imdc INCLUSIVE MEDIA & DESIGN CENTRE

Supervisory Control for a Novel Captioner-AI System

Hemanshu Bhargav

Agenda

- Introduction
- Methods
- Results & Discussion
- Conclusion

Introduction

Motivation

- Closed Captions (CC) required for the equitable participation of Deaf and Hard-of-Hearing (D/HoH) individuals consuming media
  - Verbatim transcriptions produced by human captioners
  - Can be post-production or live (focus of this thesis)
- Tradeoffs in speed, accuracy and delay in human production
  - Live broadcasts often have speaking rates of 220+ WPM (Fresno et al., 2020)
  - Human caption production 95% accurate at 133 WPM (Ruiz-Arroyo et al., 2022)
  - Delay of 3-10 seconds (Seeber, 2011)
- Captioning is mentally and physically demanding (Nam et al., 2023)

Figure 1: Live CC

Thesis Focus

- Artificial Intelligence (AI) is a form of automation (Veitch & Andreas Alsos, 2022)
- Automatic Speech Recognition (ASR) is replacing human captioners in online meetings and video content
  - Delay of 1-5 seconds (AWS, 2024; Google Cloud, 2024; Otter.ai, 2023)
  - Solves speed but still makes mistakes
- Propositions for thesis
  - AI is the typist and NOT a replacement
  - Captioner promoted to supervisor
- Combine efficiency of AI and judgement of human captioners

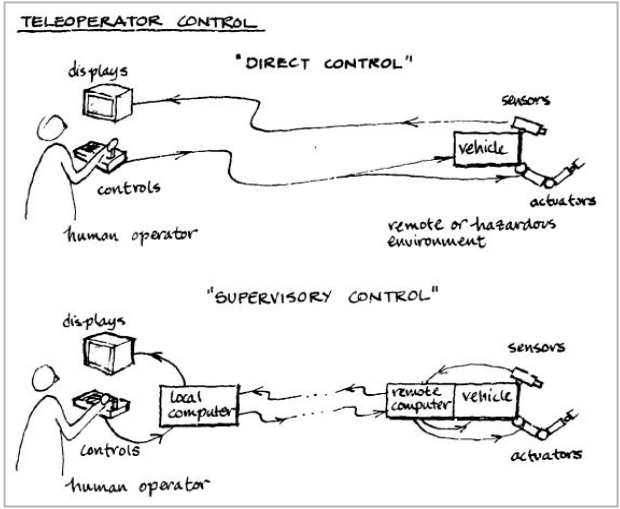

Figure 2: Operator Promoted to Supervisor
Source: Adapted from Sheridan & Verplank, 1978

Levels of Automation

	Level of automation	Monitoring	Generating	Selecting	Implementing
Level 1	Manual control	Human	Human	Human	Human
Level 2	Action support	Human/Computer	Human	Human	Human/Computer
Level 3	Batch processing	Human/Computer	Human	Human	Computer
Level 4	Shared control	Human/Computer	Human/Computer	Human	Human/Computer
Level 5	Decision support	Human/Computer	Human/Computer	Human	Computer
Level 6	Blended decision-making	Human/Computer	Human/Computer	Human/Computer	Computer
Level 7	Rigid system	Human/Computer	Computer	Human	Computer
Level 8	Automated decision-making	Human/Computer	Human/Computer	Computer	Computer
Level 9	Supervisory control	Human/Computer	Computer	Computer	Computer
Level 10	Full Automation	Computer	Computer	Computer	Computer

Figure 3: LOA. Source: Adapted from https://www.functionize.com/blog/levels-of-automation-in-testing

PAVOCAT — Supervisory Control Based Captioning

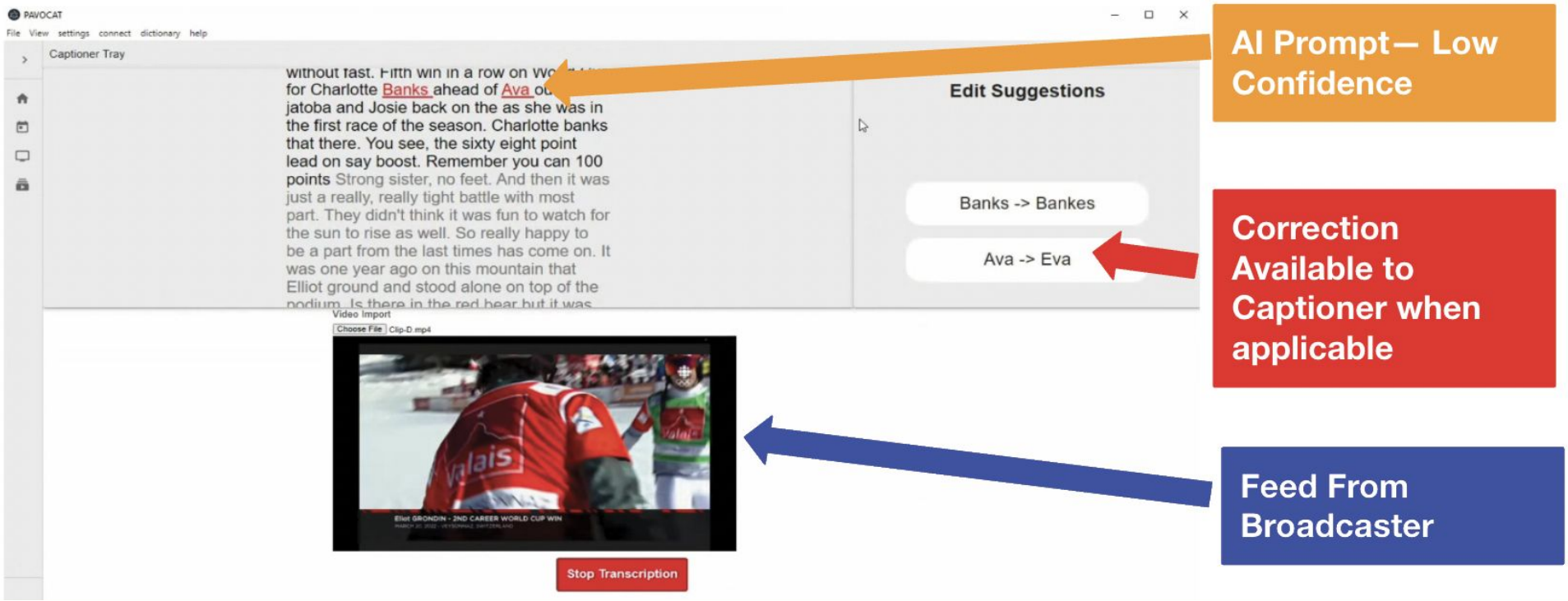

Figure 4: PAVOCAT Interface

Research Questions

1. How do novice and expert captioners work with a novel ASR-based captioning tool, PAVOCAT, when working in “captioner as overseer” mode?
2. What is the experience of captioners working with PAVOCAT?

Methodology

Data Collection

- Time-series study with 10 novice (M = 9, F = 1) & 11 experienced captioners (M = 6, F = 5)
  - 10-minute snowboarding clips (~180 WPM)
  - 3 sessions (repeated measures) analyzed using non-parametric statistics
- Pre-study, between-study and post-study measurement
- Between-Study questionnaires (after each of the 3 sessions)
  - NASA TLX Workload (Hart & Staveland, 1988)
  - Trust in Automation Scale (Jian et al., 2000)
  - System Usability Scale (Brooke, 1995)
  - Satisfaction Scale (Nielsen, 2012b)

Data Collection Continued

- 30 minute semi-structured interview
- Analyzed using Thematic Analysis
  - Also included any comments made before/after captioning sessions
  - Existing workflow, Attention, changes to Workload, etc.
  - Trust, Complexity and Engagement, comparison with other software
  - Future of the role and overriding Edit Suggestions

Human Autonomy System Oversight (HASO) Model

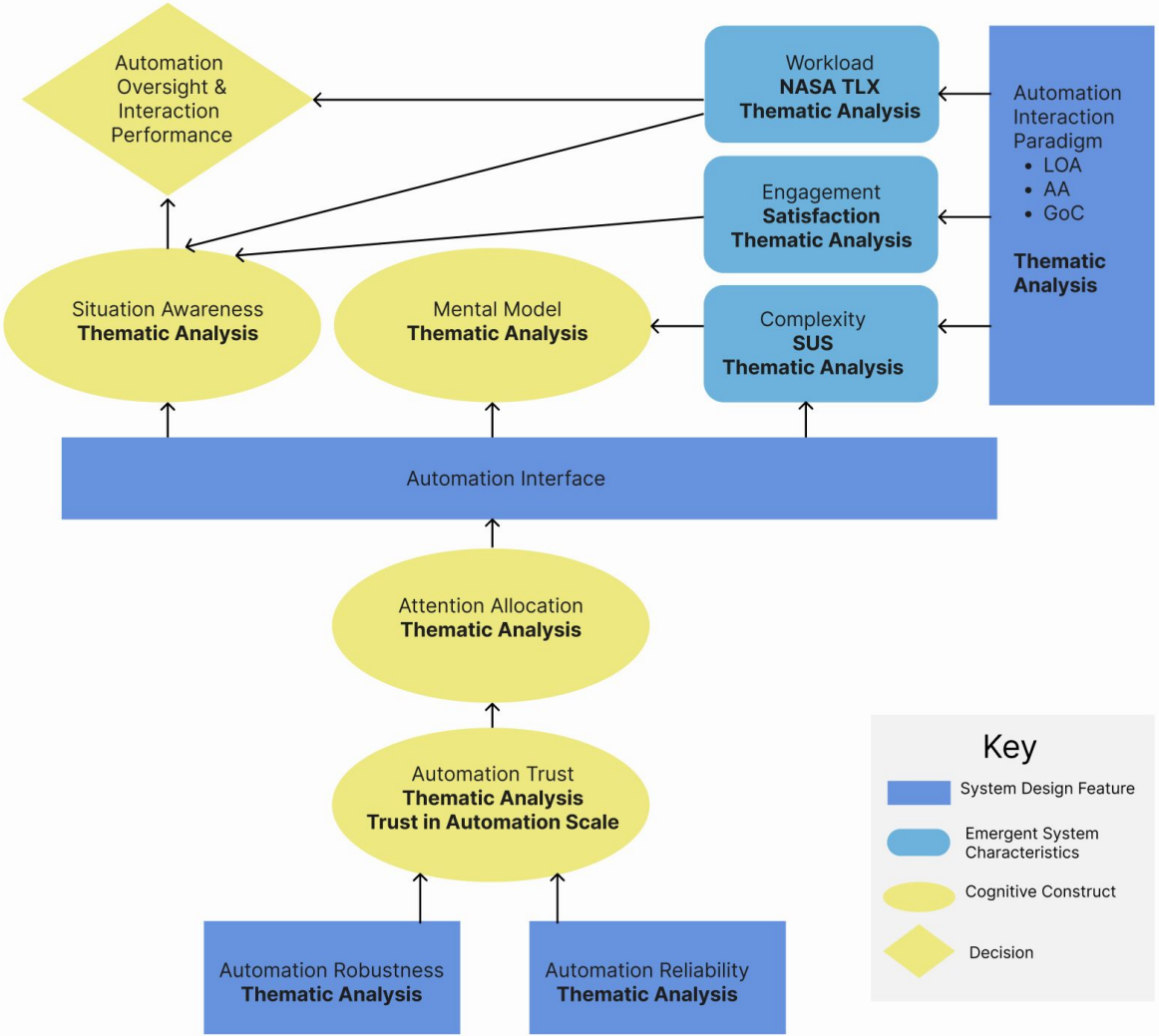

Figure 5: HASO Model w/ Measurements

Source: Adapted from Endsley, 2017, p. 9

Note. Measurement tools are in bold.

Themes Table

Theme	Modifier
1. Performance and Expectations	Positive Negative
2. Physical and Mental Workload	Positive Neutral/No Change Negative
3. Willingness to use AI	Positive Neutral/No Change Negative
4. Human AI Interaction	Override Edit Suggestions Override Typing
5. Comparison with Other Explicitly Named AI/ASR Software	
6. Other (Technical, Content etc.)	
7. User Interface Modifications	

imdc INCLUSIVE MEDIA & DESIGN CENTRE

Results & Discussion

Thematic Analysis Results

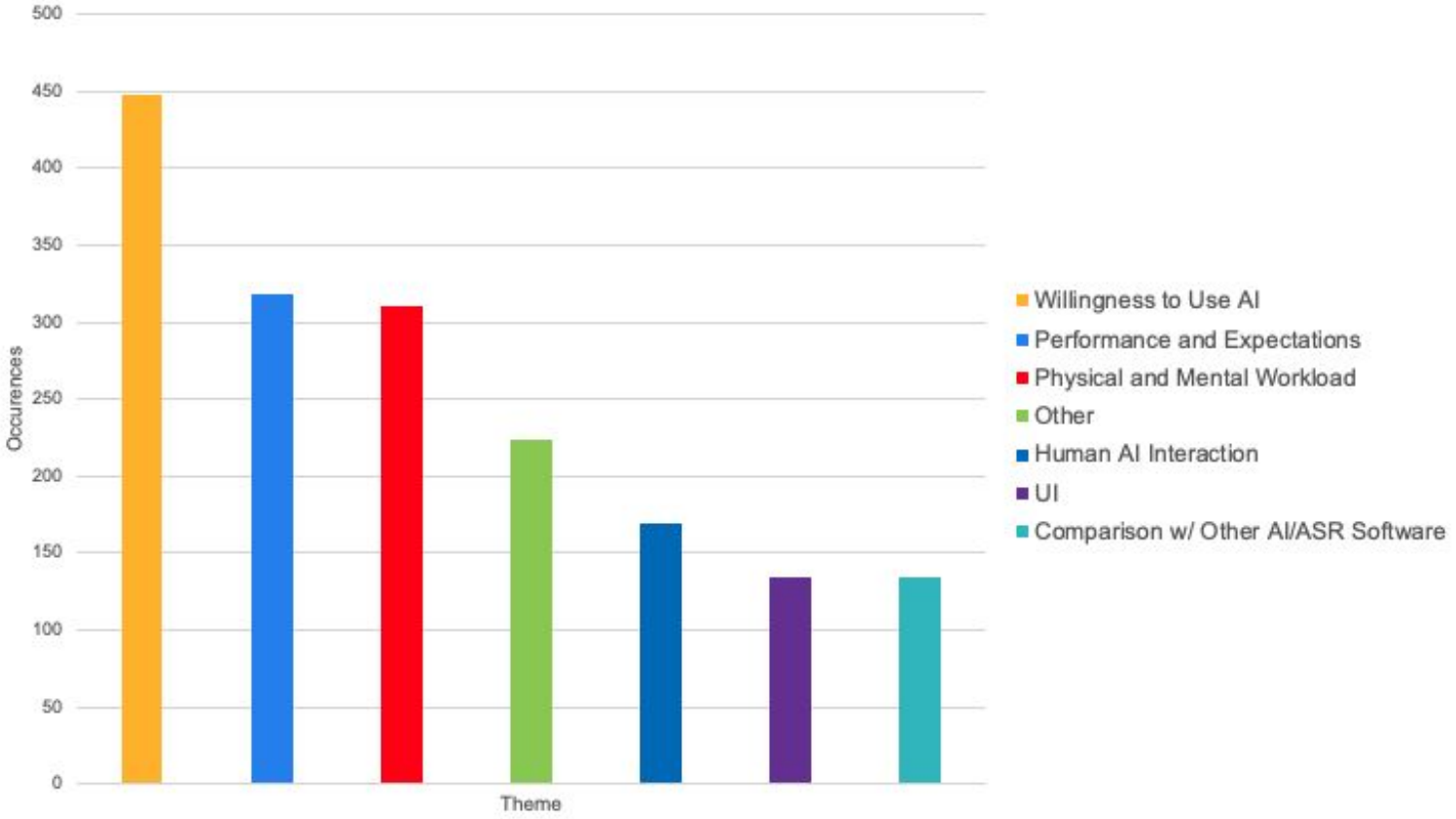

Figure 6: Thematic Analysis Results

Workload

- NASA TLX Overall Medians (/126): Session 1 = 64.5, Session 2 = 54.0 and Session 3 = 47.0
    - Previous related research: NASA TLX workload of captioners 79.3 (SD = 30.3) (Nam et al., 2023)
- Mental Demand reduced from 15 to 8 (/21)
    - Significant difference between Session 1 and Session 3
- 60% (186/311) of all Workload-related comments were positive
    - “Live captioning with this AI is very easy and I feel very relaxed” (Expert)
- AI assisted captioning was less mentally demanding than human captioning

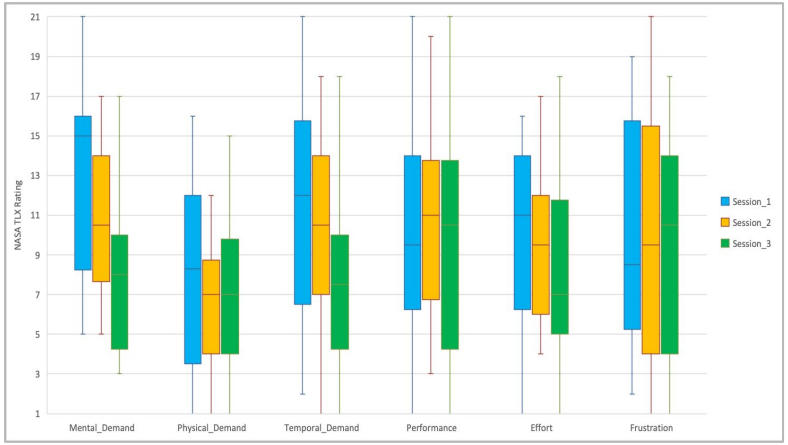

Figure 7: NASA TLX Workload Results

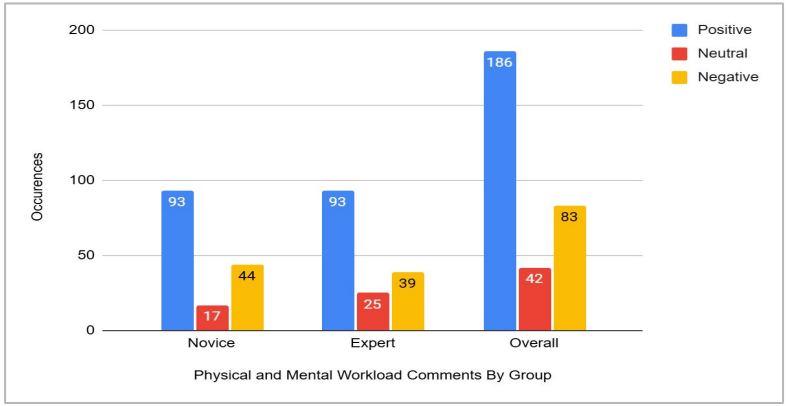

Figure 8: Workload Thematic Results

Automation Robustness and Reliability

- 42% (133/318) comments positive and 58% (185/318) comments negative
  - “At this point, it's not usable” (Expert)
- Significant correlations with Workload
  - Positive Robustness and Reliability and positive Workload ($\tau(21) = 0.55$, $p < 0.05$) and negative Workload ($\tau(21) = -0.43$, $p < 0.05$)
  - Negative Robustness and Reliability and negative Workload ($\tau(21) = 0.43$, $p < 0.05$)
- Significant correlation with Trust
  - Positive Performance and negative Trust ($\tau(21) = -0.48$, $p < 0.05$)

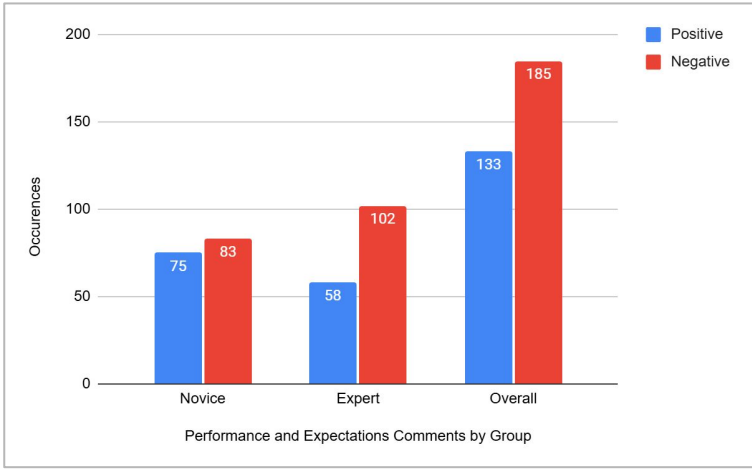

Figure 9: Automation Robustness and Reliability Thematic Results

Relationship b/w Workload and Robustness and Reliability

- Server interruptions possibly reduced Situation Awareness (SA)
  - Unexpected outputs reduce SA (Endsley, 1999, 2000; Endsley, 1995; Endsley & Kiris, 1995)
- Automation Brittleness— at least two server shutdowns per 10-minute session
- Lower workload can be related to complacency because the supervisor is idle for periods of time (Clark et al., 2019; Endsley, 2017)
- Participants did not quit the task indicating none fell Out-of-The-Loop (OOTL)
  - Maintained perception of the task
- Two participants did demonstrate ability to project how their role may adapt in the future with improvements in AI
  - “Maybe my role shifts to like I can monitor several jobs at a time” (Expert)
  - “Or helping the AI know how to weight them differently” (Expert)

Automation Trust

- Trust in Automation Scale overall median rating decreased over time: 4.30 (SD=1.27) to 3.68 (SD=1.07) out of 7
  - Suggests novelty effect of software
- 71% (318/447) comments positive
  - Scale confusing to participants
    - Many items left unanswered
- Participants approve “Captioner-as-Supervisor” workflow
  - No difference b/w novices and experts
- Correlation b/w Trust and Workload ($\tau = 0.48$, $p < 0.05$)
- Trust affects Workload
  - “We can't really trust AI, so we should have the ability to edit [the captions]” (Novice)

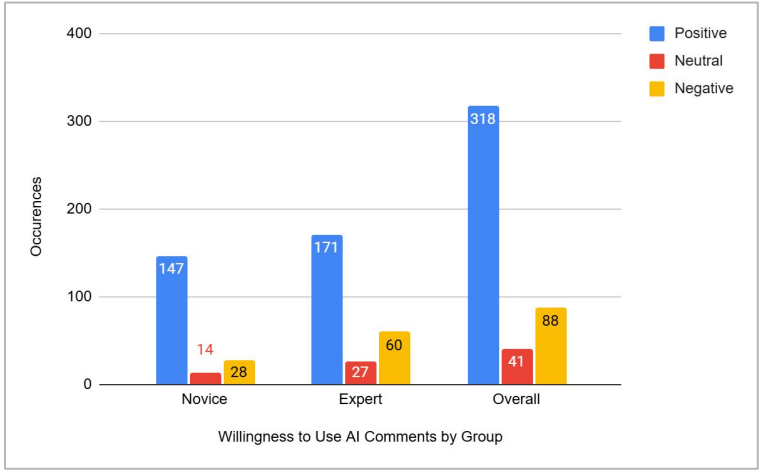

Figure 10: Automation Trust Thematic Results

Proposed HASO Diagram

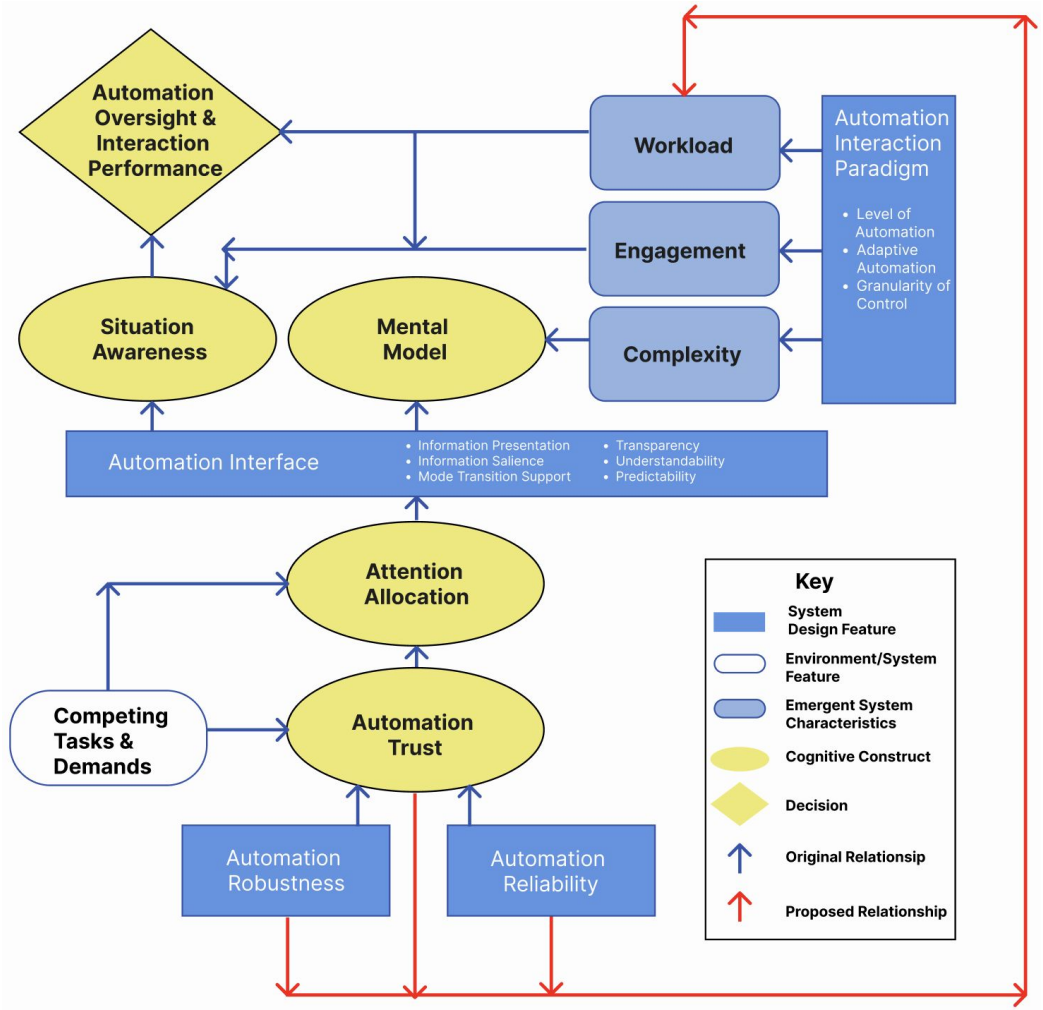

Figure 11: Proposed HASO Diagram
Source: Adapted from Endsley, 2017, p. 9

Complexity and Engagement

- Session 3 SUS = 59.05 (IQR=17.15) out of 100
  - SUS 68-70 are considered to be usable (Bangor et al., 2008; Brooke, 1995)
- Session 3 Satisfaction M = 4.50 (IQR = 1.25) out of 7
- No significant differences over time or between novices and experts
- Interface acceptable but many suggestions offered for window size, layout, etc.
  - “If you could have some options about caption placement, yeah, that would be good” (Expert)
- Comparisons made to existing AI services
  - “I have been taught how to use Amber Scripts” (Novice)

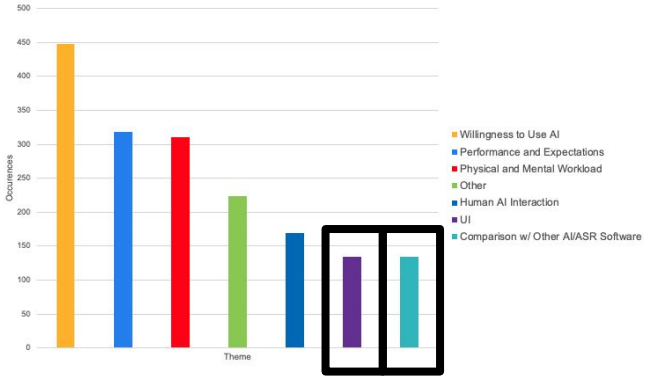

Figure 12: Complexity and Engagement Thematic Results

Automation Interaction Paradigm

- Significant correlation b/w negative Robustness and Reliability and Override Suggestions ($\tau(21) = 0.36, p < 0.05$)
  - Limited accuracy of Edit Suggestions made participants desire control over corrections
- Captioners prefer to assert more direct control
- Recommendation: PAVOCAT needs to be an Adaptive Automation (AA) with more Granularity of Control (GoC)
  - Allow captioners to decide when more control is necessary

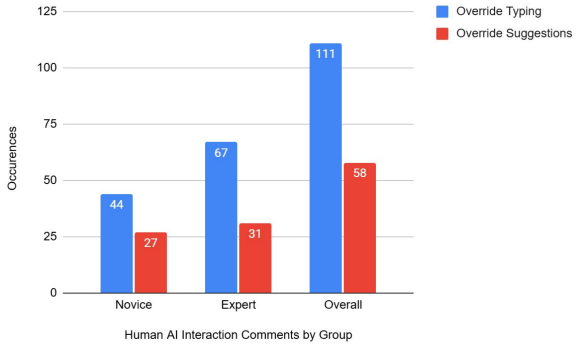

Figure 13: Automation Interaction Paradigm Thematic Results

Limitations

- Small participant pool due to small workforce
- Technical set-up
  - Audio Degradation
    - Limited accuracy
  - PAVOCAT server on limited resources caused delays which exceeded 20 seconds at times
- Virtual set-up
  - Eye-tracking, a common measure of Attention, was disqualified as a result

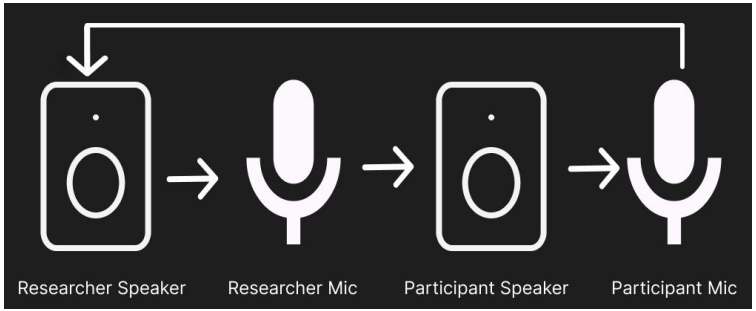

Figure 14: Audio Setup

Conclusions & Contributions

Design Recommendations

- Add and remove Suggestions
  - Suggestions did not synchronize with identified errors
- Assert control over Edit Suggestions and manually type/respeak
- Add ability to import personal dictionary
- Level 4 LOA

	Level of automation	Monitoring	Generating	Selecting	Implementing
Level 1	Manual control	Human	Human	Human	Human
Level 2	Action support	Human/Computer	Human	Human	Human/Computer
Level 3	Batch processing	Human/Computer	Human	Human	Computer
Level 4	Shared control	Human/Computer	Human/Computer	Human	Human/Computer
Level 5	Decision support	Human/Computer	Human/Computer	Human	Computer
Level 6	Blended decision-making	Human/Computer	Human/Computer	Human/Computer	Computer
Level 7	Rigid system	Human/Computer	Computer	Human	Computer
Level 8	Automated decision-making	Human/Computer	Human/Computer	Computer	Computer
Level 9	Supervisory control	Human/Computer	Computer	Computer	Computer
Level 10	Full Automation	Computer	Computer	Computer	Computer

Figure 15: LOA Recommendation. Source: Adapted from https://www.functionize.com/blog/levels-of-automation-in-testing

Contributions

1. Theoretical contribution: Proposed addition of direct relationship between Workload, Automation Trust, Robustness and Reliability to HASO Model

2. Methodological contribution: Used existing validated tools applied to HASO model

Conclusion

- Working in captioner-as-overseer mode
  - Seems to lower Mental Demand and Workload
  - Can benefit from Adaptive Automation with greater GoC than available in this study using PAVOCAT
  - No significant differences b/w novices and experts
  - Participant comments showed no clear distinction b/w novices vs experts
- Future Work
  - Longitudinal study with full length (e.g., 3-hour hockey game) broadcast content
  - Develop validated scales for Complexity and Engagement and use objective measurements for Attention Allocation

imdc INCLUSIVE MEDIA & DESIGN CENTRE

Acknowledgements

This project has been made possible by

Accessibility Standards Canada

Canada

BROADCASTING ACCESSIBILITY FUND
FONDS POUR L'ACCESSIBILITÉ DE LA RADIODIFFUSION

- The PAVOCAT Team
  - Christie Christelis, Project Manager
  - Sander Fels-Leung, Development Manager
  - Gabriella Hong, Senior UX Researcher
  - Rishabh Sharma, Developer
  - Mohammad Zaman, Developer
  - Mohammed Pasha, Developer
  - David Rose, Developer
- Funding
  - Accessible Technology Program of Innovation, Science and Economic Development Canada
  - Broadcast Accessibility Fund (BAF)
  - Queen Elizabeth II Graduate Scholarship in Science and Technology (QEII-GSST)
  - Toronto Metropolitan University Graduate Scholarship
- Research Team
  - Deborah I Fels, PhD., P.Eng.
  - Patrick W Neumann, PhD
  - Somang Nam, PhD

Thank you!