EmoScribe Camera: A Virtual Camera System to Enliven Online Conferencing with Automatically Generated Emotional Text Captions

Ari Hautasaari*
The University of Tokyo
Tokyo, JAPAN
ari@nae-lab.org

Minami Aramaki*
The University of Tokyo
Tokyo, JAPAN
aramaki@nae-lab.org

Rintaro Chujo
The University of Tokyo
Tokyo, JAPAN
chujo@nae-lab.org

Takeshi Naemura
The University of Tokyo
Tokyo, JAPAN
naemura@nae-lab.org

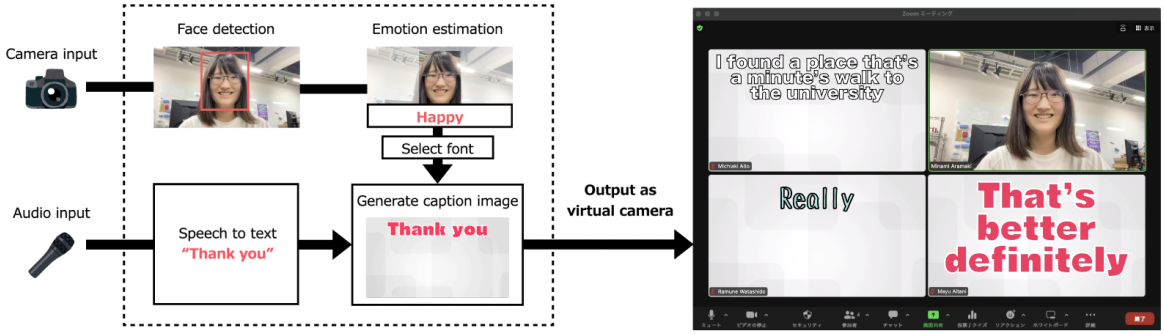

Figure 1: Flowchart of the overall EmoScribe Camera system process.

ABSTRACT

Ideally, for lively discussions to occur during online meetings, the participants should turn on both their camera and microphone. In practice this is not always possible, and meeting participants may opt to use a text chat to communicate their ideas and reactions instead. However, text messages are also time-consuming and labor-intensive to type as well as omit many of the emotional cues available through visual and audio channels. To address these issues, we propose EmoScribe Camera, a virtual camera system that generates images of automatic text captions in real time and outputs them as a software-based virtual camera that simulates a physical camera. We report on the results of a user study evaluating the efficacy of EmoScribe Camera as an alternative communication channel during online conferences when participants have their camera and microphone turned off.

*Both authors contributed equally to this research.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI EA '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650987

CCS CONCEPTS

• **Human-centered computing** → **Interaction paradigms**.

KEYWORDS

virtual camera, automated speech recognition, caption, online conferencing, emotion, font

ACM Reference Format:
Ari Hautasaari, Minami Aramaki, Rintaro Chujo, and Takeshi Naemura. 2024. EmoScribe Camera: A Virtual Camera System to Enliven Online Conferencing with Automatically Generated Emotional Text Captions. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3650987

1 INTRODUCTION

The role of online meetings in business and academia transformed due to the impact of the COVID-19 pandemic, but this shift also highlighted the technological limitations of the online conferencing mediums for social interactions and participant unity compared to face-to-face settings [1, 28]. For one, all meeting participants should ideally have their cameras and microphones turned on in order to facilitate communication and lively discussions, but in practice there are many situations where people participate in online meetings with their cameras and microphones turned off [24, 30]. In such cases, the text chat function of an online conferencing tool allows

participants to comment in text, whereas the reaction buttons offer a method for participants to engage in the conversation nonverbally. However, the text chat functionality requires operation from the user which incurs a time lag until a message is seen by others, and neither method conveys rich information, such as emotional cues in facial expressions or voice, making it challenging for detailed intentions and nuances to be conveyed [7, 29].

In this work, we propose a virtual camera system called EmoScribe Camera as an alternative communication channel that is positioned between video/audio calls and text chats. The system generates a text caption image from automatically transcribed speech voice in real time and outputs the image as a software-based virtual camera that simulates a physical camera or webcam, as shown in Figure 1, without the user having to turn on their microphone or camera in the online conferencing application. In addition, EmoScribe Camera estimates the emotional state of the speaker based on their facial expression, and changes the appearance of the caption text by selecting an emotional font based on the detected emotion. In this paper, we present the EmoScribe Camera system design and implementation, as well as user evaluation results from a group discussion conducted with EmoScribe Camera.

2 RELATED WORK

In online conferencing tools, the built-in text chat function is often used by participants who join with their cameras and microphones turned off. However, text chats are considered a lean communication medium due to the lack of available communicative cues, such as facial expressions or tone of voice [7, 8, 15]. Text chats are also prone to lower real-time characteristics (e.g., slower message transmission) compared to voice calls [19]. Furthermore, emotional expression plays a crucial role in social interactions in everyday life. Especially in face-to-face communication, people use both verbal and non-verbal cues, facial expressions, voice and gestures, to express their emotions [26], which are not available in text-based chats. Due to the lack of these cues, discerning other's emotions and emotional states is more challenging in text-based conversations compared to rich mediums, such as video conferencing, which can in turn lead to misunderstandings between meeting participants [3, 27, 29].

Therefore, in this work we propose a novel communication method that is positioned between lean text-based computer-mediated communication (CMC) mediums and rich video/audio conferencing by utilizing automated speech recognition (ASR) and emotion detection technology implemented in a virtual camera software. In the next section, we briefly cover previous literature on using ASR in video/audio conferencing.

2.1 Automatic Captions in Video/Audio Conferencing

Online meeting tools such as Zoom[1] and Microsoft Teams[2] include built-in automatic captioning features. These built-in features are currently the prevalent method for displaying real-time captions, but at the same time they cannot be used if meeting participants turn their microphones off (i.e., automatic captions are not independent from the audio channel). The effectiveness of using captions created by ASR technology as a communication aid in video and audio conferencing has been extensively investigated in previous literature (e.g., [4, 5, 10, 11, 21–23, 31]). Bulk of the recent research has focused on supporting non-native speakers during audio conferencing. For example, Pan et al. investigated the impact of ASR transcription quality on users' comprehension and subjective assessment in video/audio conferences [22, 23], while Gao et al. explored how real-time transcription can facilitate multilingual communication in audio conferencing [10, 11]. Incorporating additional features to ASR transcriptions have also been proposed, such as adding bilingual dictionaries or highlighting key parts of the transcribed conversation in the caption text [11, 21].

In addition to online conferencing tools, captions or subtitles are most commonly used in other media content, such as television programs. In these contexts, subtitle fonts are sometimes used to express emotional reactions of the program cast for production purposes [20]. DNP developed an emotional expression subtitle system [3] that automates this process based on a font dataset created by Ishii et al. [14]. This system estimates the emotions of speakers in videos based on their facial expressions and speech content and automatically adds subtitles in fonts corresponding to these emotions. While previous research is scarce on the use of emotional fonts in online conferencing systems, applications that overlay automatic captions with adjustable size and font on online meeting tools have been proposed [18]. Furthermore, Hassan et al. explored the use of typography to convey emotions in English video captions [13], whereas results reported in Chujo et al. [6] indicated that certain Japanese fonts can be used to convey emotions based on Ekman's six basic emotions (fear, joy, sadness, surprise, disgust, anger) [9].

2.2 Positioning of This Research

Conversely to current built-in ASR-based automatic captions that are tied to the audio channel and not intended for standalone use, EmoScribe Camera operates independently to the audio channel in online conferencing tools. Additionally, the appearance of subtitles (i.e., fonts) is constant in built-in systems, making it difficult to discern the speaker's emotional tone from text alone. Similarly to the emotional expression subtitle system by DNP, EmoScribe Camera adjusts the captions to match the speaker's emotions through font selection. Currently, in order to share the automatic captions provided by external software on online meeting tools, screen sharing is necessary, but concurrent screen sharing is in general limited to one user with existing systems [4] making it an impractical method for sharing captions of individual participant's speech.

EmoScribe Camera addresses the above limitations through it's design as a virtual camera application. A virtual camera is a software-simulated camera, as opposed to a physical webcam. Being software-based, it can handle video, images and text compositions as outputs while appearing as a regular webcam to other software. As a virtual camera application, EmoScribe Camera is independent and can be used in various online meeting tools such

[1] https://explore.zoom.us/products/meetings
[2] https://www.microsoft.com/microsoft-teams/group-chat-software
[3] https://shueitai.dnp.co.jp/news/detail/10161885_3733.html
[4] https://support.zoom.com/hc/ja/article?id=zm_kb&sysparm_article=KB0060608

Table 1: Fonts, caption colors and corresponding emotions used in EmoScribe Camera

Estimated Emotion	Caption Color	Font Name	Font Image
neutral	White	Noto Sans	山路を登りながら
fear	Blue	Hakidame	山路を登りながら
joy	Peach	RaglanPunchStd	山路を登りながら
sadness	Blue	851チカラヨワク	山路を登りながら
surprise	Light Blue	たぬき油性マジック	山路を登りながら
disgust	Lime Green	ab-tyuusyobokunenn	山路を登りながら
anger	Red	HGP明朝E	山路を登りながら

as Zoom and Microsoft Teams. In the following sections, we detail the design and implementation of our proposed system.

3 DESIGN AND IMPLEMENTATION OF EMOSCRIBE CAMERA

EmoScribe Camera is a virtual camera application operating on macOS and implemented using Swift[5] Camera Extension[6]. Camera Extension is a system extension available from macOS 12.3 onwards[7]. Compared to the previously used Device Abstraction Layer (DAL) plugins[8] for building virtual cameras on earlier versions of macOS, Camera Extension is easier to develop and poses fewer security risks. EmoScribe Camera outputs a video that overlays automatically generated captions on any chosen background video (i.e., either a preset background image or another camera input) as a virtual camera.

To transcribe user speech in real-time, the Speech framework[9] was used. Speech is an audio recognition framework compatible with macOS, supporting many languages including English and Japanese. Speech has a limitation of 60 seconds per request and a maximum of 1000 requests per hour[10]. Therefore, to maintain continuous automated caption functionality, requests are regenerated simultaneously with a 60-second timeout.

For face recognition and emotion detection, CoreML[11], a library for handling machine learning on Swift, was used. The standard model in CoreML[12] was employed to crop facial images for emotion recognition. The CNNEmotion model [16] was used for emotion estimation. The estimated emotions are categorized into seven basic emotions based on Ekman's research [9], and the strongest estimated emotion category determines the font and color used for the captions, as shown in Table 1.

[5]https://www.apple.com/swift
[6]https://developer.apple.com/documentation/coremediaio/creating_a_camera_extension_with_core_media_i_o
[7]https://developer.apple.com/documentation/systemextensions
[8]https://developer.apple.com/documentation/coremediaio/device_abstraction_layer_dal_plug-ins
[9]https://developer.apple.com/documentation/speech
[10]https://developer.apple.com/documentation/speech/sfspeechrecognizer
[11]https://developer.apple.com/jp/machine-learning/core-ml
[12]https://developer.apple.com/documentation/vision/vndetectfacerectanglesrequest

The optimal font size and caption display time when EmoScribe Camera is used with online conferencing systems were determined through preliminary experiments with 12 and 9 participants, respectively. A font size of 200pt was determined to be readable even with the screen divided to the maximum number of 49 meeting participants in Zoom (adjusted size of a 1920x1080 pixel output video to fit a display size equivalent to the division of a 3024x1964 pixel screen). At 200pt, EmoScribe Camera can accommodate approximately 36 Japanese characters simultaneously on one user's video feed area. The optimal display time for captions displayed with EmoScribe Camera was determined by participants viewing sentences of 10/20/30/40 characters in length and using a stopwatch to indicate when they felt the captions should disappear from their screen. As a result, we found that 4 seconds was sufficient caption display time for the longest sentences at 200pt to be read. These parameters were applied in the user study described in the following section.

4 USER STUDY COMPARING EMOSCRIBE CAMERA AND TEXT CHAT

To assess the efficacy of EmoScribe Camera to enliven online group discussions, we conducted a controlled experiment comparing our prototype with Zoom's text chat. As our proposed system is positioned between video/audio calls and text-based mediums, we chose text chat as the baseline condition in this inital study as it is commonly used by participants who connect to online meetings with their camera and microphone off.

The experiment simulated a group discussion scenario on Zoom with both camera and microphone turned off. Participants engaged in discussions using both EmoScribe Camera and Zoom's text chat (within-subjects design). We prepared two discussion themes for the participants: "The most surprising thing about my life at our university", and "The most annoying thing about my life at our university". Each participant shared their experiences related to these themes, followed by a collective decision-making process to identify the most notable experience. The order of discussion topics was counterbalanced between conditions.

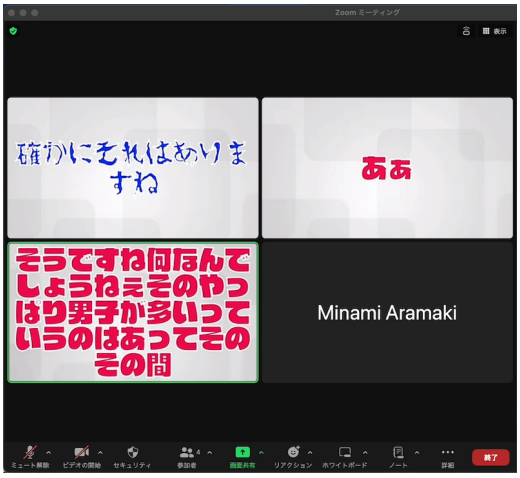

Figure 2: Scene of a group discussion conducted on Zoom with EmoScribe Camera

4.1 Participants

We recruited 24 Japanese-speaking students from the University of Tokyo with a mean age of 21.81±1.80 years (12 females, 12 males). The participants were divided into four groups of three female participants, and four groups of three male participants. This choice in group division was to limit any potential gender effects arising in mixed groups. The order in which each group engaged with either EmoScribe Camera or Zoom's text chat was counterbalanced between groups to account for order effects.

4.2 Software, Equipment and Experiment Environment

Each participant was provided with a MacBook Pro (macOS Ventura) with EmoScribe Camera (font size was set at 200pt, and caption display time at 4 seconds; Chapter 3) and Zoom installed, and they were assigned names Participant 1, Participant 2 and Participant 3 when connecting to Zoom. The participants were seated apart from each other behind partitions and wore headphones throughout the experiment to ensure they could not hear each other. Once connected to Zoom, the participants were instructed to keep the Zoom camera and microphone off throughout the experiment tasks, aside from icebreaker and self-introduction. The experiment facilitator kept their camera off but used voice (i.e., Zoom microphone on) to give instructions (Figure 2).

4.3 Experimental Procedure

For an icebreaker and as a practice for the group discussion, participants introduced themselves on Zoom followed by a question-and-answer session about their introductions. Subsequently, the participants were briefed on how to use EmoScribe Camera followed by a practice session. To ensure proficiency with the conventional text chat as well, the participants were also asked to practice typing in the Zoom's text chat field. After the practice tasks and a short break, the participants engaged in a 10-minute group discussion with either EmoScribe Camera or Zoom's text chat followed by a questionnaire. Before beginning the task, they were instructed to keep both their camera and microphone off. The participants then repeated this process with the other tool and discussion theme, after which they participated in individual semi-structured interviews about their experiences.

4.4 Measures

To assess the liveliness of discussion, we calculated the total amount of speech produced by each participant. As the experiment was conducted in Japanese, we calculated the amount of speech as the number of hiragana characters typed in Zoom's text chat and in the automated transcriptions by EmoScribe Camera after converting all kanji characters to their hiragana form. The participants answered a post-task survey on System Usability Scale (SUS) [2] for evaluating the usability of each method, and the NASA Task Load Index (NASA-TLX) [12] for assessing their workload. The SUS questions were translated into Japanese based on the Japanese translation by Sato et al. [25]. Similarly, the NASA-TLX questions were translated based on the Japanese version by Miyake et al. [17]. Additionally, the participants were asked to freely describe their impressions of the tools used during the group discussions, highlighting both positive and negative aspects, in the post-task questionnaires.

At the end of the experiment, we conducted semi-structured interviews with the following key questions:

- Why did you respond as you did in the survey?
- Which method, EmoScribe Camera or text chat, facilitated smoother discussion?
- Which method made the discussion more lively?
- What were the good and bad points of using both EmoScribe Camera and text chat?

4.5 Results

The average amount of speech produced by participants with EmoScribe Camera and Zoom's text chat are shown in Figure 3. Results from a paired t-test revealed that the participants produced significantly higher amount of speech when using EmoScribe Camera ($M=832.54$, $SD=318.14$) compared to Zoom's text chat ($M=370.08$, $SD=136.09$): $t[23] = 7.27$, $p < .05$. This result in part indicated that the group discussions were more lively with EmoScribe Camera compared to traditional text chat.

Regarding the usability of each medium, results from a paired t-test showed no significant difference in the average SUS scores between EmoScribe Camera ($M=60.46$, $SD=18.42$) and Zoom's text chat ($M=65.37$, $SD=15.79$): $t[23] = 1.16$, $p > .05$.

The average NASA-TLX scores are depicted in Figure 4. Results from a paired t-test showed that the participants experienced a significantly lower workload when using EmoScribe Camera ($M=37.60$, $SD=16.05$) compared to Zoom's text chat ($M=53.23$, $SD=17.17$): $t[23] = 4.41$, $p < .05$. Analyses comparing the results for individual NASA-TLX items revealed that the participants experienced a significantly lower physical demand (Shapiro-Wilk: $p < .05$, Wilcoxon Signed-Rank test: $Z=-3.65$, $p < .05$), temporal demand (paired t-test: $t[23] = 6.53$, $p < .05$) and effort (paired t-test: $t[23] = 2.45$, $p < .05$), and higher performance (paired t-test: $t[23] = 2.27$, $p < .05$) with EmoScribe Camera compared to Zoom's text chat. Comparisons for mental demand and frustration were not statistically significant ($p > .05$)

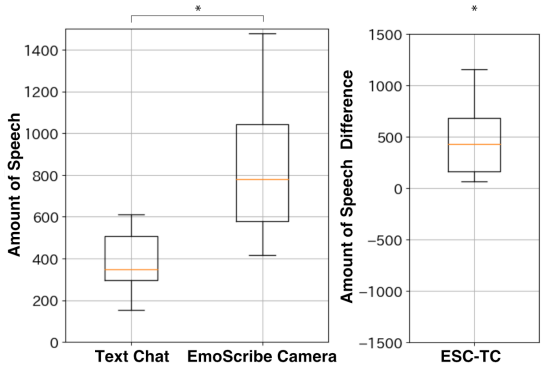

Figure 3: Left: Amount of speech for EmoScribe Camera and Zoom's text chat, Right: Difference in average amount of speech per participant when using EmoScribe Camera (ESC) and Zoom's text chat (TC).

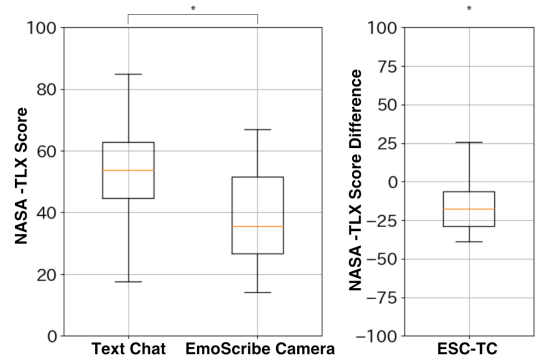

Figure 4: Left: NASA-TLX scores for EmoScribe Camera and Zoom's text chat, Right: Difference in average NASA-TLX scores per participant between EmoScribe Camera (ESC) and Zoom's text chat (TC).

Results from the free-response questionnaire items and semi-structured interviews attributed the lower workload with EmoScribe Camera to lower physical burden compared to typing (12/24 participants), as well as reduced time pressure due to being able to react to other's utterances casually (12/24 participants) and lower time lag between messages leading to a smoother conversation (8/24 participants). Furthermore, 6/24 participants specified the ease of conveying emotions as a positive aspect of EmoScribe Camera, but for some participants emotional communication was somewhat hampered by low facial recognition accuracy (4/24 participants). Lastly, the interview results indicated that 16 out of 24 participants found discussions smoother with EmoScribe Camera, and 21 out of 24 participants mentioned that discussions were more lively with EmoScribe Camera compared to Zoom's text chat.

5 DISCUSSION

We evaluated EmoScribe Camera as a novel communication channel for real-time online meetings in a controlled experiment. Compared to a traditional text chat, our evaluation results suggested that EmoScribe Camera can elicit more lively discussions than a text-based channel while incurring a lower burden on the user when meeting participants have both their microphone and camera turned off.

In terms of usability, we found no significant differences in the SUS evaluations between EmoScribe Camera and Zoom's text chat. While text chats are relatively simple tools familiar to many, if not most, users, they suffer from a lack of real-time responsiveness, as also indicated in the interview results. EmoScribe Camera offers ease of communication similar to a voice call, but is also limited by the ASR transcription accuracy which may lead to increased stress for users who are unfamiliar with the technology. Combined, these characteristics of EmoScribe Camera likely contributed negatively to the usability evaluations.

We envision that EmoScribe Camera can be utilized and evaluated in a number of communicative situations in future work. As described above, the system can also be used when meeting participants have their camera on, where the captions are overlaid on the video feed instead of a pre-set background image. In this case, the visualization of emotions based on facial expressions remains a supplementary element, but the advantage of being able to communicate in real-time without transmitting voice persists. Similarly, when using EmoScribe Camera with the microphone on, it combines normal voice calls with the emotional caption transmission provided by the system.

As the utterances made by meeting participants using EmoScribe Camera do not accompany voice, it allows for comments and reactions without hindering other participants' speech. However, in the user study described in Section 4, we observed that side conversations (i.e., conversations occurring parallel to the main conversation) between participants using the system also occurred. In other words, one future use case for EmoScribe Camera is to facilitate side conversations during online meetings through automatic captions displayed as video in parallel to the main conversation occurring over the audio channel. Lastly, EmoScribe Camera may not only be beneficial for online meetings but can also be leveraged in hybrid environments [28]. Besides supporting side conversations among online participants without interfering with the main conversation occurring simultaneously online and face-to-face, EmoScribe Camera can also convey the reactions of online participants in real-time to those present in-person facilitating smoother communication and participant unity.

5.1 Future Work and Limitations

In regards to the evaluation of EmoScribe Camera for emotional communication support, we did not quantitatively compare the extent to which emotional transmission through fonts contributes to the liveliness of discussions. Hence, future experiments should be conducted between the proposed version and a version of EmoScribe Camera where the caption font remains constant, while also considering alternative cues for emotional transmission, such as icons or emoticons. Additionally, it is imperative to investigate how the absence of voice transmission affects users' burden and thereby potentially contributes to the liveliness of online discussions. However, due to the difficulty of evaluating this aspect in a controlled experiment, we are planning to conduct field studies in more naturalistic settings, including noisy environments where

users may prefer keeping their camera and microphone off during meetings.

There are also several technical aspects that should be considered in future work. Firstly, we developed the prototype using the Speech framework for automatic transcription. However, as this framework has a limitation of 60 seconds per request, using a transcription method with less restrictive limitations could improve the usability of the system. Secondly, the parameters for font size and caption display time were experimentally determined, but they were set on a scenario where participants have both their microphone and camera off. Hence, reevaluating these parameters for other use cases is necessary. Lastly, the selection of fonts corresponding to estimated emotions was done subjectively in this study, and it cannot be definitively said that the emotional impression received from the automatic caption fonts aligns accurately with the estimated emotions. To improve on this design, it is necessary to first assess the emotional information conveyed in fonts [6], while also considering readability as well as other aspects such as the color of the text.

6 CONCLUSION

In this study, we proposed EmoScribe Camera, a virtual camera application capable of outputting automatic captions with emotional fonts that change according to the emotions detected from user's facial expressions. This application was developed to promote lively communication in online meetings even if participants have their cameras and microphones turned off. Results from our user study suggested that compared to traditional text chat in online conferencing systems, EmoScribe Camera promotes a livelier discussion while incurring a lower workload on the user. Based on the results, we discussed future directions and applications of EmoScribe Camera in more naturalistic online conferencing settings.

ACKNOWLEDGMENTS

This research is part of the results of Value Exchange Engineering, a joint research project between R4D, Mercari Inc., and the RIISE.

REFERENCES

[1] Rachel Bergmann, Sean Rintel, Nancy Baym, Advait Sarkar, Damian Borowiec, Priscilla Wong, and Abigail Sellen. 2023. Meeting (the) Pandemic: Videoconferencing Fatigue and Evolving Tensions of Sociality in Enterprise Video Meetings During COVID-19. *Computer Supported Cooperative Work (CSCW)* 32, 2 (2023), 347–383.

[2] John Brooke. 1996. SUS : A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 3 (1996), 189–194.

[3] Kristin Byron. 2008. Carrying Too Heavy a Load? The Communication and Miscommunication of Emotion by Email. *Academy of Management Review* 33, 2 (2008), 309–327.

[4] Xun Cao, Naomi Yamashita, and Toru Ishida. 2018. Effects of Automated Transcripts on Non-native Speakers' Listening Comprehension. *IEICE Transactions on Information and Systems* 101, 3 (2018), 730–739.

[5] Mei-Ling Chen, Naomi Yamashita, and Hao-Chuan Wang. 2018. Come Together: Facilitating Collocated Multilingual Group Discussion with a Language Support Tool. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, 1–6. https://doi.org/10.1145/3170427.3188595

[6] Rintaro Chujo, Atsunobu Suzuki, and Ari Hautasaari. 2024. Exploring the Effects of Japanese Font Designs on Impression Formation and Decision-Making in Text-Based Communication. *IEICE Transaction on Information Systems* E107-D, 3 (2024), 354–362. https://doi.org/10.1587/transinf.2023HCP0009

[7] Richard L Daft and Robert H Lengel. 1986. Organizational Information Requirements, Media Richness and Structural Design. *Management Science* 32, 5 (1986), 554–571.

[8] Richard L. Daft, Robert H. Lengel, and Linda Klebe Trevino. 1987. Message Equivocality, Media Selection, and Manager Performance: Implications for Information Systems. *MIS Quarterly* 11, 3 (1987), 355–366. http://www.jstor.org/stable/248682

[9] Paul Ekman. 1992. Are There Basic Emotions? *Psychological Review* 99, 3 (1992), 550–553. https://doi.org/10.1037/0033-295x.99.3.550

[10] Ge Gao, Naomi Yamashita, Ari Hautasaari, Andy Echenique, and Susan R. Fussell. 2014. Effects of Public vs. Private Automated Transcripts on Multiparty Communication Between Native and Non-native English Speakers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, 843–852. https://doi.org/10.1145/2556288.2557303

[11] Ge Gao, Naomi Yamashita, Ari Hautasaari, and Susan R. Fussell. 2015. Improving Multilingual Collaboration by Displaying How Non-Native Speakers Use Automated Transcripts and Bilingual Dictionaries. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, 3463–3472. https://doi.org/10.1145/2702123.2702498

[12] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[13] Saad Hassan, Yao Ding, Agneya Abhimanyu Kerure, Christi Miller, John Burnett, Emily Biondo, and Brenden Gilbert. 2023. Exploring the Design Space of Automatically Generated Emotive Captions for Deaf or Hard of Hearing Users. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, Article 125, 10 pages. https://doi.org/10.1145/3544549.3585880

[14] Chika Ishii, Shuichi Kurabayashi, and Yasushi Kiyoki. 2011. A Dynamic Text-Decoration System for Text by Calculating Correlation between Impressive-Words and Fonts (in Japanese). In *Proceedings of the 3rd Forum on Data Engineering and Information Management (DEIM 2011)*. article E8–3.

[15] Robert H Lengel and Richard L Daft. 1988. The Selection of Communication Media as an Executive Skill. *Academy of Management Perspectives* 2, 3 (1988), 225–232. https://doi.org/10.5465/ame.1988.4277259

[16] Gil Levi and Tal Hassner. 2015. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. Association for Computing Machinery, 503–510. https://doi.org/10.1145/2818346.2830587

[17] Shinji Miyake and Masaharu Kumashiro. 1993. Subjective Mental Workload Assesment Technique-An Introduction to NASA-TLX and SWAT and a Proposal of Simple Scoring Methods (in Japanese). 人間工学 29, 6 (1993), 399–408.

[18] Eduardo Nacimiento-García, Carina S González-González, and Francisco L Gutiérrez-Vela. 2023. Automatic Captions on Video Calls: A Must for the Older Adults. *Universal Access in the Information Society* (2023), 1–24.

[19] Jay F. Nunamaker, Alan R Dennis, Joseph S Valacich, Douglas Vogel, and Joey F George. 1991. Electronic Meeting Systems. *Commun. ACM* 34, 7 (1991), 40–61.

[20] James Ohene-Djan, Jenny Wright, and Kirsty Combie-Smith. 2007. Emotional Subtitles: A System and Potential Applications for Deaf and Hearing Impaired People. In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments: Assistive Technology for All Ages (CVHI 2007)*. Citeseer.

[21] Mei-Hua Pan, Naomi Yamashita, and Hao-Chuan Wang. 2017. Task Rebalancing: Improving Multilingual Communication with Native Speakers-Generated Highlights on Automated Transcripts. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, 310–321. https://doi.org/10.1145/2998181.2998304

[22] Yingxin Pan, Danning Jiang, Michael Picheny, and Yong Qin. 2009. Effects of Real-time Transcription on Non-native Speaker's Comprehension in Computer-mediated Communications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, 2353–2356. https://doi.org/10.1145/1518701.1519061

[23] Yingxin Pan, Danning Jiang, Lin Yao, Michael Picheny, and Yong Qin. 2010. Effects of Automated Transcription Quality on Non-native Speakers' Comprehension in Real-time Computer-mediated Communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, 1725–1734. https://doi.org/10.1145/1753326.1753584

[24] Paige Rodeghero, Thomas Zimmermann, Brian Houck, and Denae Ford. 2021. Please Turn Your Cameras on: Remote Onboarding of Software Developers During a Pandemic. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. 41–50. https://doi.org/10.1109/ICSE-SEIP52600.2021.00013

[25] Kento Sato, Nana Mitomi, Keisuke Kon, and Hirokazu Haruna. 2022. 義肢装具領域におけるSystem Usability Scale (SUS) の信頼性の検討(in Japanese). *The Journal of the Japanese Academy of Prosthetists and Orthotists* 30, 1 (2022), 32–37.

[26] Klaus R Scherer. 2005. What Are Emotions? And How Can They Be Measured? *Social Science Information* 44, 4 (2005), 695–729.

[27] John Short, Ederyn Williams, and Bruce Christie. 1976. *The Social Psychology of Telecommunications.* John Wiley & Sons, Hoboken, New Jersey, USA.

[28] John C. Tang, Kori Inkpen, Sasa Junuzovic, Keri Mallari, Andrew D. Wilson, Sean Rintel, Shiraz Cupala, Tony Carbary, Abigail Sellen, and William A.S. Buxton. 2023. Perspectives: Creating Inclusive and Equitable Hybrid Meeting Experiences. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 351 (Oct 2023), 25 pages. https://doi.org/10.1145/3610200

[29] Joseph B Walther. 1992. Interpersonal Effects in Computer-mediated Interaction: A Relational Perspective. *Communication Research* 19, 1 (1992), 52–90.

[30] Chi-Lan Yang, Naomi Yamashita, Hideaki Kuzuoka, Hao-Chuan Wang, and Eureka Foong. 2022. Distance Matters to Weak Ties: Exploring How Workers Perceive Their Strongly- and Weakly-Connected Collaborators in Remote Workplaces. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 44 (Jan 2022), 26 pages. https://doi.org/10.1145/3492863

[31] Lin Yao, Ying-xin Pan, and Dan-ning Jiang. 2011. Effects of Automated Transcription Delay on Non-native Speakers' Comprehension in Real-time Computer-mediated Communication. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13.* Springer, 207–214.