Abstract

The purpose of this research is to increase the accessibility of captions in WebRTC and RTT in order to ease the transmitting of information during a conversation between a user speaking orally and deaf or hard-of-hearing users. In this paper, we look into which settings are preferred and if they are influenced by a multitude of factors. We propose that WebRTC applications include an option for dynamic caption, which is a setting where the caption is displayed next to the speaker's head, and a texting setting that allows RTT communication in a letter-by-letter format, as captioning options.

I. Introduction & Literature Review:

WebRTC, also known as Web Real-Time Communication, is a fairly new project, having only been around 10 years [1]. The technology is widely used today on many platforms, such as Google Meet, Google Duo, Zoom, WhatsApp, Facebook, Discord, and Snapchat [2,3]. Without using plugins or external applications, WebRTC is an open-source project that allows for real-time communication between two parties through video conferencing [4]. In 2011, Google released the source code for WebRTC, which has since allowed other companies to use the technology [5]. As a result of the COVID-19 pandemic, more people have had to depend on WebRTC and, according to Dr. Jeff Jaffe, W3C CEO, makes it a crucial technology for "information sharing, real-time communications and entertainment" [6]. Data shows that there was more digital communication following the start of the pandemic (recorded on April 4-8, 2020), with a 43% increase in texting, 36% in voice calls, 35% in social media, and 30% in video calls [7]. While around only 40% of people met virtually prior to the pandemic, this number has risen to nearly 100% over the pandemic [8]. Most forms of digital communication and virtual meetings use WebRTC which makes it a valuable technology.

With the increasing demand for WebRTC, our review of the literature intends to explore the capabilities and accessibility of WebRTC and further discuss the technology to the benefit of the general public that chat using WebRTC and captions over the Internet. WebRTC is an already highly competent technology as it can readily support more than 100 clients without any significant latency or loss of quality [9]. Furthermore, WebRTC does not require a high-speed Internet connection or high-end computer specifications in order to maintain adequate video and audio quality, having been shown to use less than 55 kbit/s and 20% central processing unit (CPU) usage [10]. The average Internet speed in the U.S. is 42.86 Mb/s which is equivalent to 42860 kbit/s [11]. Thus, on average, most people have more than 750 times the resources needed for a stable Internet connection. Every computer has a CPU, which can handle up to 100% usage and is sufficient for an average WebRTC session. We intend to move away from focusing on WebRTC's capabilities and more on its accessibility, usability, and quality of experience, specifically for deaf and hard-of-hearing (DHH) users.

WebRTC has proved its usefulness during the pandemic, especially as it is also used for Virtual Remote Interpreting (VRI). For instance, it was considered unsafe at one point for

interpreters to show up in person at hospitals, so VRI was used as an alternative to that [12]. An impressive feature of WebRTC is its speed and ability to deliver data quickly allowing it to be used for live captions, as well as Real-Time Text (RTT) in Zoom, Google Meet, and other video conferencing platforms. As such, we are interested to know whether or not captions differ between WebRTC video conferencing and news broadcasting and entertainment, including TV shows, movies, or other media. User experience (UX) research on captioning shows that most people prefer when captions are positioned outside of the video rather than overlaying a portion of the video, which may block content [13]. The third approach to caption positioning is to display captions dynamically inside the frame and placing them closer to the speaker to minimize eye movement. Most of the frequent caption users saw this third approach as a helpful improvement while people who did not use captions as often found it distracting [14]. Yet another way to implement captions in an in-person environment is through using Augmented Reality (AR). Users of AR captions preferred when captions were near the head, particularly to the right, rather than at the bottom of the screen [15]. In-person captioning may also be implemented by using a projector where text is displayed on a board near the speaker's head, which is being tracked using Real-Time-Text Display (RTTD). The study's results showed that DHH participants also prefer to have the text appear next to the right side of the speaker's head [16].

While there are many studies addressing the capabilities and limitations of WebRTC as a technology, there are little-to-no studies that focus on improving its user experience and none that talk specifically about the WebRTC caption-user interface (UI). Needless to say, there are no current studies that study DHH user experience with WebRTC. In exploring how other mediums address caption-UI, we plan to incorporate the findings into our study design to see if similar principles can be applied to the WebRTC interface. We want to look specifically at how the types of WebRTC caption-UI affect the understanding among DHH participants in a video conference environment. For example, we can adjust the caption position or the RTT frequency to find the most optimized setting for users' understanding. From this study, we intend to identify ways to improve WebRTC UI for DHH users, and thus improve its accessibility.

II. Method

This study was approved by Gallaudet University's Institutional Review Board. We started recruitment for participants residing in the United States in July of 2021 through social media (e.g.:Facebook and Discord) and word of mouth. Participants were self-selected and compensated $25 for their time in an hour-long study. Participants filled out the demographic questionnaire in Google Forms before the Zoom meeting. Short questionnaires were implemented into the WebRTC demonstration to gather their responses based on the Likert scale. The data we collected include a sample size of 21 adult participants. A demographic questionnaire administered to participants asked about their gender, age, ethnicity, education level, deaf identity, hearing ability, sign language skills, lip-reading skills, and experience with

using technology and captions. The last few questions ask about their experience with using WebRTC. Questions include “How well do you understand speakers in TV, videos, and other media without using captions or subtitles?” with a rating scale of 1 (Not at all) to 5 (Very well).

Using a premade WebRTC demonstration shared by Gallaudet University, we developed and coded new video captions-UI conditions that we wanted to test. We adjusted the demonstrations to make them more suitable for testing purposes. We added buttons and links to our questionnaire and the next demonstrations. We also implemented the RTT style of displaying captions word-by-word and line-by-line within the caption area. All of these changes were added and edited using Visual Studio Code. To create dynamic captions, we attempted to code face tracking that would attach captions to the right of the speaker's head, however, the video began to lag and degrade in quality. It was possible to adjust the tracking interval and make it a background task as well as making a few other tweaks to make it less laggy. However, due to time constraints, we did not do that. As an alternative, we used video editing to create the illusion of live captions that move according to the position of the speaker’s head. Finally, we also used HiveQL to script the speed and timing of our typing in the chat with participants, ensuring conversational consistency among different participants. The same idea was applied within our first part of testing by using pre-recorded conversations and playing them accordingly as the participants respond.

Once participants joined the Zoom meeting and provided consent, we introduced ourselves and explained what our research is about and our agenda for Phase One and Two. During the Zoom meeting, they were asked to share their screen to allow us to follow their progress and their response in using the WebRTC demo. Participants were directed to open the demonstration link shared via Zoom chat and asked to turn the Zoom camera off for the duration of the demonstration so the WebRTC camera would work properly. Phase One, which was designed to show the location preference for captions, consisted of four pre-recorded videos of a hearing speaker posing a series of open-ended questions to the participant, conversation-style. The caption locations were inside the bottom of the video, outside the bottom of the video, a transcript in a new window, and dynamic captioning that changed location when the speaker moved. The speakers in the videos, male and female, spoke clearly and to the camera as if they were talking to the participants themselves. The spoken audio would be transcribed live using Microsoft Azure’s API to convert them into words to display on the caption area. We controlled the video timing of the videos in real-time so that there was enough time for the participant to finish talking or for the video to skip forward if their replies were short to keep the conversation speed as natural as possible. Participants could respond either in ASL, orally, or by typing their answers. Upon completion of all four videos and respective questionnaires, we transitioned to the second part of testing.

Phase Two looked at the type of RTT style, in a constant location, during a conversational style live interview with one of the authors. In this phase, the author typed questions to the participants which showed up on their screen in the different RTT styles. The participants were required to type their responses, while their cameras remained on for us to track their responses

and eye movements. All captions were at the inside bottom of the video, but the RTT style changed from letter-by-letter to word-by-word, then to line-by-line. For the last two RTT styles, participants saw a status of "[author] is typing.." to reduce the speculation of whether the interviewer was still typing. As in Phase One, participants completed short questionnaires after completing each scenario.

Between both Phases, there were seven scenarios participants were asked to evaluate. The captioning conditions are listed in Table 1.

Table 1:

Video (location of captions)	**RTT (style)**
Inside & bottom of the video	letter-by-letter (by typing)
Outside & bottom of the video	letter-by-letter (by typing)
Transcript off to the side	letter-by-letter (by typing)
Next to the speaker's head	N/A style (no typing)
Inside & bottom of the video	letter-by-letter (by typing)
Inside & bottom of the video	word-by-word (by typing)
Inside & bottom of the video	line-by-line (by typing)

Between each minute-and-a-half-long video or RTT scenario, participants had the opportunity to respond to at least three questions. For example, we asked: "How easy was it for you to understand the captions?" with Likert scale response options, "Difficult to understand," "Neither difficult nor easy to understand," "Easy to understand," and "Very easy to understand." We also included an open-ended question asking if the participant had any additional comments about the scenario they just participated in. Similarly, we asked, "How often did you know what the other person was going to say before they finished?" with a Likert scale-style response. This question provided us insight into how quickly a user may understand information using various RTT settings. From the data we gathered, we analyzed if and how participant demographics had any correlation with caption UI preferences.

III. Results

The data we gathered from participants for each video condition is summarized in Table 2. "Mean understanding" was the average rating provided among participants for the question "How easy was it for you to understand the caption?" Mean visibility" was the average rating

that participants gave for the question “How easy was it for you to see the captions and the speaker at the same time?” “Mean favorites” was the average response to the question “Can you organize the live captions settings you've seen from most favorite to least?” where the most favorite was given 4 points, the second favorite was 3 points, the third favorite was 2 points and only one point for the least favorite.

Table 2:

Video Conditions	Mean understanding	Mean visibility	Mean favorites
Inside the video	4.428571429	3.904761905	3.095238095
Outside the video	3.952380952	3.333333333	2.19047619
Transcript	4.095238095	2.666666667	1.476190476
Dynamic	4.714285714	4.666666667	3.238095238

The data for RTT conditions are summarized below in Table 3 where “Mean predictability” was the average rating that people gave for the question “How often did you know what the other person was going to say before they finished?” “Mean favorites” was the average value to the question “Can you organize the texting settings you've seen from most favorite to least?” where the most favorite was given 3 points, the second favorite is 2 points, and the least favorite is only one point.

Table 3:

RTT Conditions	Mean predictability	Mean favorites
letter-by-letter	3.904761905	2.571428571
word-by-word	3.666666667	2.285714286
line-by-line	2.428571429	1.142857143

Table 4, below, indicates how many people use a certain language as their main method of communication. The percentages were used to compare how many people like various settings in each group to see if there are any correlations.

Table 4:

Main Language	#	Percent (%)	Dynamic %	Transcript %	Inside %	Outside %
ASL	12	57.14%	66.66%	8.33%	16.66%	8.33%
Spoken	5	23.81%	60%	0%	40%	0%
SimCom	4	19.05%	50%	0%	50%	0%

Other demographic information, such as self-rated ASL ability, self-rated lip-reading level, as well as ability to understand a speaker without captions are included as potential factors, as well. Figures 1, 2 and 3 below show a quick summary of that information:

Figure 1:

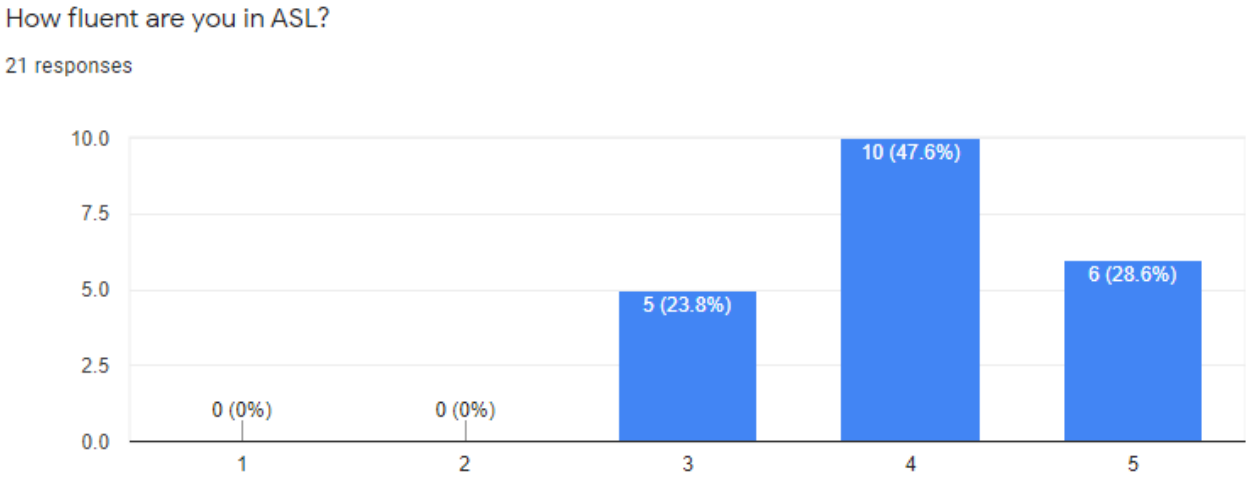


Figure 2:

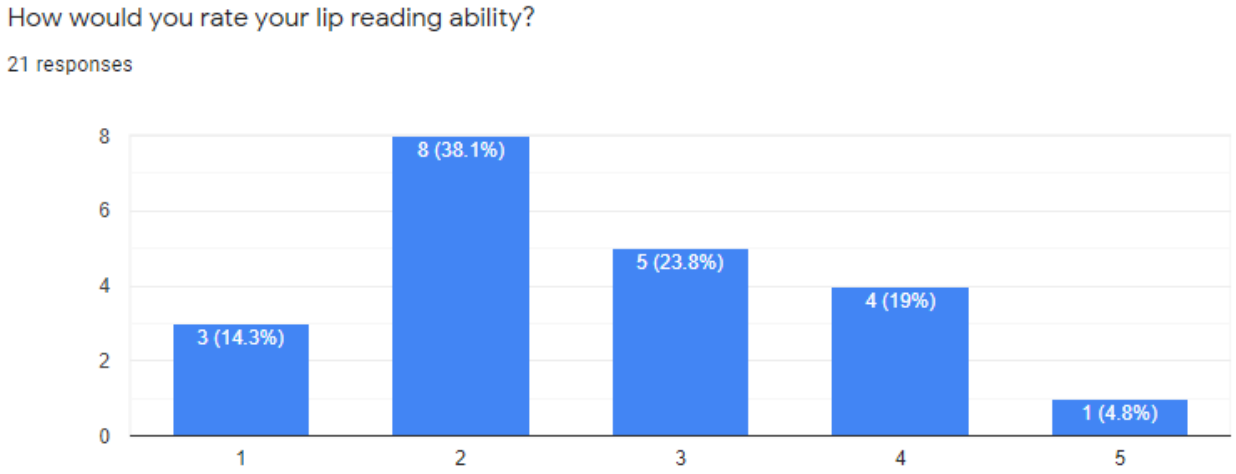
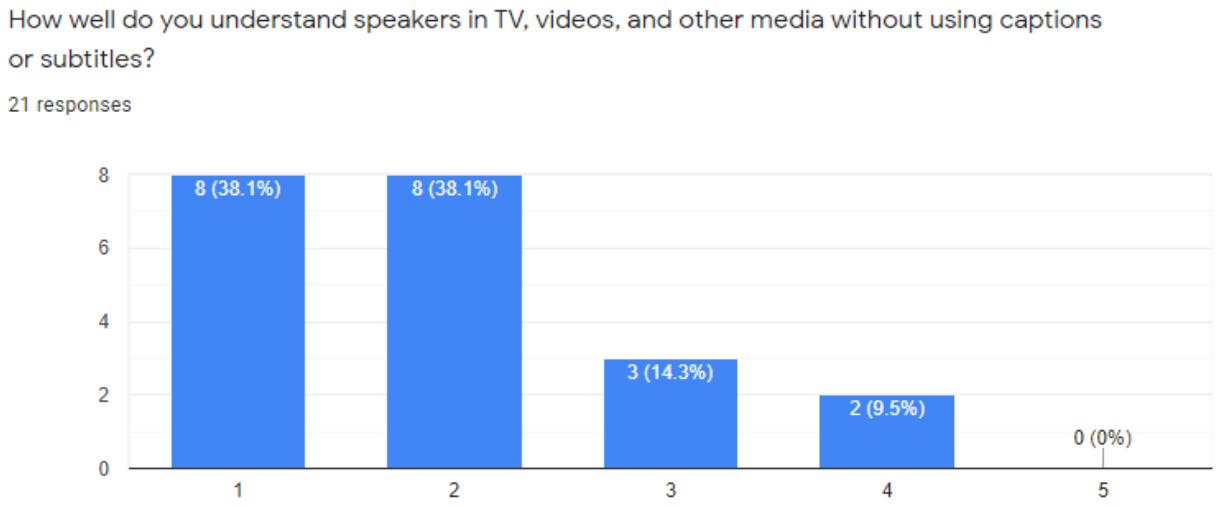


Figure 3:

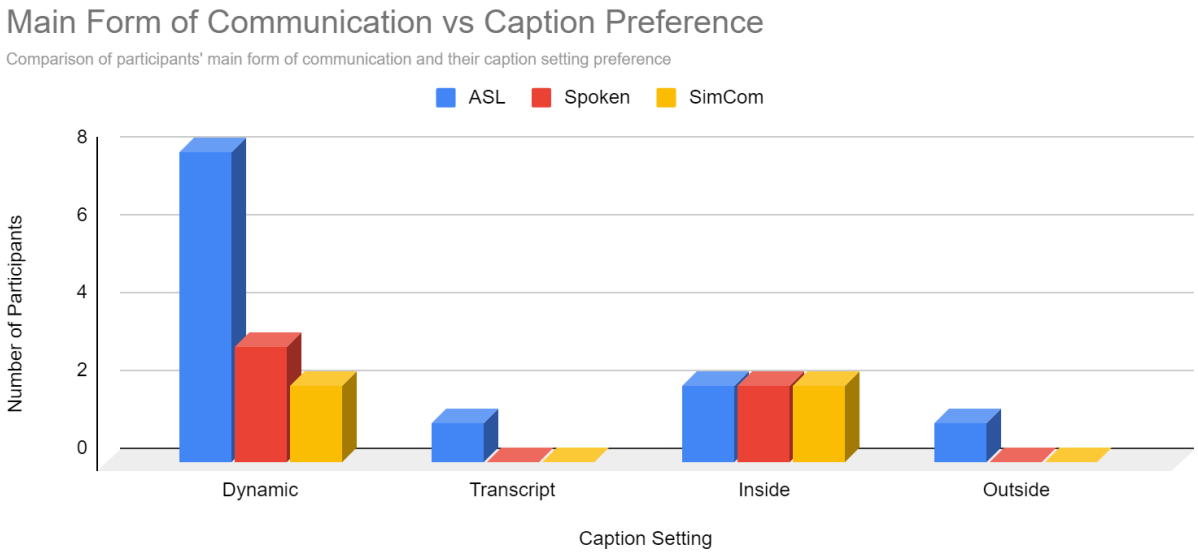


IV. Discussion and Analysis

Our data shows that the majority of participants prefer dynamic captions regardless of their main form of communication shown. This could be because the closer the captioning is to the speaker's head, the easier it is to watch facial expressions and identify emotions. There were a couple of outliers among the participants who preferred the captions on a transcript window or outside of the video instead to avoid obscuring any content. Sixty percent of the participants that used spoken English as their main language of communication preferred the caption style of dynamic captions, where the captions move location when the speaker moves to stay near the speaker's head. Around forty percent of spoken English participants preferred captions inside of the video, instead of other location options. Again, dynamic captions could make it easier for spoken English participants to read lips and captions at the same time.

To see if other factors such as the participants' main form of communication is statistically relevant, we compared that to their caption preferences, shown in Figure 4 below. While it would seem that all participants liked dynamic captions the most, the SimCom group is split between dynamic captions and captions inside the video. Since the sample size of SimCom was only four, it is difficult to draw conclusions based on just that. ASL users certainly liked dynamic captions the most, whereas the spoken English group is somewhere in between favoring dynamic captions and captions located at the inside bottom of the video. We acknowledge that some participants choose ASL as their main form of communication but yet choose to speak during the study and that this may have influenced the data.

Figure 4:



We compared the main form of communication among our participants and their preference for the RTT settings in Figure 5 below. Out of all RTT settings we tested, we noticed that almost all of our participants disliked line-by-line because they were unsure if a speaker was still typing despite our typing indicator at the top of the video. Most participants preferred the letter-by-letter format with two-thirds of ASL participants, three-fifths of spoken English participants, and three-fourths of SimCom participants favoring letter-by-letter. We believe that it is because they liked the predictability of RTT with letter-by-letter which allowed them to know what the typing user may want to convey before they've finished, which gives the participant more time to think and respond.

Figure 5:

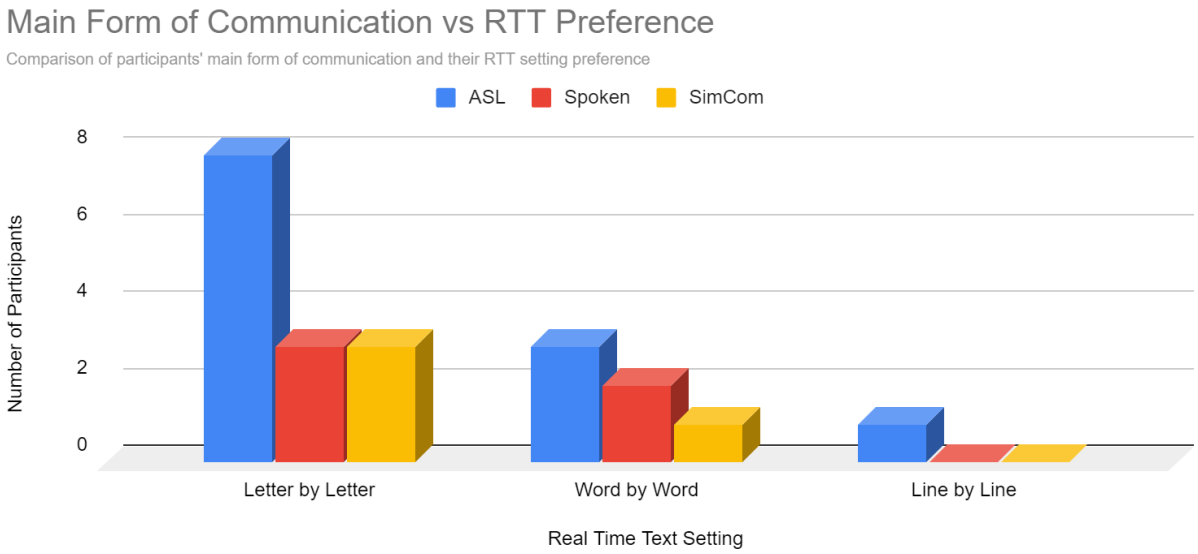
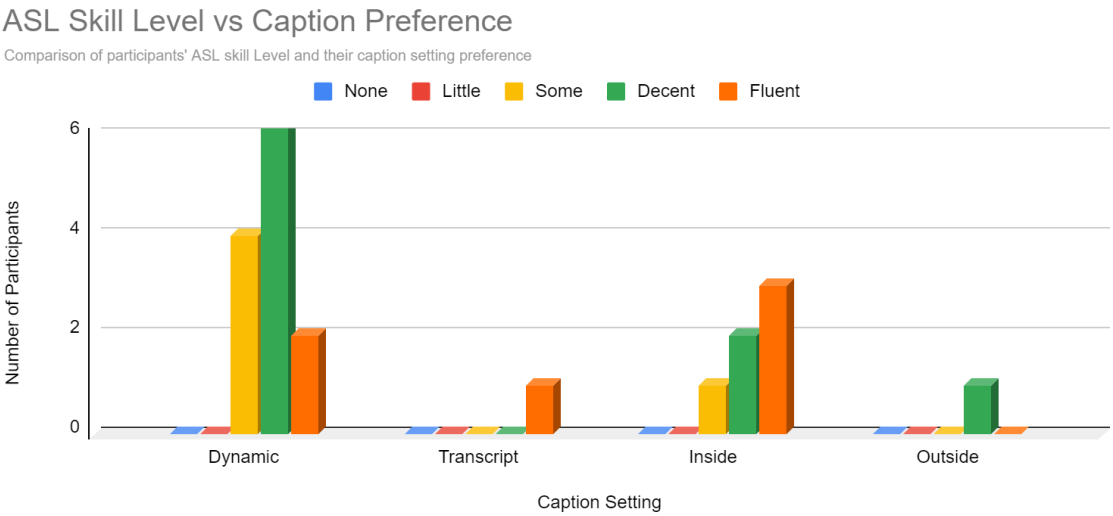


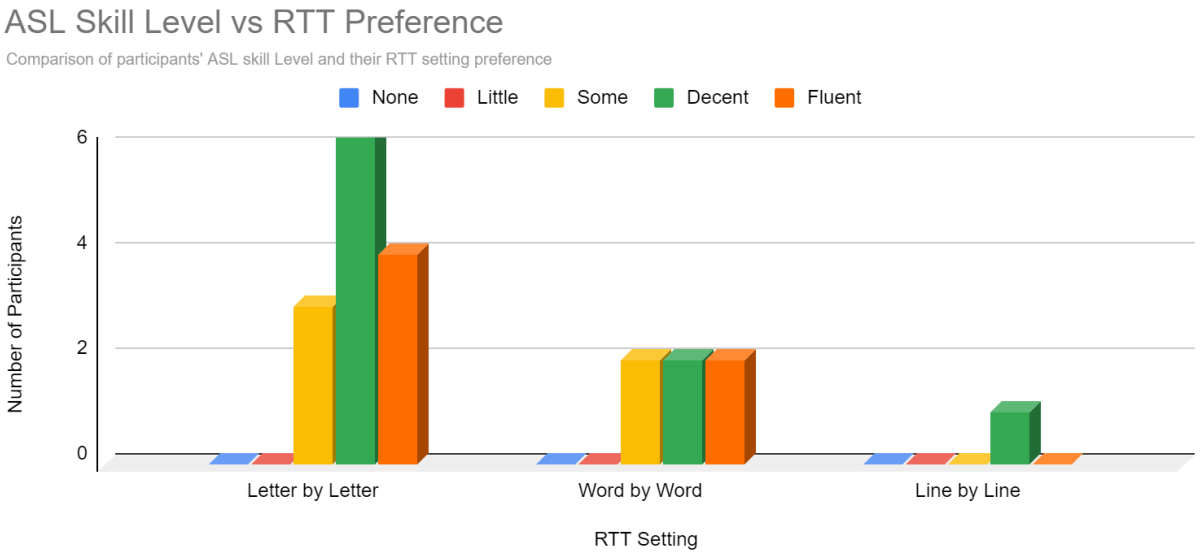
Figure 6 looks at the participant's self-rated ASL skill level compared to their caption preferences. Interestingly, the more fluent a person rated their ASL level, the less they preferred dynamic captions in favor of captions inside the bottom of the video. Participants "decent" in ASL strongly preferred dynamic captions.

Figure 6:



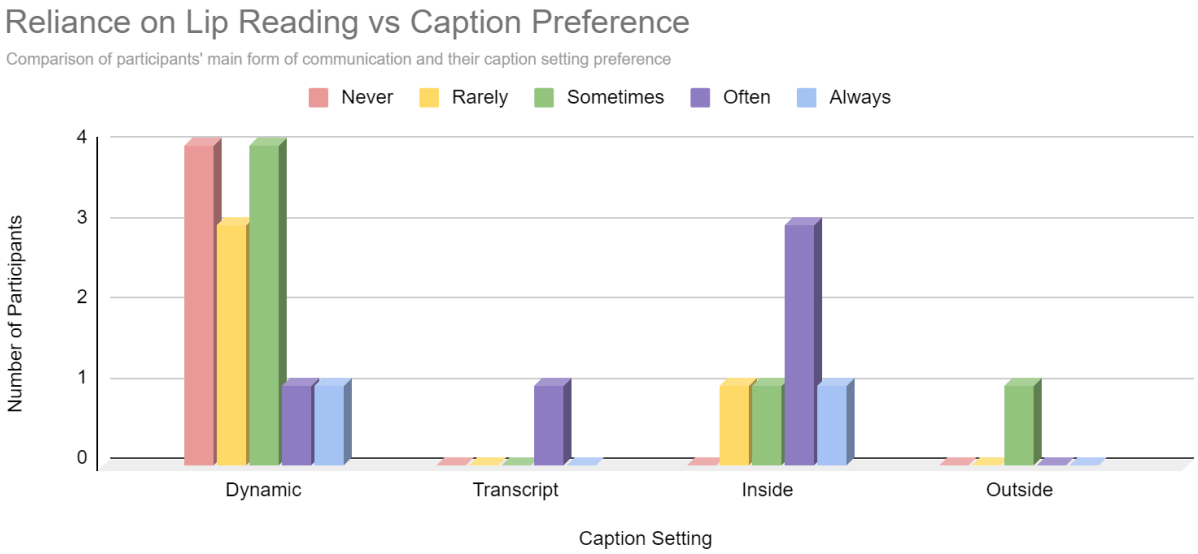
As for ASL skill level in the RTT settings as displayed in Figure 7 below, the majority of participants preferred letter-by-letter RTT style than the two other settings.

Figure 7:



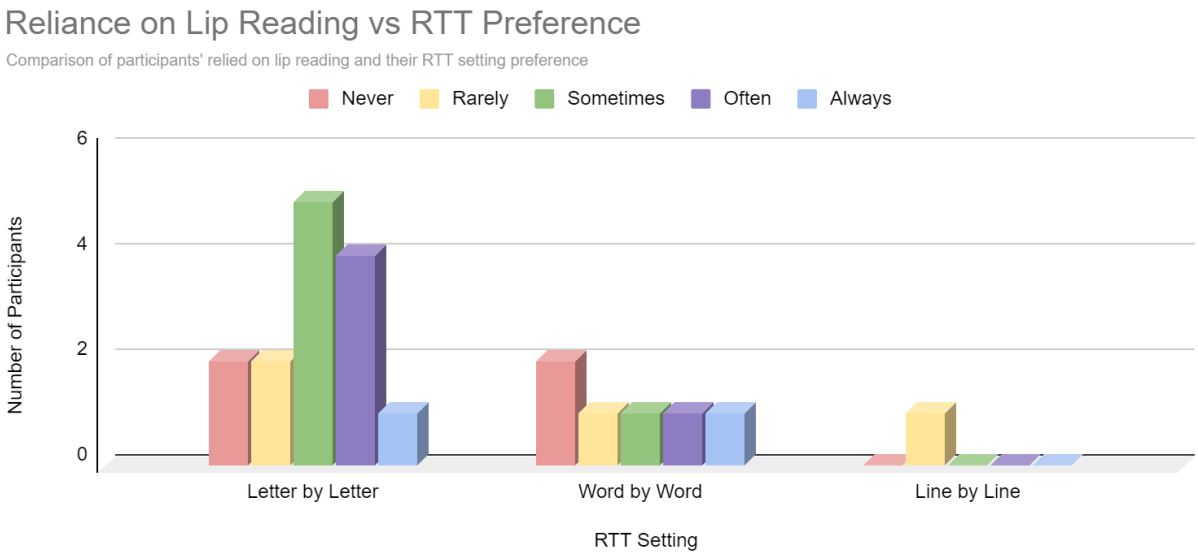
Participants that "often" rely on lip-reading to improve their understanding of the speaker preferred inside captions as shown in Figure 8 below. Conversely, people that "sometimes" or "rarely" use lip-reading preferred dynamic captions. This data is interesting as we expected that using dynamic captions would improve one's ability to read the speaker's lips and would have been preferable. However, many of the more lip-reading-reliant participants have dismissed that conjecture as they said dynamic captions were distracting for them, preferring to focus on the speaker and read the captions if they feel like they've missed something. The less lip-reading-reliant participants enjoyed dynamic captions because they were able to see the speaker's facial expressions. These participants relied more on facial expression to understand the speaker's tone which helped them to grasp the context and understand the material better.

Figure 8:



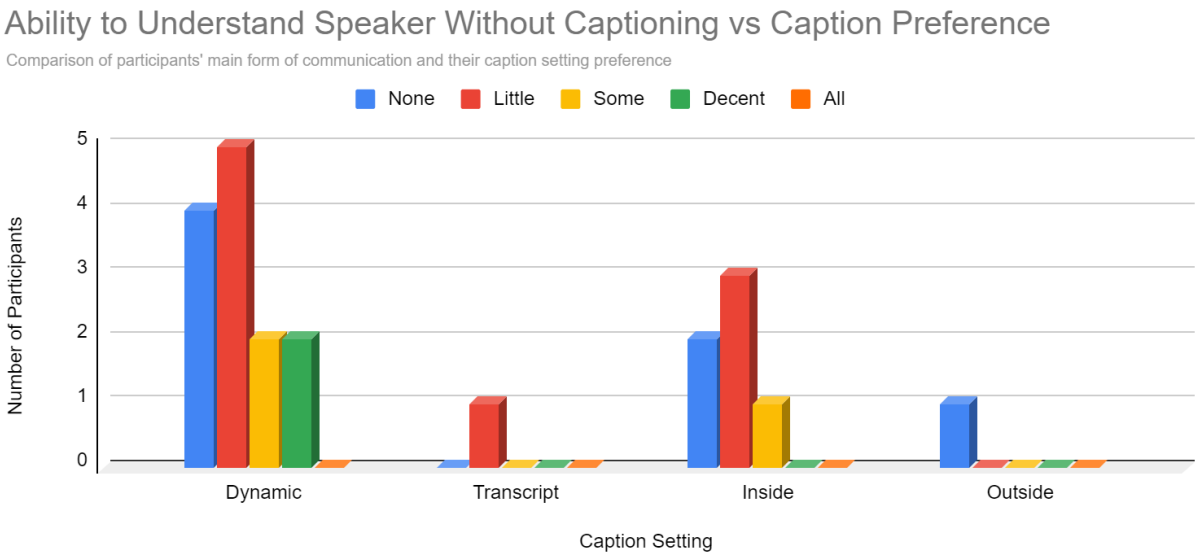
For the people who rely on lip-reading, the preference between letter-by-letter and word-by-word RTT varies. In Figure 9, below, the graph shows that participants that "never" rely on lip-reading prefer either letter-by-letter or word-by-word RTT settings. However, participants that "often" use lip reading, preferred letter-by-letter by a landslide. One of the participants said, "This [letter-by-letter] is pretty much my preferred communication method. So the communication is equal since we read and type." Like this participant, other participants seemed to enjoy having a communication method that both speaker and participant can use equally and be able to express their exact words whilst understanding the intended messages from the other side.

Figure 9:



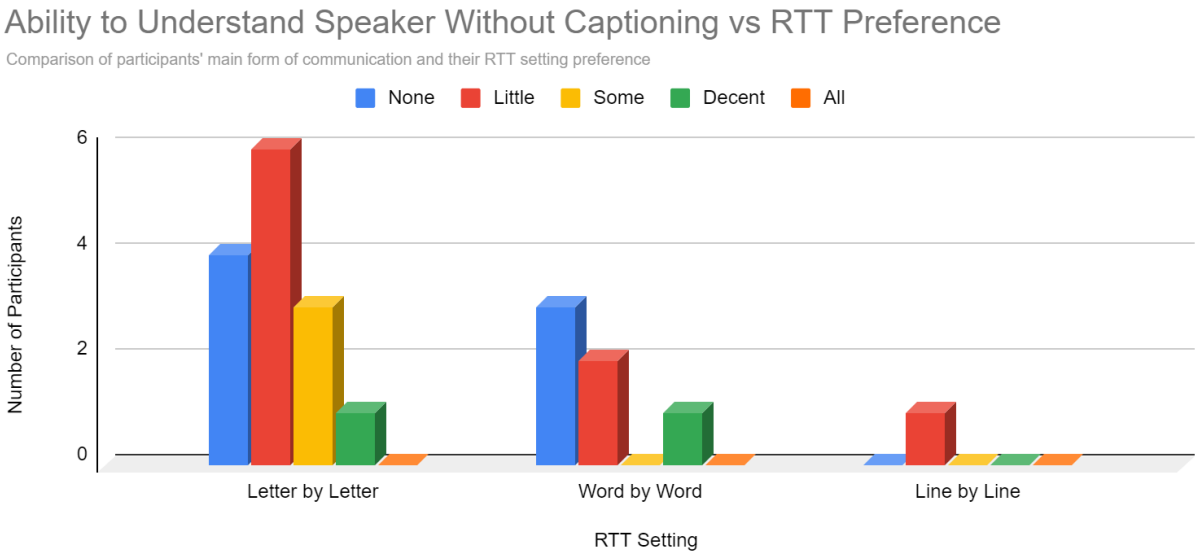
Our participants who understood “none” or “little” of the speaker’s message without captions preferred dynamic captions. However, those who understood “some” or a “decent” amount of the message without captions did not show as strong of a preference for one caption style over another. Figure 10 shows the less caption-reliant participants had lesser deviation between caption options as opposed to other groups.

Figure 10:



In all data for comprehension between any common among participants, they all but one agreed that they dislike line-by-line easily. In Figure 11 below, this graph shows participants' preference on RTT settings with their ability to understand a speaker without captions. The data shows that the majority of participants prefer a letter-by-letter RTT setting with only one participant preferring line-by-line.

Figure 11:



V. Limitations

Several issues arose during the study which may have affected the accuracy of the data. The most notable one was the Internet speed on our side as well as the participants' side. Contrary to our literature review, our study required more speed and stability in the connection in order for us both to see each other smoothly with no latency.

As mentioned in the method section, due to time constraints, we made the dynamic setting's captions pre-recorded, which was perfectly accurate, unlike the other settings that relied on live captions. Participants may perceive this setting as easier to understand and slightly skew the data.

Additionally, the errors of the live captions weren't consistent. For example, when the speaker says, "without them," the captions sometimes show, "without the map," while at other times completely omit a word, or on rare occasions convey the speaker's words correctly. Hence, participants saw different versions of the captions which we did not intend.

Another possible limitation is the screen size that participants might have affected the UI look of WebRTC during sessions. One of the participant's browser views was set to portrait UI view which meant the video of the participant themselves and of the researcher was at the top

and bottom in a 2x1 column layout rather than the 1x2 row layout that we intended. We asked them to go in full-screen mode and the view was changed to the correct view.

Lastly, the questions were not to our satisfaction. We felt that better questions could be refined to be more specific in order to collect more detailed information and insights from the participants for data analysis. For example, one question we asked was, “How much did the typos affect your understanding of what the speaker wanted to say?” The question was trying to grasp how the participants may have understood more or less based on the intentional typos and subsequent fixes that the scripted typing displayed. Since the words were fixed, to some participants they were not technically typos so most participants answered that it did not affect their understanding at all. We could ask instead, “How many typos or typo corrections did you notice?” That way, the results should reflect that they noticed fewer typos in letter-by-letter than word-by-word and even fewer with line-by-line.

VI. Conclusion and Future Work

Based on our study, we found that in one-on-one or speaker-to-audience video conference settings, the dynamic captions based on face tracking settings were preferred over outside of video captions, inside of video captions, and transcripts in another window. We believe that this is because it is easier to read captions and at the same time keep track of the speaker’s facial expressions as well as their lips for understanding when the captions are close to the speaker's head as he moves. The letter-by-letter style is ranked first for RTT preference. We believe it is because it is much quicker for participants to know what the other side is about to say which allows them to respond quickly for faster communication As such, we propose that dynamic caption be one of the options in the caption setting for the WebRTC environment along with a letter-by-letter RTT setting.

We acknowledge that it is possible that deaf and hard-of-hearing participants may have different preferences if the video conference was set in an academic or workplace environment versus a casual environment one-on-one. There should be a future study based on those environments, and further studies looking at a few other issues such as Internet speed and screen size. There were a lot of errors in the captions but we did not ask participants if they noticed the errors or if the errors impacted their understanding of the speaker’s message. The first three video-caption settings used WebRTC’s auto speech recognition (ASR) captioning which caused errors in every scenario. The ASR live captions did not always output the same captions despite having the exact same audio. Some participants saw the speaker’s words correctly whereas other participants saw the wrong words displayed for the same audio. That could have influenced their understanding and responses. One participant did catch that she missed the speaker’s name and informed us that the live caption skipped the introduction part. The focus of this study was not on the errors of the captioning so ideally, our caption should have been ninety percent correct in each setting. These errors should be the same intentional errors for each participant to get truly consistent results.

The conversations script in the video setting and RTT setting currently felt inconsistent and it could be beneficial for our data to be more consistent. For example, the first three caption video settings had the conversation shared equally between speaker and participants but the fourth video (set in dynamic captioning) was not. The speaker was speaking the whole time during the fourth setting, giving the participant little chance to reply. Additionally, the script of the second scenario (set at the outside and bottom of the video captioning) was the only one that involved heavy visuals with the hands. Some participants complained that they had difficulty tracking the hands, the speaker, as well as the captions at the same time. “I have a hard time to see what she’s doing with her hands while reading the captions.” We did not include Internet speed as a factor to test the quality of WebRTC during sessions.

Lastly, we hope future study will increase the sample size based on some factors such as age groups, race, main language preferences, or caption reliance. Additionally, we can focus on things such as the font size, font color, or font family as some participants feel like the fonts are too big whereas other participants feel like it is too small. Dynamic captioning is also a possibility to focus on in a future study by testing various conditions such as the locations for the dynamic captions to be placed near the speaker, and the captions’ movement pattern, as well as their frequency in the movement to make it less jerky or more jerky. Overall, we believe that this study creates a foundation for future research that can improve the accessibility for deaf and hard-of-hearing users who rely on captions for information. There are many more possibilities in improving the accessibility of WebRTC captions for deaf and hard-of-hearing people.

VII. Acknowledgements

This work has been generously supported by an NSF REU Site Grant (#1757836) awarded to Dr. Raja Kushalnagar, PI, and Dr. Christian Vogler, Co-PI. We would like to thank all of our mentors (Dr. Raja Kushalnagar, Dr. Christian Vogler, Mr. Norman Williams, and Katja Jacobs) from Gallaudet University who helped and invested their time supporting our research and we would like to thank REU AICT for the opportunity to do this research.

VII. Reference

[1] Bergkvist, A., Burnett, D. C., Jennings, C., & Narayanan, A. (Eds.). (2011, October 27). WebRTC 1.0: Real-time Communication Between Browsers. https://www.w3.org/TR/2011/WD-webrtc-20111027/.

[2] Gross, G. (2020, December 8). *WebRTC technologies prove to be essential during pandemicGrant Gross.* IETF. https://www.ietf.org/blog/webrtc-pandemic/.

[3] Morton, A. (2015, August 31). *Top 19 Companies in WebRTC*. AT&T Developer. https://developer.att.com/blog/top-19-companies-in-webrtc.

[4] WebRTC. (n.d.). https://webrtc.org/.

[5] Alvestrand, H. (2011, May 31). *Google release of WebRTC source code*. Mailing lists archives. https://lists.w3.org/Archives/Public/public-webrtc/2011May/0022.html.

[6] *Web Real-Time Communications (WebRTC) transforms the communications landscape; becomes a World Wide Web Consortium (W3C) Recommendation and multiple Internet Engineering Task Force (IETF) standards*. W3C. (n.d.). https://www.w3.org/2021/01/pressrelease-webrtc-rec.html.en.

[7] Nguyen, M. H., Gruber, J., Fuchs, J., Marler, W., Hunsaker, A., & Hargittai, E. (2020). Changes in Digital Communication During the COVID-19 Global Pandemic: Implications for Digital Inequality and Future Research. Social Media + Society. https://doi.org/10.1177/2056305120948255

[8] Willem Standaert, Steve Muylle, Amit Basu, Business Meetings in a Post-Pandemic World: When and How to Meet Virtually?, Business Horizons, 2021, ISSN 0007-6813, https://doi.org/10.1016/j.bushor.2021.02.047.

[9] B. Garcia, F. Gortazar, L. Lopez-Fernandez, M. Gallego and M. Paris, "WebRTC Testing: Challenges and Practical Solutions," in IEEE Communications Standards Magazine, vol. 1, no. 2, pp. 36-42, 2017, doi: 10.1109/MCOMSTD.2017.1700005.

[10] N. M. Edan, A. Al-Sherbaz and S. Turner, "Design and evaluation of browser-to-browser video conferencing in WebRTC," 2017 Global Information Infrastructure and Networking Symposium (GIIS), 2017, pp. 75-78, doi: 10.1109/GIIS.2017.8169813.

[11] "What Is a Good Internet Speed?" *HighSpeedInternet.com*, 2021, www.highspeedInternet.com/how-much-Internet-speed-do-i-need.

[12] A. J. Henney and W. D. Tucker, "Video Relay Service for Deaf people using WebRTC," 2019 Conference on Information Communications Technology and Society (ICTAS), 2019, pp. 1-6, doi: 10.1109/ICTAS.2019.8703631.

[13] Michael Crabb, Rhianne Jones, Mike Armstrong, and Chris J. Hughes. 2015. Online News Videos: The UX of Subtitle Position. In Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15). Association for Computing Machinery, New York, NY, USA, 215–222. DOI:https://doi.org/10.1145/2700648.2809866

[14] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic Subtitles: The User Experience. In Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '15). Association for Computing Machinery, New York, NY, USA, 103–112. DOI:https://doi.org/10.1145/2745197.2745204

[15] T. Kurahashi, K. Suemitsu, K. Zempo, K. Mizutani and N. Wakatsuki, "Disposition of captioning interface using see-through head-mounted display for conversation support,"

2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), 2017, pp. 1-4, doi: 10.1109/GCCE.2017.8229441.

[16] Behm, G. W., & Kushalnagar, R. S., & Stanislow, J. S., & Kelstone, A. W. (2015, June), Enhancing Accessibility of Engineering Lectures for Deaf & hard-of-hearing (DHH): Real-time Tracking Text Displays (RTTD) in Classrooms Paper presented at 2015 ASEE Annual Conference & Exposition, Seattle, Washington. 10.18260/p.23995