Accessibility Standards Canada

Understanding user perspective of the speed/accuracy/delay tradeoff for captioning fast-paced media

Research Report #17324237

Submitted by:

Deborah Fels, PhD., P.Eng.

Abstract

Closed captioning (CC) in Canada has made great strides since the 1990s but there remains much to do to meet the benchmarks of equity and inclusion set out in the Accessible Canada Act. In this project, the main goal is to examine the tradeoffs between speed, accuracy and delay that have been identified by people who are Deaf or Hard of Hearing (D/HoH) as their top three concerns for live broadcasts. Specifically, in this project, we focus on fast-paced sports as this content is particularly problematic because there are few opportunities for correcting errors. In addition, the advent of more reliable and accurate speaker independent Speech-to-Text (STT) and artificial intelligence (AI) may offer methods to resolve some of these issues. We report on three user studies 1) play-by-play (PBP) versus commentary-only (CO) for two fast-paced sports; 2) captioner workload; 3) user study of PAVOCAT, one software development tool resulting from these studies. Two additional user studies that are a continuation of the original four studies remain in progress as master's theses. The main findings are that: 1) captioner workload for fast-paced live events could be considered high and should be reduced; 2) captioning the CO portions of a live broadcast fast-paced sport instead the PBP announcing and the commentary seems promising as one solution to speed and accuracy issues for CC those games; 3) PAVOCAT that uses an AI system to generate live CC with captioner supervision seems to offer one solution to the delay and speed issues that exist with conventional CC although addition research and development is required to ensure that the system is robust and usable; and 4) captioners do not trust AI captioning. The two remaining studies will investigate the acceptability of commentary-only CC over the long term, and the Deaf perspective on trustworthiness, believability, quality and comprehension of AI-generated CC compared with conventional CC for live, fast-paced broadcast content.

Table of Contents

Abstract 2
Introduction 4
Research Objectives 4
Objective 1. Comparative study 4
Objective 2. Longitudinal study 10
Objective 3. Caption tool enhancement 10
Part 1. Understanding captioners workload 10
Part 2. Design and development of an automatic caption tool 13
Objective 4. Develop recommendations 17
Objective 5. Project findings dissemination 19
Limitations 20
Comparative study limitations 20
Captioner study limitations 20
AI-supported captioner tool usability study limitations 21
Conclusion 21
Acknowledgements 22
References 22

Introduction

Since 1993, providing closed captioning (CC) for all broadcast media content has been part of the licensing regulations of the Canadian Radio-television, and Telecommunications Commission (CTRC). Since that time, there has been an evolution in these standards and regulations regarding the quantity of programming requiring CC, to more recently, addressing the quality of those CC; the latest update to the CC quality standard for live programming occurred in October 2019. In addition, the mandate has been broadened to online content that remains under the purview of the broadcasters, such as streaming on broadcaster websites. A common complaint, however, has been the lack of the users' perspective in defining the assessing quality that can then inform these standards and regulations. With the introduction of new technologies including automatically generated CC, speaker-independent speech recognition and artificial intelligence (AI), traditional production and quality measures are being redefined and challenged, particularly for live content. Towards addressing these gaps and challenges, we are reporting the results of a series of studies and a software development project funded in 2020 by Accessibility Standards Canada and based on an earlier survey of D/HoH viewers on their priorities in defining the quality of CC for live television. The current set of studies explores user viewing behaviour and experience of live CC in two fast-paced sports (hockey and basketball), captioner workload and user perceptions of AI-generated versus human-generated CC for fast-paced live sports. All studies are focused on the speed/accuracy/delay trade-off as it poses a considerable challenge in providing high-quality and meaningful CC. Unavoidable delays in the production processes and speaking rates (e.g., play-by-play (PBP) announcers that are considerably faster than either CC production or a viewer's reading abilities result in errors and low-quality CC that interfere with a D/HoH user's comprehension, enjoyment and equitable access to the content. For example, the recommended maximum CC rate is 180 words per minute (WPM) ((Jensema, 1998; Ofcom, 2005; Szarkowska, 2013) but often the PBP announcing rate can be well above 220 WPM. Users are unable to read CC at the rate and watch and understand the game simultaneously. Automatically generated CC using Speech-to-Text (STT) and AI currently available for video conferencing applications, such as Zoom, may be one solution to these issues with fast-paced content but, these types of CC may introduce other issues that only amplify the readability and understandability of the CC. Finally, STT and AI for CC live content requires tools that integrate the skills that a human can contribute with the efficiencies of an AI system to evaluate the respective preferences and performance of the different CC generation methods.

In this report, we present the research that was conducted for the objectives outlined in the research proposal funded by Accessible Standard Canada. It should be noted that there were considerable delays in planned face-to-face research due to COVID-19. Also, the figures and tables remain in English as the study information, such as the theme table (Table 2), was only in English and used English-only content.

Research Objectives

Objective 1. Comparative study

Carry out a comparative study with a set of modified caption accuracy factors and assess the user experience of D/HoH participants for fast-paced live media content for three sports and/or Olympic events with male and female teams, and two talk shows (1 general interest and 1 sports talk show in English). The factors that will be adjusted include PBP announcing and

commentary versus commentary-only CC for sports, removing high-frequency versus low-frequency words, and quantity of paraphrasing.

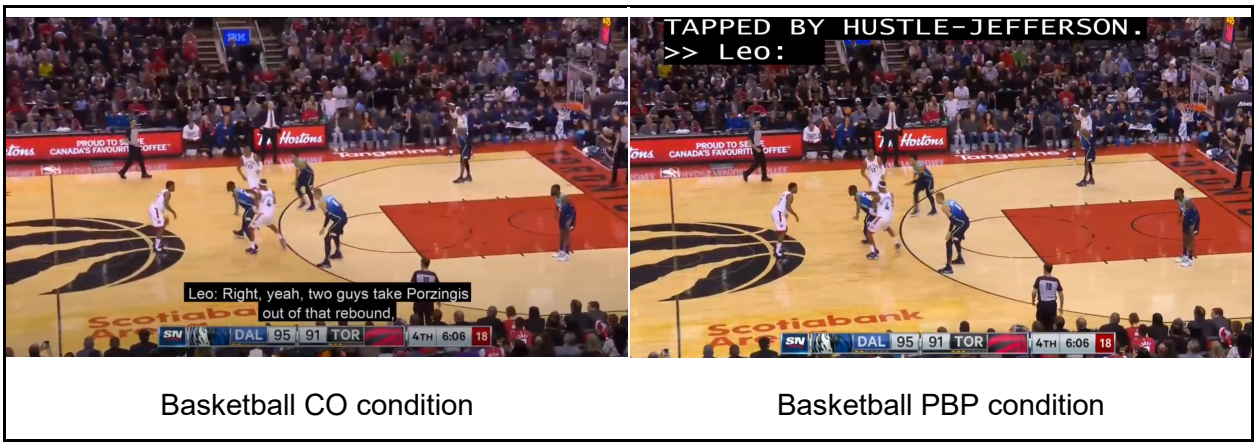
A study was designed as a mixed-methods study, combining precise eye-tracking data for participants taking part in the study with qualitative preference and comprehension data. The study design was approved by the Ryerson (now Toronto Metropolitan University) Ethics Board. Participants watched the video segments for the sports, and subjective ratings of quality, engagement and interest were collected through questionnaires. In addition, a conversation-style interview was conducted with participants to determine their comprehension of the game and non-gameplay information; key events that occurred during the gameplay (e.g., penalties and goals) and comments that were made by the announcers about non-gameplay information were discussed. The study included two segments of the CC games (either hockey or basketball), with each segment presenting a different CC condition to the participants. The entire study took about 1-hour for each participant to complete and was conducted at the Toronto Metropolitan University campus in one of the usability labs. Participants were given an honorarium of $60 plus transportation costs for their participation.

This study introduced a novel form of CC style as an approach to reducing the number of words presented, without reducing the accessibility of the information. To illustrate how the different CC styles appeared on the screen, Figure 1 shows that during a PBP call, there is no caption present in the commentary-only (CO) condition. When the gameplay pauses/stops, the CO captions are displayed, as per the commentary taking place during the game. The PBP captions are always displayed whenever one of the commentators is speaking. Note that in the images below, the CO captions are centred on the page, while the PBP captions on the right are spread across the entire top part of the screen.

Figure 1: On-screen appearance of CO captions and PBP captions during gameplay.

Hockey CO condition | Hockey PBP condition

Figure 2: On-screen appearance of the CO and PBP captions during commentary.



Basketball CO condition | Basketball PBP condition

Using the Tobii eye-tracker, the duration and frequency of visits in the Areas of Interest (AOI) were captured, where each visit was defined as the sum of fixations, saccades, and glances in milliseconds. The AOI were defined as the areas containing the captions and the areas not containing captions (e.g., gameplay area, stands, scoreboard, etc.). Based on the eye tracking records, it was possible to perform statistical analysis, which revealed that there were no statistically significant differences in the duration and frequency of visits between CO and PBP captions regardless of the participant's hearing status.

The comparative study was completed successfully with 16 Deaf and 11 HoH participants. The duration for visits for CO captioning was 11% and 10% for PBP captioning. Although statistical significance was not found, Figure 3 shows the average values for D/HoH views for each caption condition. Participants seem to visit the CO captions more than PBP captions regardless of the sport or hearing status. One way to interpret this would be that PBP captions interfere with watching the gameplay so that participants did not spend much time on or visit the area of the screen at all regardless of whether there were captions for PBP or for the CO. Gameplay CC were not present in the CO condition so participants may have thought that these CC were worth reading. The average number of visits for the CO captions was 124, whereas the average number of visits for the PBP captions was 106 (see Figure 4).

Figure 3: Mean and standard deviation (lines) of the duration for visits on the two AOI for CO and PBP conditions.

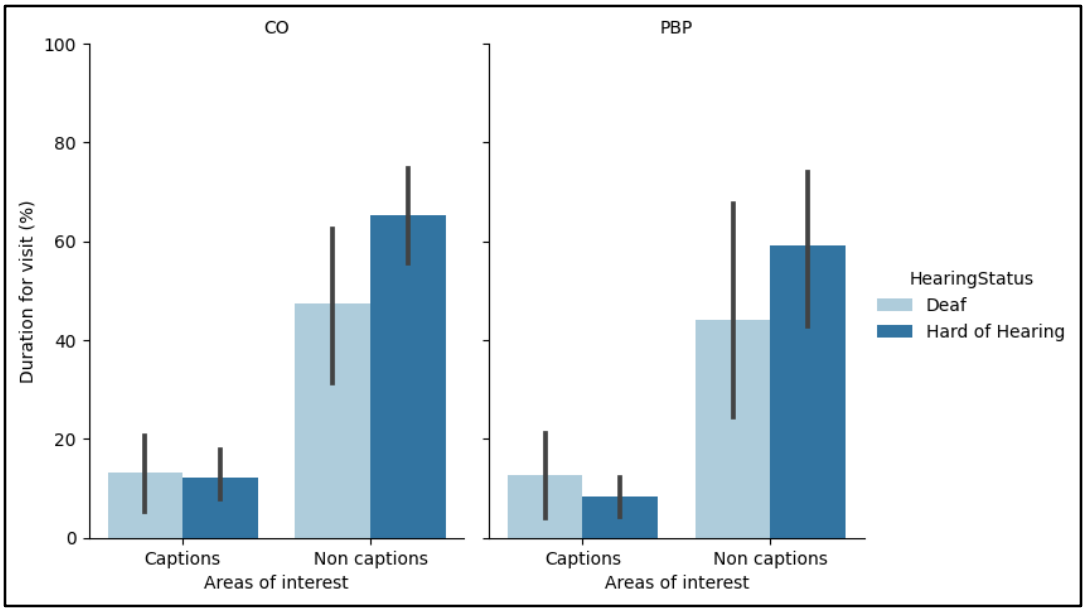
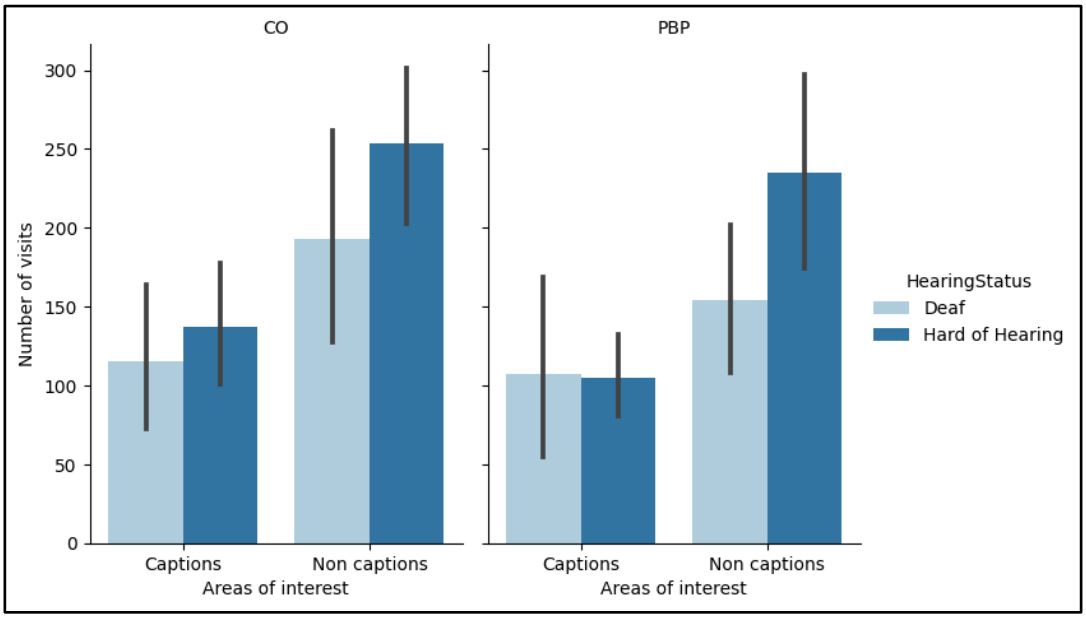


Figure 4: Mean and standard deviation for the frequency of visits on the two AOI for CO and PBP conditions.



Given that the PBP condition represented conventional CC that participants were used to watching, finding no statistical difference between the two conditions revealed the potential for using CO captioning as the novelty effect likely influenced viewer experience negatively (e.g., “this is different from what I am used to”). This potential appears to be more evident from the positive participant feedback on CO captioning compared with the negative feedback on PBP captioning after viewing the two conditions.

“I would prefer the second clip's captioning [referring to CO]. It was good because I understood the game. I do not need the play-by-play. When live play is happening the commentators are not even talking about that, so I get distracted easily. I like [CO] because there were no captions during the live play and because I can't hear, I cannot tell if they are talking, and I could catch up on what was going on when live play was dead (replays and timeouts).”

Another participant described how CO captioning helped follow the game without being worried about missing out.

“I could concentrate on the game and not worry that I was missing out on something by trying to keep up with both (...) It was at an eye level that even when the play was starting, I could finish reading and still see what was happening in my peripheral [vision].”

For PBP captioning, participants mentioned how difficult it was to follow the game and the CC displayed on the screen. Some of the participants' comments about PBP captioning were:

“too distracting” / “stressful, too much.”

“seriously overwhelming” / “a lot of work for my eyes” / “ongoing.”

“a lot of visual back and forth [looking between the captions and game screen.]”

“feel[ing] a little stressed because I do not want to miss the important information before it disappears.”

“It's like reading a book and when I looked up, I would miss the captions.”

Some participants also added how the conventional captioning [PBP] resulted in them avoiding the CC as a whole and going *“straight to watching the game.”* The comments from participants support the aspect of interpreting duration for visits in the non-caption area, where participants want to focus on watching the actual game visuals and play, with fewer distractions if possible.

However, not all comments for CO were positive and not all comments on PBP were negative. One participant preferred PBP captioning because she was used to watching PBP captioning, especially because she “knew that everything was on it”. Other comments referred to how CO captioning missed the player identification/name and numbers as participants used the CC to recall names.

From the questionnaire responses, most (65%) participants selected no preference when asked about which CC condition they preferred. The remaining participants preferred CO captioning (27%) more than PBP captioning (8%). Sixty percent of the D participants and 73% of the HoH participants showed no preference. In the group that stated a preference, about 33% of the D participants and 18% of the HoH participants preferred CO captioning, and 7% of D and 9% of HoH preferred PBP captions. The perceived quality, viewing experience, and CC helpfulness ratings after watching the clips were higher for the CO condition than the PBP condition, although these differences were not statistically significant (see Table 3).

Table 1. Descriptive statistics of the overall satisfaction and viewing experience rating after watching the clips in each CC condition**.**

Sports	Watched condition	Hearing Group	Caption satisfaction Mean (SD)	Viewing experience Mean (SD)

Basketball	PBP	Deaf	3.75 (0.96)	4.50 (0.58)
		Hard of Hearing	2.86 (1.35)	4.13 (0.64)
	CO	Deaf	3.75 (0.96)	4.25 (0.96)
		Hard of Hearing	3.63 (1.19)	4.38 (1.06)
Hockey	PBP	Deaf	2.91 (1.38)	3.45 (1.13)
		Hard of Hearing	3.33 (1.53)	3.00 (1.00)
	CO	Deaf	3.36 (1.75)	3.73 (1.27)
		Hard of Hearing	2.33 (1.53)	3.33 (1.53)

This study was a first of its kind in Canada and has revealed some insights into the user experience of watching fast-paced live sports CC for hockey and basketball. While the industry and viewers know the need to address the general live CC issues, such as synchronization delay, accuracy, speed of captions, speaker identification, caption placement, and readability, a specific aspect of live CC for fast-paced sports involves the challenges of generating and reading the high number of WPM. In this study, we introduced a new CC style for live sports broadcasting, which only provided CC for commentary with fewer words than the conventional PBP style.

Results from the eye-tracking portion of the study showed that there were no statistically significant differences between the two CC styles, however, the descriptive analytics revealed the difference in duration and frequency of visits, where participants spent longer watching and visited more frequently on CO captioning. Furthermore, participants' responses in CC quality satisfaction, viewing experience, and CC support in understanding, showed evidence of preferences for the CO captions. In addition, while many reported no preference for the style of CC, participant comments allowed a deeper probe into the reasoning behind subjective preferences. It was surprising to find that participants had a considerable amount of positive feedback for CO captioning from this short study despite the study being skewed towards PBP because it was, up until now, the only type of CC available to viewers.

Finally, the findings from this study provide sufficient evidence to pursue the notion of the 'colour commentary' only style of CC in a longitudinal study to determine whether CO captioning is viable and whether viewers are satisfied after overcoming the novelty of this style of CC. In addition, a longitudinal study could indicate whether CO captions sufficiently reduce some of the tradeoffs between speed, accuracy and delay in fast-paced live sports CC and improve the viewer experience.

A second study consisting of an ongoing master's thesis project, that will extend beyond the completion of this project, has been designed to investigate the impact of automatic CC for fast-paced sports that are less familiar to viewers, such as those from the Olympic games. In this study, CC will be produced

using speaker-independent speech recognition technology, such as Google's Speech-to-Text AI (Google Cloud Speech-to-Text, 2024). Participants will watch four clips for four different winter Olympic high-speed sports such as the Women's Half-pipe which will have a randomly assigned CC type, either human-generated or automatic generated. Five factors, believability (Bowers & Phillips, 1967; Simpson & Kahler, 1981), satisfaction (Bhattacherjee, 2001) trustworthiness (Bowers & Phillips, 1967), quality rating and comprehension, will be measured using standardized metrics and a comprehension method developed in the first comparative study.

Objective 2. Longitudinal study

Carry out a longitudinal study (Study 2) with 30 D/HoH viewers for user preferences indicated from objective 1 to understand the longer-term effects on user experience and quality assessment. It is anticipated that this will be a one to three-month study per participant.

As the first comparative study results revealed a potential to use the CO version of CC to be preferable to D/HoH audience groups compared to the conventional PBP captioning, a longitudinal study investigating the longer-term exposure of CO captioning has been designed for a second master's thesis, currently underway. Unlike the first study, this longitudinal study will aim to provide participants with a more extended period of exposure (20 minutes per day over three days) to CO captioning for fast-paced sports. It is expected that individuals will have sufficient time to habituate to watching with fewer words. The ethics review application was prepared and submitted to the University of Toronto's Research Ethics Board.

Objective 3. Caption tool enhancement

Develop and evaluate CC tool enhancements that allow CC to be generated according to the results of objectives 1 and 2.

Part 1. Understanding captioners workload

In addition to understanding the CC viewers perspectives, a study investigating the CC perspectives was designed. The purpose of this study was to assess the subjective mental workload (SMW) of live CC. The research questions for this study are:

1. What is the SMW experience of live captioners who caption different live-content genres;
2. What is the impact of paraphrasing on the SMW?

Methodology

Following university ethics approval, participants were asked to complete a 55-question online survey containing 21 paired weighting and six rating questions comprising the standard NASA-TLX Task Load Index instrument for evaluating SMW (Hart & Staveland, 1988), four demographic questions, such as gender and education, 12 questions about training and general experience with live CC and 10 questions about attitudes and preferences for paraphrasing in live CC. The NASA-TLX comprises six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration) to evaluate the mental workload of an individual performing a task. Each dimension is rated on a scale from 0 to 20. In addition, there is a weighting component that asks people to rate the importance between pairs of all six dimensions.

A set of interviews were conducted to collect further qualitative data with people who consented to participate. Interviews were semi-structured and asked questions about workstation setup, challenges, and their opinions about paraphrasing.

Participants

Thirty respondents (27 women and 3 men) completed the entire survey. The participant group represented four different countries: Canada (16 of 30), the United States (11 of 30), Australia (2 of 30) and one from the European Union (1 of 30). Age groups ranged from 18 to more than 60 years old. The education levels varied from high school to post-graduate degrees. The majority of participants (17 of 30) had more than eight years of live CC experience, where they captioned a variety of live programming including broadcasted news, weather or entertainment, fast-paced sports (such as ice hockey and soccer), slower sports (such as golf(, sports talk shows, day or night time talk shows, and captioned other types of live shows (such as government or business meetings) and live religious events. Finally, 23 of 30 respondents used stenography for live CC, and seven used re-speaking. Five participants agreed to participate in the follow-up interview session.

Results

The total SMW was determined as a score out of 100, which was calculated using the mean of weighted ratings. The weighted adjusted ratings for the six dimensions were calculated by multiplying the raw ratings by the corresponding weights.

The total SMW per participant ranged from 16.7 to 91.5 (M = 62.87, SD = 24.04). Figure 5 shows the means and standard deviations between the six subscales that make up the total SMW.

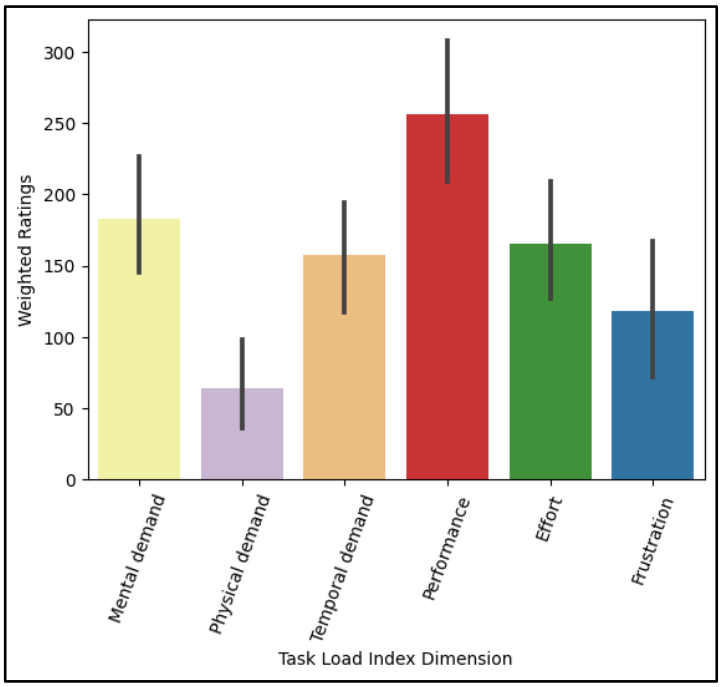
The perceived workload of live captioners is considered high compared to the literature benchmarks (Grier, 2015). Live CC's SMW ranks in the top 25% compared to a wide array of job types, including those in the medical field. Performance was the most significant factor influencing SMW. This was followed by mental workload, effort and the need to meet deadlines, with physical demand having the least impact. The minor role of physical demand is expected as tasks, like stenographic keyboarding or re-speaking, become more routine and automatic with practice. The emphasis on performance as a key factor in mental workload is largely due to the stringent standards set by government regulations, broadcaster clients and company policies. Achieving these standards at high speeds is challenging and can significantly heighten captioners' stress levels, especially when coupled with the pressures of timely delivery.

“Error responsibility is more to stenographers (...) and I try to be 100% correct.” - P1
“Once it’s an error, you can’t take it back.” - P2

During the interviews, participants indicated that their mental stress was primarily due to the obligation of providing accurate, word-for-word CC. They were aware that the quality of their CC played a crucial role in ensuring D/HoH individuals' access to broadcast media and that any lapses in quality could lead to complaints. These complaints are directed at broadcasters and can impact their compliance with governmental standards, potentially harming the captioner's reputation and future work opportunities.

Another crosstab analysis was performed to determine the association between SMW categories and captioner employment status. There was a statistically significant association between the SMW category and the employment condition; $\chi 2(4, 30) = 10.72$, $p < .05$, $\varphi = .423$. A follow-up Kruskal-Wallis test indicated significant differences in total SMW scores for each employment category, $\chi^2(2) = 7.015$, $p < 0.05$ and Dunn's test using Bonferonni correction revealed a significant difference in SMW scores between the freelancer and company groups: Freelancer/self-employed (M=48.59, SD=26.17), company employment (M=68.61, SD=22.31) and captioners who were freelancers and employed by a company (M=76.18, SD=10.65). This finding revealed the association between employment status and SMW clusters, where freelancers seemed to have lower SMW scores compared to the captioners employed by companies or agencies. This could be interpreted as a phenomenon where captioners who control their schedules tended to perceive less stressful workloads than the ones who were obligated to follow assigned tasks. Other factors influencing captioners' stress included job finding, task allocation and scheduling and preparing for the live events, including their dictionary mapping for names and unique terms.

Figure 5: Mean and standard deviation of the weighted ratings for each of the NASA-TLX dimensions. The weight of each dimension was acquired from the trade-off questionnaire and was multiplied by the corresponding raw rating (ranging from 0 to 100).



Caption Paraphrasing

From the survey, we asked participants if their work policy allowed them to paraphrase or not. Most participants reported that their work policy allows paraphrasing (21/30), a few perform paraphrasing as needed even if it is not allowed (6/30), and the remaining participants (3/30) were not allowed to paraphrase. From those who reported that they paraphrased, the majority paraphrased ‘only when necessary’ across all genres, such as fast sports, talk shows, weather forecasts and news. Some participants reported that they ‘always paraphrase’, which was found for all genres excluding slow

sports. For slow sports, 14 participants ‘never paraphrased’ whereas, the remaining 13 reported that they ‘paraphrased when necessary’. Lastly, a majority of participants reported that they believed that paraphrasing improved the quality of CC ‘somewhat’ (12/27) or ‘a lot’ (4/27). Seven participants reported the paraphrasing impact was ‘minimal’ (8/27) or made ‘somewhat more difficult’ (making CC lower quality) (3/27). A related question was, “*for what reason do you omit words from your live captions?*”, which a majority of participants answered 'to keep up with the speed' (24) followed by 'to reduce redundant speech' (11), and 'when paraphrasing' (9).

A crosstab analysis using the Chi-square test for independence was performed to determine the association between SMW high, medium and low categories and paraphrasing use for the different content genres. There was a statistically significant association between fast sports CC paraphrasing frequency and the SMW category; $\chi^2(4, 27) = 11.25$, $p < .05$, $\varphi = .456$. No other genres and their paraphrasing use frequency were found to be statistically significantly associated with the SMW categories.

Keeping up with the high-speed speaking rates was the main justification for using paraphrasing by study participants. As survey responses regarding CC paraphrasing frequency by genre suggest, the faster speech rate the greater the frequency of paraphrasing. Another common reason for paraphrasing mentioned by participants is when multiple speakers are talking simultaneously. Live CC are usually displayed line-by-line as speaking occurs, called roll-up captions, and occasionally they are pop-on type depending on the broadcaster. When more than one person speaks at a time, captioners must decide what words to CC and how to paraphrase so that they can try to CC more of the speech. Not only is this a difficult auditory task but a new tracking task; keeping track of who is talking is then added to a captioner’s mental workload. They need to reduce the number of words to be CC when they could not catch up with the spoken words, regardless of regulations or policies. In addition, listening and remembering what was heard appeared to be challenging. This would suggest that the paraphrasing task adds more stress and mental workload on captioners.

> *“If it's a reasonable rate of speed, I will capture them all, but other than that, people are speaking too fast and they talk at the same time. It is impossible to caption without paraphrasing.”* - P3

This research has revealed the contribution of the different mental workload factors, which influence live closed captioners' SMW. The findings provided initial evidence to support the need to reconsider live CC tasks due to the mismatch between audience preferences, regulatory requirements, and the high mental workload captioners experience.

Part 2. Design and development of an automatic caption tool

PAVOCAT

In order to mitigate some of the difficulties with live CC, an automatic caption tool, PAVOCAT, was developed to support the captioners in generating live and post-production CC by using machine learning algorithms. It is a live CC software tool that uses AI technology to produce text CC, monitors for errors and provides suggested corrections. The captioner then is responsible for supervising PAVOCAT and selecting error corrections. If the captioner determines that the suggested corrections

are insufficient, they can make corrections via a steno keyboard, a QWERTY keyboard or through re-speaking.

The system is composed of three main elements:

1) User interface and functionality for the captioners and broadcasters for scheduling programs requiring CC, assigning and accepting caption jobs, communication between broadcasters and captioners, system status tracking and monitoring the AI system and correcting errors (see Figures 6a and b). PAVOCAT produces the initial CC along with a list of errors it thinks exist in those CC. It will also provide two possible recommendations to correct each error in real-time. Captioners can upload their dictionaries with specialized terminologies and names, etc. for various programs which, will be maintained through the Carl database;

2) Carl is the backend application programming interface (API) that receives the audio track from the broadcaster, transcribes it into text, determines errors and provides suggestions and sends it to PAVOCAT to display; and

3) PAVOCAT is the signaling server that establishes and manages the signal between captioners and broadcasters. Figure 7 provides the PAVOCAT system diagram of the three elements and the signal processing that occurs between those elements.

Figure 6a: Captioner view of PAVOCAT interface.

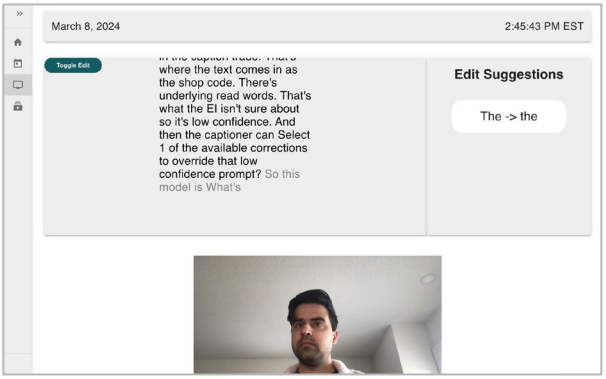


Figure 6b: Broadcaster view of PAVOCAT interface.

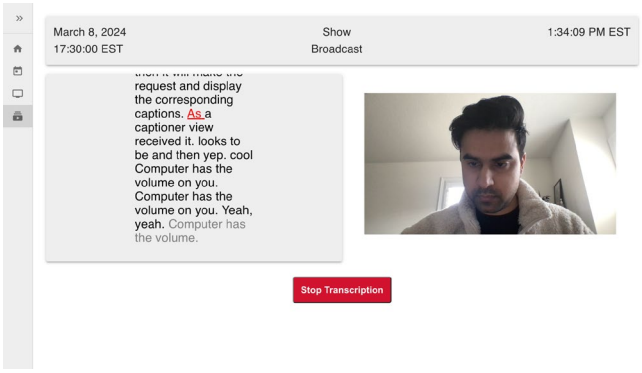
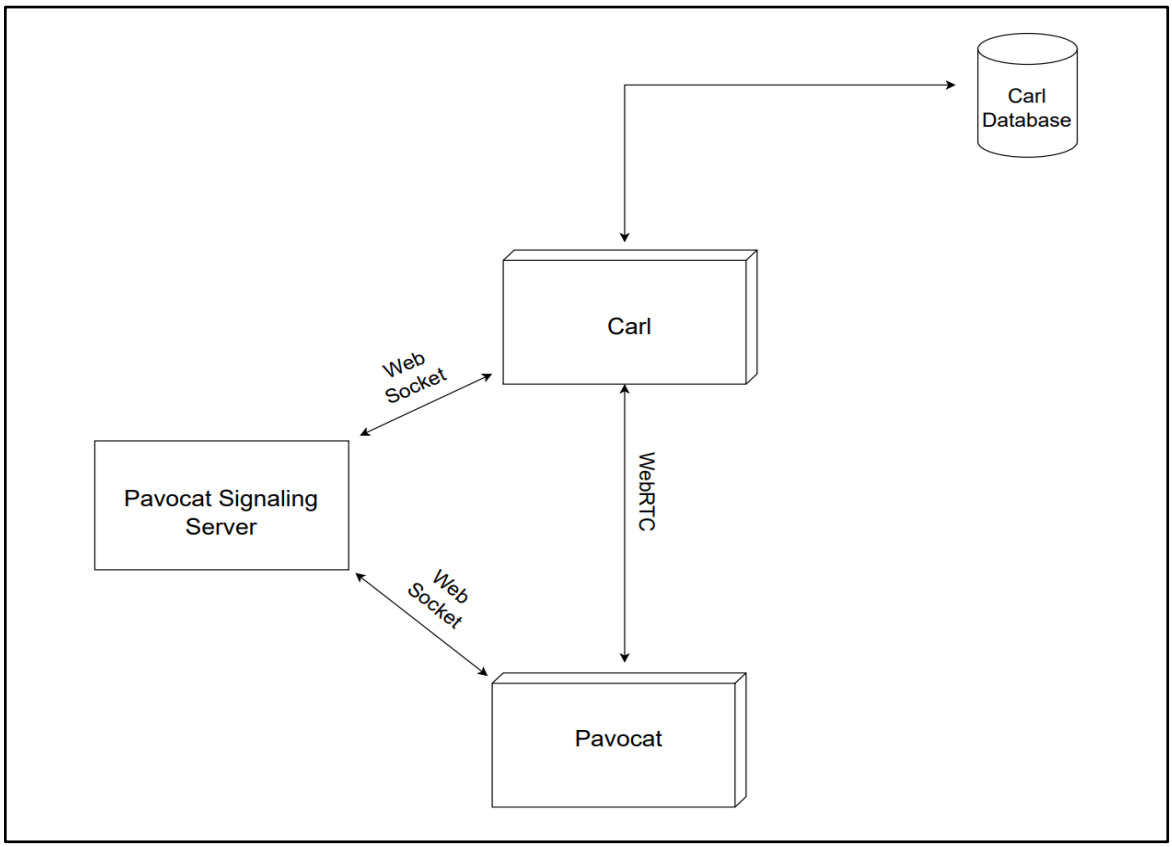


Figure 7: PAVOCAT elements and signal processing flow.



User study

Following university ethics approval, a user study with novice and experienced professional captioners was designed and conducted with the first fully working prototype of PAVOCAT.

The purpose of the study was to understand captioners' experiences when using an AI tool to aid in their CC.

The research questions for this study are:

1. How do novice and expert captioners work with a novel AI tool in a "captioner as overseer" mode?
2. What is the experience of captioners working with the novel AI tool PAVOCAT?

Participants were presented with three 5-minute fast-paced video clips from the Winter 2022 Olympic games over a 5-day period where PAVOCAT automatically produced the live CC, monitored for errors and then provided correction options for detected errors from which the captioner could choose. No typing was allowed. Our hypothesis was that the role of live captioners can be modified such that the live captioners act as supervisors to or overseers of AI tools, rather than acting as operators (stenographers and re-speakers).

Experiment Design

A time-series experiment was conducted with 21 participants. Ten novice captioners (1 female) with less than six months of experience as a captioner, and 11 experienced captioners (5 female) with more than six months of experience. Participants completed a 19-question pre-study questionnaire, participated in a training and practice session to learn how to use PAVOCAT, and used the PAVOCAT software to caption three 5-minute segments (performed once per day over five days) of Olympic games content. After completing each ten-minute captioning session, the participants completed a 25-item questionnaire consisting of four merged validated tools. The NASA-TLX (Hart & Staveland, 1988) was used to assess SMW. The System Usability Scale (SUS) consists of 10 Likert-style questions which measure ease of use, ease of learning, error correction and satisfaction (Brooke, 1996). The Trust in Automation scale (Jian et al., 2000) consists of a 12-item questionnaire which rates levels of trust felt by the participant with the system, over a seven-point scale. The Satisfaction scale is a one-question item which explicitly asks if the participant felt satisfied with the system (SUS does not *explicitly* measure satisfaction).

At the end of the study each participant completed a 30-minute semi-structured interview expressing their opinions and experience with the PAVOCAT system and the growing impact of AI and Automatic Speech Recognition (ASR) on the field of captioning.

Table 2: Themes and modifiers for PAVOCAT user study.

Themes	Modifiers
Performance and Expectations Performance and quality of PAVOCAT's automated captions	**Positive** Captioning was accurate and/or the system's performance or the participant's familiarity with the system improved over time
	Negative Captioning was not accurate and/or the system's performance or the participant's familiarity with the system did not improve over time
Other (Video Content and Technical) Content chosen for captioning, including speed of captioning provided, proper names, languages, accents or other linguistic features affect AI captioning accuracy Technical issues with captioning in general as well as with the system	
User Interface Modifications Suggestions and omissions of the interface such as control over audio/visual elements including, layout, organization and other adjustments	
Physical and Mental Workload Describing the stress and ability to stay aware and in control of the PAVOCAT environment compared to their current stenography/respeaking work Note: If physical workload was positive and mental workload was negative then that comment should be coded in BOTH positive and negative modifiers	**Positive** PAVOCAT was not physically or mentally stressful to operate
	Neutral/No Change Physical and mental demands of PAVOCAT were as stressful to operate
	Negative PAVOCAT was more physically or mentally stressful to operate
Comparison with Other Explicitly Named AI/ASR Software Comparisons of PAVOCAT with other AI/ASR software	
Willingness to use AI Participants' trust in AI and any changes in trust during the study	**Positive** Participant's trust in AI to delegation of captioning tasks was either high or improved
	Neutral/No Change Participant's trust in AI or opportunity to delegate captioning tasks is neutral and/or did not change as a result of the study
	Negative Participants' trust in AI or delegate captioning tasks is low and/or reduced as a result of the study OR concerns about losing employment to AI
Human AI Interaction How humans and AI will or will not work together	**Override Typing** Participants would like to option to override the AI's captioning and respeak/type instead
	Override Edit Suggestions Participants would like the option to provide Edit Suggestions back to the AI or force the AI to remember or dismiss Edit Suggestions

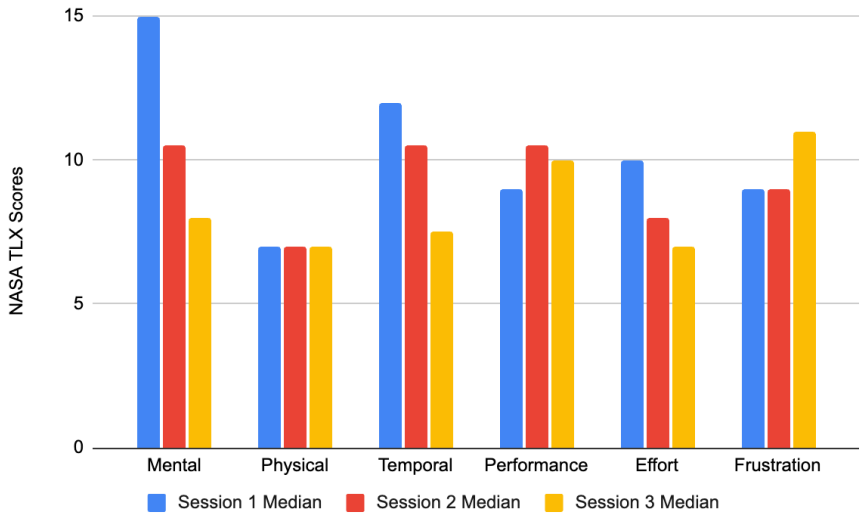
Data analysis

The post-study interviews were analyzed using thematic analysis via Noldus Observer (Noldus, 2024). The entirety of the study was recorded, so observations and video and audio recordings could be analyzed.

Results

Non-parametric Friedman tests were performed on all measurements (NASA-TLX, SUS, Trust in Automation and Satisfaction) in aggregate and using cross-tab analysis for novice/expert classifications. Mental workload saw a statistically significant reduction between the three sessions, M=10.33 (SD = 5.46) for session 1, M=9.25 (SD = 4.78) for session 2, and M=8.42 (SD = 5.17) for session 3. The other factors of NASA-TLX were not found to be statistically significant but did show improvements over the duration of the study. The overall SUS scores yielded a median of M=57.5 (SD = 18.75) for session 1, M=55 (SD = 17.30) for session 2 and M=52.5 (SD = 12.93) for session 3. The Trust in Automation scale yielded an overall median of M=32 (SD = 11.75) for session 1, M=33 (SD = 11.39) for session 2 and M = 39 (SD=10.28) for session 3. The Satisfaction scale yielded an overall median of M=5 (SD = 1.34) for session 1, M=4 (SD = 1.56) for session 2 and M=4.5 (SD = 1.54) for session 3. Thematic analysis is still in progress and is undergoing an inter-rater reliability verification process.

Figure 8: NASA-TLX sub-factor scores.



Objective 4. Develop recommendations

Develop recommendations with appropriate empirical evidence and examples for supporting CC standards and regulation improvement as one policy brief. This will also include the impact on various distribution technologies and processes.

A preliminary set of recommendations was developed based on the results of user studies designed to understand the perspectives of captioners and viewers. Findings revealed evidence and potential solutions to improve CC standards and regulations. These recommendations were written in technical and academic reports for dissemination. Future work could involve presentations to the Canadian Radio-television and Telecommunications Commission (CRTC) on the findings of this research for the next round of CC standard hearings.

From the PBP caption user study conducted with the CC viewers, we recommend:

1) A reduced number of words in CC may provide a better user experience from viewing live CC for fast-paced sports broadcasting. A longitudinal study is needed to confirm these results.

   Tentative recommendation: Caption the commentary portion of fast-paced sports and do not caption the PBP announcing during gameplay.

Based on the findings from the captioner workload study, PAVOCAT user study, and principles of ergonomics and industrial engineering to reduce the subjective workload of live captioners, improve their job satisfaction, and enhance their overall work performance, the recommendations are:

1. Support Collaboration: A captioner team working in pairs can support as well as offer breaks to each other. A standard of practice for live language translation services, including sign language interpreters, is to work in pairs (i.e., Booth-mates) to support each other by monitoring for errors and providing breaks (Syahputra et al., 2017). Forming a group of two (or more) live captioners for each session may help reduce their SMW. A common argument against employing more than one captioner is the cost of paying two people instead of one, or job sharing. Captioners tend to be fairly low-paid and are often unwilling to job-share, particularly as freelancers. In addition, there is no professional organization dedicated to advocating for captioner rights or developing workplace norms. To change the culture of live CC work, advocacy and more leadership are required.
2. Schedule Breaks: Having regularly scheduled breaks is another common recommendation to reduce mental workload (Boucsein & Thum, 1997; Krueger, 1989). In live broadcasting, commercial breaks provide that opportunity as commercials are CC as post-production content so that the live captioner is not required to CC them. A typical commercial break lasts between 2-3 minutes and they occur about every 10 minutes in a broadcast segment (Bollapragada & Garbiras, 2004). The duration of breaks needs to be determined after investigating the relationship between cognitive recovery, work demands, resources, and stress (Hobfoll, 2001; Meijman & Mulder, 2013). However, with live sports broadcasting, commercial break occurrences and length vary, depending on the gameplay.
3. Optimize Work Schedules: Schedule work in a way that takes into account the cognitive load and processing requirements of live CC. Avoid scheduling back-to-back sessions and aim to provide adequate recovery time between sessions. It is a known challenge that scheduling and worker allocations have an impact on SMW (Martens et al., 1999; Trumbull, 1966). Similar to other types of work, live captioners' SMW may be mitigated by modulating schedules and having balanced work allocations.
4. Ensure Flexibility: Encourage flexibility in work arrangements to enable live CC workers to balance work and personal commitments. This could include flexible working hours, remote work options and job-sharing arrangements (Russell et al., 2009).
5. Provide Adequate Resources: Ensure that live CC workers have the resources they need to perform their jobs effectively. This could include broadcasters providing research information pertinent to the topic before the broadcast (e.g., squad information of live soccer game, guest information of talk shows, etc.), advanced CC software, or even hardware instruments (such as high-quality headphones), ergonomic chairs and desks and appropriate lighting.

6. Training and Support: Provide adequate training and support for live CC workers to help them manage the cognitive demands of their jobs. This could include training in keyboard shortcuts and other productivity tools, as well as training in stress management and relaxation techniques.
7. Feedback and Continuous Improvement: Encourage feedback from live CC workers and use this feedback to improve the work environment and processes continuously. Broadcasters, freelancers and CC companies should regularly review work schedules, job demands and available resources to ensure that their workers are working under optimal conditions.
8. The combination of TTS, AI and captioner as supervisor may reduce the SMW and delays and improve accuracy for fast-paced live television. However, additional research and development are required to, not only provide sufficient evidence for this claim, but also put in place fully functional and supportive tools to achieve high-quality live CC.

Objective 5. Project findings dissemination

The project findings were disseminated through two peer-reviewed journals or conferences (one peer-reviewed submission will be to an open-source journal that is freely available to all with no need to access a library; open-source journals have a publication cost to authors who have been listed in the project budget), one trade publication and one media release.

The results of the captioner workload study have been published in one journal and one conference venue. Note: students are indicated in bold.

1. Nam, S., Karam, M., Christelis, C., **Bhargav, H**., Fels, D.I. (2023). Assessing Subjective Workload for Live Captioners. *Journal of Applied Ergonomics*. *113,* 104094.
2. Karam, M., Christelis, C., Hibbard, E., **Leung, J., Kumarasamy, T**., Whitfield, M., Fels, D.I. (2022, July). Workload evaluations for closed captioners. *ICCHP 2022*.

The comprehension methodology paper has been submitted to one conference.

1. **Kumarasamy, T.**, Nam, S., Fels., D.I. (2024) Using a Novel Conversational Method for Measuring Comprehension of D/HoH Viewers Consuming Captioning Content for Entertainment Purposes. *ICCHP 2024*.

Results from the PBP study have been published in one trade publication and submitted to one journal:

1. Nam, S. Karam, M., **Kumarasamy, T**., Whitfield, M., Fels, D.I. (April 2023). Colour commentary versus gameplay captions of live fast-paced sports for Deaf and Hard of Hearing television viewers. Live Subtitling and Accessibility Symposium.
2. Nam, S. **Kumarasamy, T**., Karam, M., Whitfield, M., Hibbard, E., **Leung, J**., Fels, DI. (2024 submitted). Eye gaze behaviour and comprehension of color commentary and gameplay captions of live fast-paced sports for Deaf and Hard of Hearing television viewers. *Behavior and Information Technology.*

It is expected that the master's students will publish their thesis work in appropriate academic venues.

The research reports are available on the Live Captioning Canada website: https://www.livecaptioningcanada.ca/

Limitations

Comparative study limitations

This study produced some insightful results that could be considered for the future of CC for live fast-paced sports broadcasts. However, there were a few limitations that should be noted. The first limitation is the small sample size, which may have resulted in insufficient data to show conclusive differences between conditions and/or participant groups. There were also a number of different issues with the eye tracker and resulting data. For instance, it was easy to lose the eye tracker calibration due to head or body movements, which would then require that the study be paused and the system re-calibrated. In addition, the system did not consistently capture the same quantity of eye movement data for each participant. While some of these limitations could be addressed with normalization, some could not. The eye-tracking data were not further investigated through interviews to infer cognitive processes through self-reports, or to explain why participants were looking at something or what they were perceiving. There were also many different types of data (e.g., pupil diameter) that could be captured from the eye tracker and we only used two data types: frequency and duration of visits. These other eye-tracking data could provide more detailed indications of some of the viewing behaviours we found (such as how much attention was being paid while looking at a specific area that can be indicated using pupil diameter). Further studies that capture and use these other data along with complimentary techniques, such as probing interviews, may be required to understand detailed viewing behaviour. Participants only watched a single period of the games so the novelty effect of the CO captions may have not been overcome. In addition, many participants were not only familiar with the game, but they were also fans who knew about the teams, players, referees and game rules, which may have, in turn, influenced their knowledge, interest in watching the game and willingness to overlook errors or difficulties. Further research should consider longitudinal studies using entire games and from other, perhaps less-known sports, such as women's hockey or rugby. Lastly, there were limitations regarding the translation of ASL to spoken English, some translations may not have accurately reflected the meaning of what D/HoH participants were trying to convey. Encouraging Deaf researchers to participate in this research may assist in mitigating the potential translation issues.

Captioner study limitations

Our study revealed some insightful findings for the work environment, workflow, and workload of live captioners. However, there were a few limitations. The first limitation would be the sample size, which limits the generalizability of findings. Captioners were only asked to rate their SMW on their last live CC job; which was different in time, genre and job type between participants. A future study could have captioners 'live' CC the same content sample. However, even controlling for content type and length will incur differences in experience, familiarity with the subject matter, and comfort with the topic. Subjective measurements, such as SMW, can be affected by individual differences, and cognitive and social desirability biases. One finding from our study was the difference between genres in terms of frequency of use. Given this evidence, it would be beneficial to study stress levels measured in each genre. In addition, while the original NASA-TLX aims to measure a single task, many previous studies used the NASA-TLX for multi-task activities. The results from the interview provided evidence of live CC

being a multi-task situation, however, further studies on sub-tasks of live CC can be conducted. Lastly, user studies can only provide a snapshot of the participants' experience at a specific point in time. In our study, we asked participants to recall the last CC programming. Therefore, the study results may not capture the full range of SMW experiences that live captioners face in their daily work. To better understand captioners' perspectives and workload factors, further research, observation, and in-depth discussions with captioners are needed.

AI-supported captioner tool usability study limitations

In this study, some limitations are similar to the other studies, namely sample size, limited variety of content and captioner experience. In addition, there were other limitations specific to this study, including the fact the study was virtual which resulted in technical issues (such as Internet connectivity) that interfered with the speed and synchronization of the PAVOCAT TTS and resulted in inconsistent system performance. Participants may have become frustrated with these technical issues, which may have translated into poorer usability and trust ratings as a result. It is recommended that face-to-face studies could resolve some of these connectivity issues although this would limit the participant pool. The PAVOCAT system itself also had some instability and would cease functioning after about 10 minutes, so it had to be restarted to continue the study. The PAVOCAT development team is now making revisions to the software to address the issues that arose from this study. It is expected that system crashes will not occur in any future studies.

Conclusion

Much of the work carried out in this research project was novel and original. The main findings from this project suggest that there were tradeoffs from speed-accuracy-delay that must be addressed in order to improve CC quality for D/HoH viewers. One main conclusion was that live captioners experience a high SMW when CC fast-paced sporting events. Human factors and psychology practices can provide methods to mitigate this workload including increased breaks, and paired/shared work. However, other options, such as Automatic Speech Recognition (ASR) and AI may provide alternatives. PAVOCAT was developed to use ASR and AI for live CC with captioners as supervisors of the ASR and AI to monitor the ASR/AI performance and select error corrections posed by PAVOCAT. Captioners were wary of this system and wanted to be able to override either the suggestions or the CC when they believed the system performance was inadequate. However, they also believed that a system such as PAVOCAT was almost inevitable and were positive in participating in its further development. Another option investigated in this project was to reduce the amount of CC for fast-paced sports, such as hockey and basketball, where the delays, errors and speech of the CCmake them almost unusable. A study where CC were provided for the commentary announcing (not PBP) was carried out using eye-tracking and a novel method of comprehension assessment. This first study found that D/HoH participants preferred CO captions, however, further longitudinal studies were necessary to reduce the potential novelty effect. The conversation-style interview to measure D/HoH comprehension in a more naturalistic way provided potential improvement from standard comprehension assessment techniques, although a full comparison remains to be carried out. As a result of the innovative and novel work in this project, new questions, issues and gaps also arose that require further research.

Acknowledgements

We gratefully acknowledge the Accessible Standards Canada for their generous funding. We also thank all of the participants who volunteered for all of the studies. We also gratefully acknowledge the advisory board members who provided advice, experience and wisdom for the project.

References

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, 351–370.

Bollapragada, S., & Garbiras, M. (2004). Scheduling Commercials on Broadcast Television. *Operations Research*, *52*(3), 337–345. https://doi.org/10.1287/opre.1030.0083

Boucsein, W., & Thum, M. (1997). Design of work/rest schedules for computer work based on psychophysiological recovery measures. *International Journal of Industrial Ergonomics*, *20*(1), 51–57.

Bowers, J. W., & Phillips, W. A. (1967). A note on the generality of source-credibility scales. *Speech Monographs*, *34*(2), 185–186. https://doi.org/10.1080/03637756709375542

Brooke, J. (1996). SUS: a “quick and dirty” usability scale. In *Usability evaluation in industry* (pp. 189–194). Taylor Francis.

Grier, R. A. (2015). How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 1727–1731. https://doi.org/10.1177/1541931215591373

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkat (Eds.), *Human Mental Workload* (1–Book, Section, pp. 139–183).

Hobfoll, S. E. (2001). The Influence of Culture, Community, and the Nested-Self in the Stress Process: Advancing Conservation of Resources Theory. *Applied Psychology*, *50*(3), 337–421. https://doi.org/10.1111/1464-0597.00062

Jensema, C. J. (1998). Viewer reaction to different television captioning speeds. *American Annals of the Deaf*, *143*(4), Article 4.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Krueger, G. P. (1989). Sustained work, fatigue, sleep loss and performance: A review of the issues. *Work & Stress*, *3*(2), 129–141. https://doi.org/10.1080/02678378908256939

Martens, M. F. J., Nijhuis, F. J. N., Van Boxtel, M. P. J., & Knottnerus, J. A. (1999). Flexible work schedules and mental and physical health. A study of a working population with non-traditional working hours. *Journal of Organizational Behavior*, *20*(1), 35–46. https://doi.org/10.1002/(SICI)1099-1379(199901)20:1<35::AID-JOB879>3.0.CO;2-Z

Meijman, T. F., & Mulder, G. (2013). Psychological aspects of workload. In *A handbook of work and organizational psychology* (pp. 5–33). Psychology press. https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.4324/9780203765425-2&type=chapterpdf

Noldus. (2024). *Behavioral coding—Event logging software | The Observer XT. Behavioral Coding—Event Logging Software | The Observer XT*. https://www.noldus.com/observer-xt

Ofcom. (2005). *Guidance on Standards for Subtitling*. http://www.ofcom.org.uk/tv/ifi/guidance/tv_access_serv/archive/subtitling_stnds/

Russell, H., O’Connell, P. J., & McGinnity, F. (2009). The Impact of Flexible Working Arrangements on Work–life Conflict and Work Pressure in Ireland. *Gender, Work & Organization*, *16*(1), 73–97. https://doi.org/10.1111/j.1468-0432.2008.00431.x

Simpson, E. K., & Kahler, R. C. (1981). A Scale for source credibility; Validated in the selling context. *Journal of Personal Selling & Sales Management*, *1*(1), 17–25.

Syahputra, B. P., Saragih, A., Lubis, S., & Muchtar, M. (2017). INTERPRETING TECHNIQUES BY A TOUR GUIDE AT THE ANCIENT TOMBS OF RAJA SIDABUTAR. *Researchers World*, *8*(1), 151.

Szarkowska, A. (2013). Auteur Description: From the Director's Creative Vision to Audio Description. *Journal of Visual Impairment & Blindness*, *107*(5), Article 5.

Trumbull, R. (1966). Diurnal Cycles and Work-Rest Scheduling in Unusual Environments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *8*(5), 385–398. https://doi.org/10.1177/001872086600800502.