandemic." *World Economic Forum*. *www.weforum.org*, https://www.weforum.org/agenda/2020/03/infographic-apps-pandemic-technology-data-coronavirus-covid19-tech/. Accessed 17 July 2020.