Exploring the Design Space of Automatically Generated Emotive Captions for Deaf or Hard of Hearing Users

Saad Hassan
Rochester Institute of Technology
Rochester, NY, USA
sh2513@rit.edu

Yao Ding
Meta Platforms, Inc.
Menlo Park, CA, USA
yaoding@meta.com

Agneya Abhimanyu Kerure
Christi Miller
agneyakerure@meta.com
christim@meta.com
Meta Platforms, Inc.
Redmond, WA, USA

John Burnett
Meta Platforms, Inc.
Redmond, WA, USA
john.burnett.c@gmail.com

Emily Biondo
Meta Platforms, Inc.
Silver Spring, MD, USA
emilybiondo@gmail.com

Brenden Gilbert
Meta Platforms, Inc.
TX, USA
bgilbert73@gmail.com

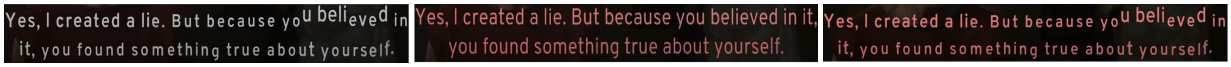

Figure 1: Caption text automatically rendered by emotion-detection models based on the three design schemas we explore: Typography (Left), Coloration (Middle), and Hybrid (Right). Additional visual details of the designs are provided in appendix.

ABSTRACT

Caption text conveys salient auditory information to deaf or hard-of-hearing (DHH) viewers. However, the emotional information within the speech is not captured. We developed three emotive captioning schemas that map the output of audio-based emotion detection models to expressive caption text that can convey underlying emotions. The three schemas used typographic changes to the text, color changes, or both. Next, we designed a Unity framework to implement these schemas and used it to generate stimuli videos. In an experimental evaluation with 28 DHH viewers, we compared DHH viewers' ability to understand emotions and their subjective judgments across the three captioning schemas. We found no significant difference in participants' ability to understand the emotion based on the captions or their subjective preference ratings. Open-ended feedback revealed factors contributing to individual differences in preferences among the participants and challenges with automatically generated emotive captions that motivate future work.

CCS CONCEPTS

• **Human-centered computing** → *Accessibility design and evaluation methods*; **Empirical studies in accessibility**.

KEYWORDS

captions, emotive captions, Deaf or hard of hearing users, Applications of emotion recognition

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585880

ACM Reference Format:
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), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3544549.3585880

1 INTRODUCTION

There are over 360 million Deaf or Hard of Hearing people worldwide [19]. In the U.S. alone, 15% of adults experience some degree of hearing loss [15]. Advances in automatic speech recognition have made it possible to generate high-quality captions, and many video streaming services and social media platforms allow content creators to add captions manually. Although there have been improvements in caption quality and usability, they still do not convey the acoustic or emotive cues present in speech, such as changes in loudness, pitch, or emotional tone. As a result, viewers who rely on captions often require assistance in understanding the emotions conveyed through speech. For DHH viewers, sign language interpreters often provide the missing audio-based emotive information in speech, but most online content is not interpreted. Occasionally, facial expressions give viewers an understanding of the speaker's emotional state, but it is not always accurate. Further, a speaker's facial expressions are sometimes absent in videos, e.g., in a documentary with a narrator in the background or in an action movie featuring a character who wears a mask.

Variations in speech intonation can signify several valuable pieces of information for the audience. Speech intonation is a vital part of the video-watching experience for audiences with access to audio, and it impacts how they perceive the video as well [16]. Therefore, it is important to investigate approaches that can enable

DHH viewers to access this missing information. Expressive captions are capable of conveying more than the text to users. Prior research has shown that hard of hearing users prefer enhanced animated text captions capable of showing intonation in speech over standard captions [21]. However, most of this prior work has focused on using prosodic information in speech to enhance captions [7]. Some prior work has evaluated the use of captions for conveying emotions, but it has usually been limited to a small set of discrete emotions [8]. Moreover, this work has largely employed wizard-of-oz techniques to generate stimuli videos to evaluate with users and assumed perfect emotion-detection.

Advancements in artificial intelligence (AI) have enabled emotion detection in audio using self-supervised speech representations and joint audio-visual information, but it's still imperfect. The use of recognition technology instead of wizard-of-oz approaches in the design and evaluation of emotive captions can help uncover challenges related to the imperfect nature of the underlying emotion recognition. In this work, we leverage two existing emotion recognition models to design emotive captions [25, 30]. It is important to clarify that our goal is not to compare emotive captions with standard captions, but rather explore the design space of emotive captions in order to determine the most effective designs for future studies that will compare them to the standard captions baseline.

There are two main contributions of this work:

- We present our process of designing three mappings schemas between different properties of emotions detected using state-of-the-art acoustic-emotion detection models and stylistic enhancements to caption text. We present findings from an evaluation with 28 DHH users, which highlight the need for further research to improve the legibility and understandability of automatically generated emotive captions.
- We release a Unity interface that would allow future researchers to design stimuli videos using output from emotion-recognition models to conduct evaluations with DHH users.

2 BACKGROUND AND PRIOR WORK

Prosody in speech conveys emotions through vocal inflections such as pitch, cadence, intensity, and volume. It can have its own meaning or alter the meaning of speech. A non-native speaker listening to a speaker, e.g., a football commentator or a poet, in another language can still decipher emotions. Similarly, the same sentence in English can change meaning based on how it is uttered, e.g., the sentence *"I can't believe you did that."* can convey happiness, shock, anger, or a combination of different emotions. This information is not conveyed in standard captions.

Prior work has investigated the use of typographic changes to convey prosody and emotions in written text. Some of this work has explored mapping sound features to typographic variables, e.g., font weight, size, style, etc. [6, 22, 28, 32]. Researchers have explored both automatically generating expressive text based on audio signals (like pitch), and other tools that allow artists to manipulate the text to convey emotions manually. Some of this research has been in the context of captions to convey prosody [7], or emotion [13, 21]. Research has explored using color-emotion associations in written text to express emotion-laden words, such as using red-colored text to convey anger [27]. In the context of caption text, it remains unknown whether typography, color, or a combination of both is more effective in expressing emotions in caption text.

Several prior studies have investigated visual factors and design variables that affect DHH viewers' perception of captioned videos. The legibility of captions remains an important concern for all caption styles, including styles that attempt to convey more than text. Amin et al. have investigated the issue of captions occluding crucial visual information on screen [3]. Other researchers looked at the effect of changing caption text width [9]. Researchers have also investigated the benefits of different types of enhancements to caption text, e.g., highlighting important words [12]. Most of the evaluations of emotive captions have been with hearing participants who are not the primary user group, so they fail to reveal potential challenges that DHH users might face with emotive captions [8]. Few studies that recruited DHH participants have largely conducted studies with young DHH students and did not include caption users with age-related hearing loss [2]. To the best of our knowledge, no prior research has investigated DHH viewers' perception of different automatically generated caption text designs that convey complex emotions in videos. We investigate designs of emotive captions as well as the challenges faced by DHH users in our two research questions.

RQ1 In a comparison between videos with captions stylistically enhanced to convey emotion using typography, coloration, or both, is there a difference between:
(a) DHH viewers' ability to determine the emotional valence of a video?
(b) their subjective judgments about whether captioned videos are useful and easy-to-follow?

RQ2 What challenges did participants face when viewing videos with emotive captions based on state-of-the-art emotion recognition models?

3 VIDEO STIMULI PREPARATION AND EMOTIVE CAPTIONS DESIGN TOOL

3.1 Phase 1: Caption Text Design

Our design team consisting of a Deaf accessibility expert, a product designer, and a Human-computer Interaction (HCI) researcher, held three design sessions to finalize the visual design of captions used in our study. The team aimed to pick a framework for depicting emotions and decided to use emotional state models that decompose emotions into sub-constituents, given the limitations of emotion-recognition models. The design sessions were conducted over video conference, and the team spent one hour on each session. The team picked the PAD emotional state model, which uses three numerical dimensions **P**leasure (how pleasant a speaker is?), **A**rousal (how energized a speaker is?), and **D**ominance (how in-control the speaker is?) to decompose and represent all emotions [4]. The PAD model of emotion includes a dominance dimension that distinguishes emotions with similar pleasure and arousal levels but different levels of dominance. It is an advanced version of the circumplex model of emotion experience [24], and it can help distinguish between dominant and submissive emotions such as anger and fear. Next, the team chose the visual design for the modulated captions with

input from a Deaf team member and previous research on speech-modulated typography. All the design choices were checked to be W3C Accessibility Guidelines (WCAG) 3.0 complaint [18][1].

3.1.1 Typography. A team member reviewed previous research on affective representation in typography and chose three typographic features to convey emotion in captions [6, 29, 32]. We used font-weight to depict emotion along the pleasure axis and hue to show polarity. For arousal, we used baseline shift (displacement of text above or below the line) to depict quietness of an utterance, based on previous research on speech-modulated typography [8]. We ensured that the spacing between two lines of caption text met the WCAG guidelines when using baseline shift. Finally, our third choice was font size to depict dominance was also motivated by prior research on prosodic mapping of visual properties of text [23, 29]. The font size variation was limited to a range of 4 pixels. Figures 5, 6, and 7 in the appendix describe our designs.

3.1.2 Coloration. Our second design interest was the use of color, inspired by prior research on color-emotion associations and the emotional effects of color dimensions (hue, saturation, and brightness) [11, 31]. We mapped the three color dimensions to three emotion dimensions (PAD) and used red for negative, green-blue for positive, and gray for neutral. We avoided using only green to ensure individuals with color vision deficiency could distinguish it from other colors. We mapped arousal to saturation to depict intensity and brightness (tone) to dominance. Figures 8, 9, and 10 in the appendix describe our designs.

3.1.3 Hybrid. The hybrid design in the study combined typography and coloration designs, allowing for free combination to ensure clear emotive information to viewers. Our choices for typography and coloration allowed free combination. In the case of pleasure, we used both weight and hue to depict positive or negative emotions and their intensity[2]. For depicting arousal, we use both baseline shift and changes in saturation. Similarly, we use changes in font size and brightness for dominance. Figures 11, 12, and 13 in the appendix describe our designs.

3.2 Phase 2: Selection of Videos

We selected 6 video genres, including documentaries, late-night shows, movies, news, sports, and social media videos [1]. From each genre, we chose 4 videos, extracted segments of 15-90 seconds, and used emotion detection models to process them [25]. To select the final set of videos, a researcher considered three factors: the videos' length, covering a range of 15-90 seconds to allow for feedback on at least six videos during the experimental session, and a wide range of values for three emotive sub-components: pleasure, arousal, and dominance. Table 1 in the appendix describes our stimuli videos.

3.3 Phase 3: Extraction of Emotion Features from Videos

We used the acoustic emotion detection model [25], to get the per bit pleasure and arousal ratings in speech. For dominance, we used a separate model [30]. In most cases, the values were normalized to a range between 0 and 1 except in case of arousal ratings used to implement baseline shift for the typography and hybrid conditions as well as pleasure ratings used to determine the hue in coloration and hybrid conditions. More details are shared in the appendix.

3.4 Phase 4: Generating Emotive Captions and Augmenting in Videos

We developed a framework in Unity to augment emotive captions into the videos. Figure 2 shows an annotated screenshot of the interface. The interface allowed us to define a subtitle, add text, add a corresponding file with the emotional information, and then render the captions with appropriate latency[3]. The tool also allows us to apply the stylistic enhancements to captions at different levels, including character, word, and selected syllables. We applied the enhancements at a syllable level to entire videos. This meant that the ratings corresponding to the audio bits for characters in each syllable were averaged out. We release this Unity framework with this paper. More details about the interface and how to use it are provided in the electronic supplement. We used the commonly used block approach for captioning (entire caption text showing at once) instead of a one word-at-a-time approach.

4 STUDY DESIGN

Participants signed IRB-approved consent forms before participating in a remote experimental study with 28 participants. The study used a within-subject design and was administered using Qualtrics survey. The first few pages of the survey familiarized the participants with our three emotive captioning schemas using examples. Participants had to acknowledge their understanding of the three schemas before continuing with the study. If they had any confusion, a research assistant was available for assistance. The rest of the pages on the survey consisted of a link to the video and the word containing the syllable participants had to focus on to respond to questions regarding emotions, followed by the questions on emotion ratings and subjective preference. Participants watched six videos for each of the three conditions, each from a different genre. The same six videos were used for all three conditions with different captioning styles [17]. Participants could re-watch videos if they liked to. The sequence of conditions and videos within each condition were Latin-square randomized across our 28 participants. Participants answered questions after watching each video to assess their understanding of the emotive content and their subjective preferences. For task performance questions, participants rated pleasure, arousal, and dominance on a 9-point scale adapted from the Self-Assessment Manikin (SAM) scale, with pleasure ranging from unpleasant to pleasant, arousal from calm to excited, and dominance from being controlled to in control. The following two subjective judgment Likert-scale questions were adapted from [2].

(1) I found these captions useful. (From *Strongly Disagree* to *Strongly Agree*)
(2) It was easy to follow the captions. (From *Strongly Disagree* to *Strongly Agree*)

The study included an open-ended feedback field on each survey page, and a "final thoughts survey" that asked participants to select

[1] https://www.w3.org/WAI/GL/WCAG3/2022/how-tos/captions/
[2] Note that we also had hue in the typography condition, but it was only three different shades of gray. For the hybrid condition, we only take the font size for our combination
[3] The resource exists as a project file instead of an executable.

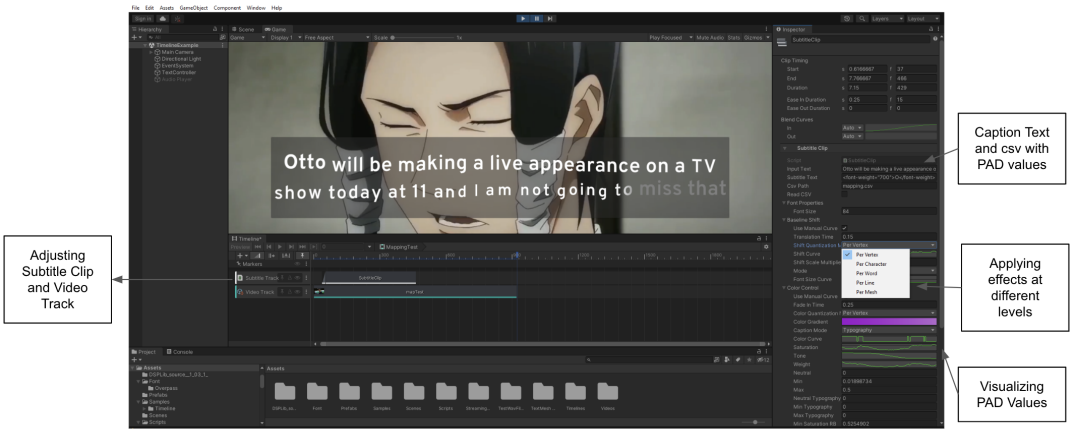

Figure 2: Unity framework for generating emotive captions based on acoustic-emotive data. The display includes a timeline editor and a signal visualizer on the right (Size of the caption text has been increased to easily test the variations in styles).

their favorite schema. On average participants took 74 minutes to finish the three surveys and 26 minutes to understand the schemas[4].

4.1 Recruiment and Participants

We had two key recruiting criteria that informed our screening questions: (1) "Do you identify as Deaf or Hard of Hearing or an individual with hearing loss?" and (2) "Do you use captioning when viewing videos or television?" Our 28 participants included 11 men, 16 women, and one individual who identified as non-binary. Our IRB only allowed asking participants' age range, and they were: 18-24 (1), 25-34 (4), 35-44 (8), 45-54 (6), 55-64 (3), and 65+ (6). Six participants reported having severe hearing loss; eight said moderate, three reported mild, and eleven did not specify. Nine participants reported that American Sign Language (ASL) is their primary language, whereas 19 participants said English. Participants received $250 for participation in the study.

5 FINDINGS

5.1 RQ1: Comparison of Task Performance and Subjective Judgements

Our question responses were collected with 5-point Likert response scales or 9-point responses in case of emotion sub-constituents. We considered and modeled this data as ordered categorical (ordinal) responses. To analyze the data and account for the ordinal nature, we built cumulative model (CM) frameworks for each response question [14]. We constructed separate hierarchical multivariate ordinal regression models of response outcome under a Bayesian framework for each item using an R package [10]. We also included three conditions, six genres, and their interaction as population-level effects with varying (group-level) effects of participants. More details about the model design are provided in the appendix B. Figure 3 displays the results of the task performance questions, while Figure 4 displays the results of the two subjective judgment questions, for all three captioning conditions.

Our findings did not reveal a significant difference across the subjective questions' conditions. However, the median model outcome ratings were highest for the typography condition. Our analysis of the three emotion ratings for the three conditions also did not reveal any significant differences. We found that arousal was best picked by participants across all the three conditions followed by dominance. Surprisingly, the pleasure sub-component was not picked well by the participants even in conditions where we used discrete colors to depict positive, negative, and neutral emotions. On the final survey 13 out of our 28 participants (including 5 out of 6 participants over the age of 60) preferred coloration, 11 preferred typography, and 4 preferred hybrid.

5.2 RQ2: Challenges Experienced

Open-ended feedback from participants was largely positive. Participants mentioned how the emotive captions allowed them to understand the emotive aspects of speech that they cannot using standard captions: *"they did a better job of conveying the tension and emotionality of the scene" - P9.* We primarily focused on challenges reported by participants to motivate future work on enhancing emotive captions, rather than thematically analyzing the entire open-ended data. Eight participants reported that emotive captions could be more distracting than standard captions and take away their attention from the important content. P11 explained, *"I needed to take my focus off the captions ... to look at the images in each scene. The other captions with color changing took all of my screen time to process and didn't allow me any time to watch the show."* Distraction was a bigger issue in the hybrid condition where typographic styles were combined with color changes. P3 commented, *"I'm liking the colors of red to convey how he is talking, but the other subscripts bothered me."* Participants also pointed out how distraction was more problematic for certain genres, e.g., short-form videos: *"It felt like the captions were the star of the video and not the singer." - P9*

[4]https://www.qualtrics.com/support/survey-platform/survey-module/editing-questions/question-types-guide/advanced/timing/

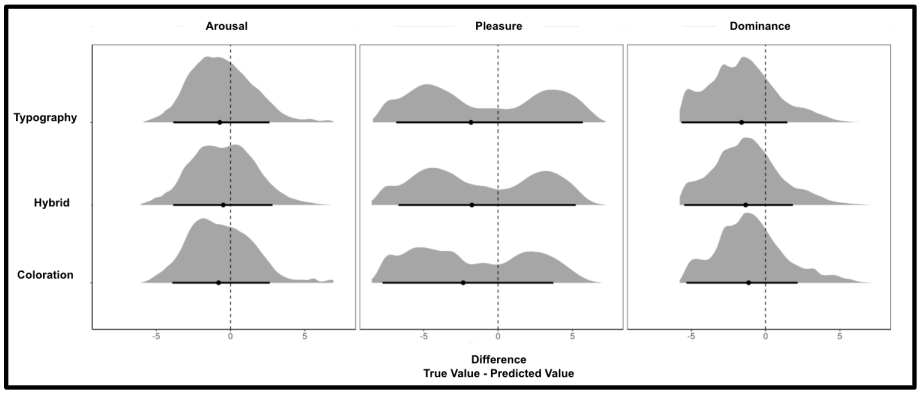

Figure 3: Comparison of the model outputs for three conditions across emotion sub-component rating questions. The graphs show the distribution of differences between the True Value (what acoustic-emotion detection model predicted and was used to modulate the captions) and Predicted Value (what the cumulative model based on participant feedback predicted). A difference of 0 would show that participants assessment of emotion sub-component is exactly the same as what caption showed.

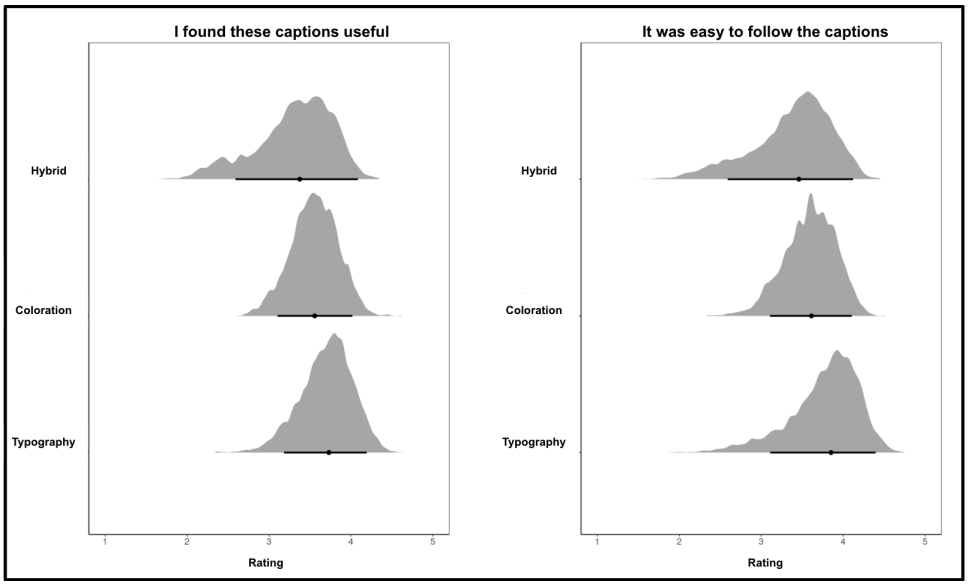

Figure 4: Comparison of the model response outcome based on participants' responses to the two subjective judgment questions. The higher variation in the hybrid condition might be due to some participants finding the captions distracting.

Participants also shared feedback on the preferred granularity and their perception of changes to caption text. Some participants, e.g., P8, did not like the enhancements applied at a syllable level: *"It was really distracting to read some words a raised or lowered, I would like all same font and size per word."* Participant 18 commented that they found it hard to notice any changes to the caption text in the typography condition where no color was used. They stated, "I didn't really notice a difference from the first time I watched this." Participants also highlighted how caption occlusion can be exacerbated due to the use of color in emotive captions and longer staying background panels: *"At times they covered up the news logo, so I would prefer they go to the top so I don't miss anything but these are the best I've seen so far, absolutely." - P9.* Six participants aged 65 and above preferred the coloration schema over typography and hybrid schemas because it was less distracting. P4 commented, *"I did not like the font, but the coloring was okay, and I liked that it was solid and didn't move."* Three participants suggested providing emotive captions on-demand: *"I don't like the color changing but I can see how it could be used in a creative way... for something important I would hate to see color changing..." - P27.*

6 HOW DO OUR FINDINGS MOTIVATE FUTURE WORK ON EMOTIVE CAPTIONS?

Our work can be best conceived as a prequel to future research on fully automatic emotive captions. It motivates research and provokes novel conversations on at least three unsolved problems

in the domain of emotive captions. Our released Unity tool can be used to generate video stimuli for proposed future studies.

Our results revealed that all three designs roughly perform the same in terms of both identifying emotions and subjective preferences. The typography condition performed slightly better, but the results were insignificant. Surprisingly, we found that emotive captions did not influence participants' perception of the pleasure sub-component of emotion in the video (as evident from figure 3). Prior research has shown that both audio and visual information channels affect viewers' emotional experience while watching television [26]. Previous research has also shown that genre of a video affects the emotional reception from the viewers [20]. Perhaps some deaf participants may have used preconceived notions about the content of the video when rating pleasure, leading to higher ratings than what was indicated by the captions. Specifically, the ratings for the documentary video were much higher than expected, possibly due to the pleasant nature of the content, which may have caused participants to disregard the emotive captions. This suggest that future studies should **explore strategies to guide users in disambiguating between emotions in the audio channel and emotional valence in the visual channel**.

Our qualitative findings revealed that participants found our emotive captions useful for determining affective connotation in speech, especially for movies, short-form videos, and documentary genres. Some participants found the stylistic enhancements to be distracting, especially older adults with age-related hearing loss. While the 14-point font size was larger than what older adults are accustomed to, changes to font shape and weight were still problematic for them. Prior research has shown that font type can affect text legibility and reading time among older adults [5]. Feedback from these participants suggests that emotive captions should always be available as an option rather than the default in video streaming services. Future research should **investigate design approaches to reduce text legibility and distraction concerns with emotive captions**, evaluating the legibility of different typographic and coloration changes to pick the least distracting ones.

Finally, the study was the first to use emotion recognition models to generate emotive captions for evaluation. However, the models were imperfect and the pleasure, arousal, and dominance values varied abruptly due to rapidly changing emotions in speech. The caption design tool allows for normalization of the output of the acoustic-emotion detection model and application at various levels, such as character vertex, whole character, selected syllable, whole word, whole line, and individual subtitle. However, our choice of applying it at the syllable level caused some participants to find the changes in the caption text too frequent. Future research needs to **investigate differences in legibility and subjective preferences when stylistic enhancements to caption text are applied at different levels**. Future research can also consider different smoothing or emotion signal normalization techniques to help the users cope with the imperfect emotion recognition models.

7 CONCLUSION

Standard captions fail to provide underlying emotive information in speech to DHH participants. Emotive captions based on automatic acoustic-emotion detection models can provide DHH viewers access to much-needed emotive sub-text in speech. We present findings from a study that investigates three designs for emotive captions with 28 DHH participants. Although no statistically significant differences were observed across conditions, our qualitative findings revealed why participants preferred different designs and uncovered challenges related to the legibility and understandability of emotive captions, suggesting at least three avenues for future research. Our publicly released emotive caption-generating tool would allow future researchers to experiment with different design configurations of emotive captions to support experimental studies.

ACKNOWLEDGMENTS

We would like to thank Amanda Stump and Sokol Zace from Contact Design, and Sarah Partridge from IPSOS UX for helping with recruitment and data collection. We would also like to thank Scott Selfon, Muhammad Ikram, Khia Johnson, Jan Zikes, Stavros Petridis, W. Owen Brimijoin, Zachery Schramm, and Jeff Crukley from Meta Platforms, Inc. for feedback on various stages of this project. This material is also partially supported by the National Science Foundation under Grant numbers 2235405, 2212303, 1954284, and 2125362, and by the Department of Health and Human Services under Award No. 90DPCP0002-0100.

REFERENCES

[1] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Caption-Occlusion Severity Judgments across Live-Television Genres from Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 18th International Web for All Conference* (Ljubljana, Slovenia) *(W4A '21)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3430263.3452429

[2] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Effect of Occlusion on Deaf and Hard of Hearing Users' Perception of Captioned Video Quality. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 202–220.

[3] Akhter Al Amin, Saad Hassan, Sooyeon Lee, and Matt Huenerfauth. 2022. Watch It, Don't Imagine It: Creating a Better Caption-Occlusion Metric by Collecting More Ecologically Valid Judgments from DHH Viewers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 459, 14 pages. https://doi.org/10.1145/3491102.3517681

[4] Robert Freed Bales. 2017. *Social interaction systems: Theory and measurement*. Routledge, Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK.

[5] Michael Bernard, Chia Hui Liao, and Melissa Mills. 2001. The Effects of Font Type and Size on the Legibility and Reading Time of Online Text by Older Adults. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (Seattle, Washington) *(CHI EA '01)*. Association for Computing Machinery, New York, NY, USA, 175–176. https://doi.org/10.1145/634067.634173

[6] João Couceiro e Castro, Pedro Martins, Ana Boavida, and Penousal Machado. 2020. «Máquina de Ouver»-From Sound to Type: Finding the Visual Representation of Speech by Mapping Sound Features to Typographic Variables. In *Proceedings of the 9th International Conference on Digital and Interactive Arts* (Braga, Portugal) *(ARTECH 2019)*. Association for Computing Machinery, New York, NY, USA, Article 13, 8 pages. https://doi.org/10.1145/3359852.3359892

[7] Caluã de Lacerda Pataca and Paula Dornhofer Paro Costa. 2023. Hidden Bawls, Whispers, and Yelps: Can Text Convey the Sound of Speech, Beyond Words? *IEEE Transactions on Affective Computing* 14, 1 (2023), 6–16. https://doi.org/10.1109/TAFFC.2022.3174721

[8] Caluã de Lacerda Pataca and Paula Dornhofer Paro Costa. 2020. Speech Modulated Typography: Towards an Affective Representation Model. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 139–143. https://doi.org/10.1145/3377325.3377526

[9] Abraham Glasser, Joseline Garcia, Chang Hwang, Christian Vogler, and Raja Kushalnagar. 2021. Effect of Caption Width on the TV User Experience by Deaf and Hard of Hearing Viewers. In *Proceedings of the 18th International Web for All Conference* (Ljubljana, Slovenia) *(W4A '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 5 pages. https://doi.org/10.1145/3430263.3452435

[10] Rainer Hirk, Kurt Hornik, and Laura Vana Gür. 2020. mvord: an R package for fitting multivariate ordinal regression models. *Journal of Statistical Software* 93, 4 (2020), 1–41.

[11] Domicele Jonauskaite, Ahmad Abu-Akel, Nele Dael, Daniel Oberfeld, Ahmed M Abdel-Khalek, Abdulrahman S Al-Rasheed, Jean-Philippe Antonietti, Victoria Bogushevskaya, Amer Chamseddine, Eka Chkonia, et al. 2020. Universal patterns in color-emotion associations are further shaped by linguistic and geographic proximity. *Psychological Science* 31, 10 (2020), 1245–1260.

[12] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the Benefit of Highlighting Key Words in Captions for People Who Are Deaf or Hard of Hearing. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 43–55. https://doi.org/10.1145/3308561.3353781

[13] Daniel G Lee, Deborah I Fels, and John Patrick Udo. 2007. Emotive captioning. *Computers in Entertainment (CIE)* 5, 2 (2007), 11.

[14] Torrin M Liddell and John K Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* 79 (2018), 328–348.

[15] Frank R Lin, John K Niparko, and Luigi Ferrucci. 2011. Hearing loss prevalence in the United States. *Archives of internal medicine* 171, 20 (2011), 1851–1853.

[16] Sydney L Lolli, Ari D Lewenstein, Julian Basurto, Sean Winnik, and Psyche Loui. 2015. Sound frequency affects speech emotion perception: Results from congenital amusia. *Frontiers in Psychology* 6 (2015), 1340.

[17] Jon D Morris. 1995. Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research* 35, 6 (1995), 63–68.

[18] Liddy Nevile and Brian Kelly. 2008. Web Accessibility 3.0: learning from the past, planning for the future.

[19] World Health Organization. 2022. Deafness and hearing loss. https://www.who.int/health-topics/hearing-loss#tab=tab_1

[20] Deidre Pribram. 2012. *Emotions, genre, justice in film and television: Detecting feeling*. Routledge, Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK.

[21] Raisa Rashid, Quoc Vy, Richard Hunt, and Deborah I Fels. 2008. Dancing with words: Using animated text for captioning. *Intl. Journal of Human–Computer Interaction* 24, 5 (2008), 505–519.

[22] Tara Rosenberger and Ronald L. MacNeil. 1999. Prosodic Font: Translating Speech into Graphics. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania) *(CHI EA '99)*. Association for Computing Machinery, New York, NY, USA, 252–253. https://doi.org/10.1145/632716.632872

[23] Tara Rosenberger-Shankar. 1998. *Prosodic Font: The space between the spoken and the written*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[24] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[25] Abhinav Shukla, Stavros Petridis, and Maja Pantic. 2020. Learning speech representations from raw audio by joint audiovisual self-supervision. *arXiv preprint arXiv:2007.04134* 1 (2020), 8 pages.

[26] Johnny V Sparks, Wan-Chu Chuang, and Sungwon Chung. 2012. Continuous emotional responding to audio, video, and audiovisual sensory channels during television viewing. *Southwestern Mass Communication Journal* 28, 1 (2012), 31 pages.

[27] Tina M Sutton and Jeanette Altarriba. 2016. Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behavior research methods* 48, 2 (2016), 686–728.

[28] Dimitrios Tsonos and Georgios Kouroupetroglou. 2016. Prosodic mapping of text font based on the dimensional theory of emotions: a case study on style and size. *EURASIP Journal on Audio, Speech, and Music Processing* 2016, 1 (2016), 1–16.

[29] Dimitrios Tsonos and Georgios Kouroupetroglou. 2016. Prosodic mapping of text font based on the dimensional theory of emotions: a case study on style and size. *EURASIP Journal on Audio, Speech, and Music Processing* 2016, 1 (2016), 1–16.

[30] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn Schuller. 2022. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *ArXiv* abs/2203.07378 (2022), 25 pages.

[31] Lisa Wilms and Daniel Oberfeld. 2018. Color and emotion: effects of hue, saturation, and brightness. *Psychological research* 82, 5 (2018), 896–914.

[32] Matthias Wölfel, Tim Schlippe, and Angelo Stitz. 2015. Voice driven type design. In *2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, Bucharest, Romania, 1–9. https://doi.org/10.1109/SPED.2015.7343095

Figure 5: Key for depicting "Pleasure" dimension in the Typography schema. The color becomes more dull gray as we go from positive to negative. The weight of the font shows the magnitude of the "Pleasure" dimension on a normalized -1 (extremely negative) to 1 (extremely positive) range.

Figure 6: Key for depicting "Arousal" dimension in the Typography schema. Baseline shift is used to move the caption text above the baseline if there is high speech arousal and downwards if there is low arousal on a normalized -1 to 1 range.

Figure 7: Key for depicting "Dominance" dimension in the Typography schema. The size of the caption text increases to show higher dominance and decreases to show lower dominance on a normalized 0 to 1 range.

A SCHEMAS

A.1 Typography Schema

Figures 5, 6, and 7 show how captions were modulated based on the pleasure, arousal, and dominance values predicted by emotion-detection models.

A.2 Coloration Schema

Figures 8, 9, and 10 show how captions were modulated based on the pleasure, arousal, and dominance values predicted by emotion-detection models.

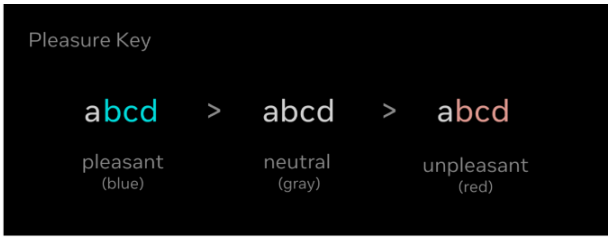

Figure 8: Key for depicting "Pleasure" dimension in the Coloration schema. There are three discrete colors corresponding to three ranges in the -1 to 1 normalized pleasure rating scale: green-blue (greater than or equal to 0.333), gray (between -0.333 and 0.333), and red (less than or equal to -0.333).

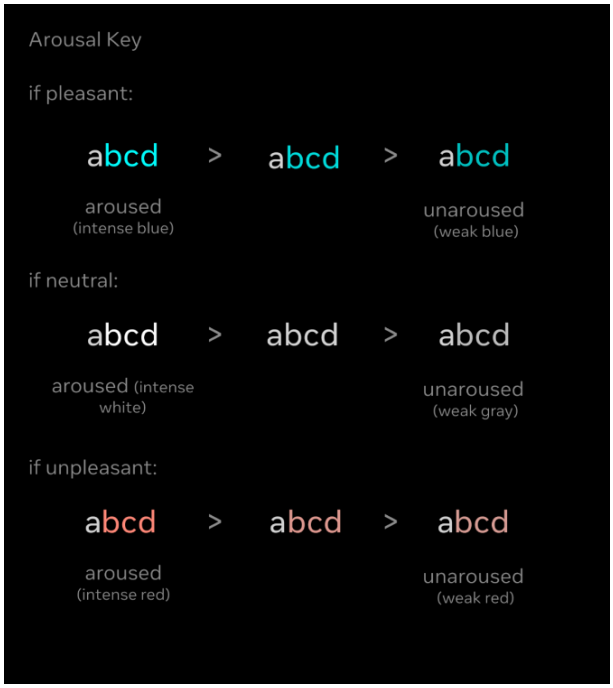

Figure 9: Key for depicting "Arousal" dimension in the Coloration schema. Saturation was used for all three colors to depict arousal.

A.3 Hybrid Schema

Figures 11, 12, and 13 show how captions were modulated based on the pleasure, arousal, and dominance values predicted by emotion-detection models.

B BAYESIAN MODELING

The cumulative models assume that the observed ordinal variable Y, the Likert response rating or response on the 9-point scale, originates from categorizing a latent (not observable) continuous variable $\hat{Y}$. To model this categorization process, the CMs assume that there are K thresholds τ_k which partition $\hat{Y}$ into K+1 observable, ordered categories of Y. In our case, there are K+1=10 response

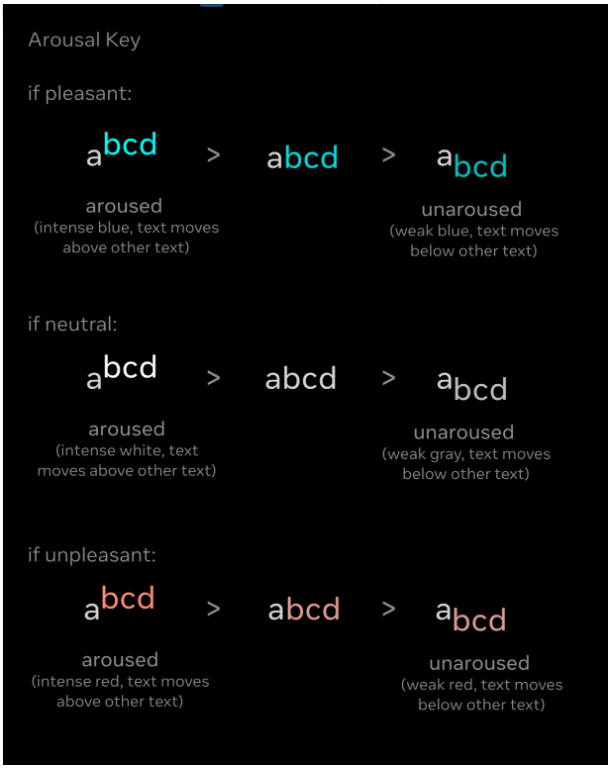

Figure 12: Key for depicting "Arousal" dimension in the Hybrid schema. Saturation was used for all three colors to depict arousal. In addition, baseline shift is used to move the caption text above the baseline if there is high speech arousal and downwards if there is low arousal on a normalized -1 to 1 range.

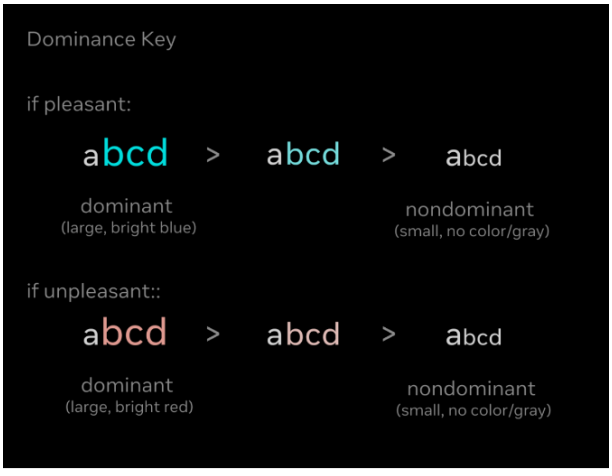

Figure 13: Key for depicting "Dominance" dimension in the Hybrid schema. Tone changes were used for all three colors to depict dominance. Further, the size of the caption text increases to show higher dominance and decreases to show lower dominance on a normalized 0 to 1 range.

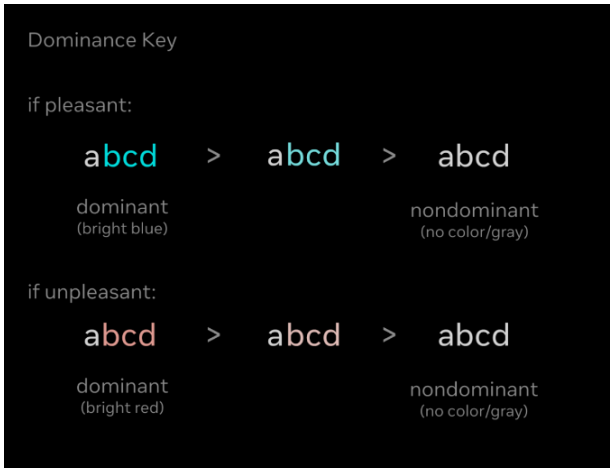

Figure 10: Key for depicting "Dominance" dimension in the Coloration schema. Tone changes were used for all three colors to depict dominance.

Figure 11: Key for depicting "Pleasure" dimension in the Hybrid schema. There are three discrete colors corresponding to three ranges in the -1 to 1 normalized pleasure rating scale: green-blue (greater than or equal to 0.333), gray (between -0.333 and 0.333), and red (less than or equal to -0.333). Further, the weight of the font shows the magnitude of the "Pleasure" dimension on a normalized -1 (extremely negative) to 1 (extremely positive) range.

categories for emotion questions, and K+1=5 response categories for Likert questions, and therefore giving us K=9 (or K=4) thresholds. If we assume $\hat{Y}$ to have a certain distribution (e.g., a normal distribution) with cumulative distribution function F, we can write down the probability of Y being equal to category k via: $Pr(Y = k) = F(\tau_k) - F(\tau_k)$. To expand this into a regression, we formulate a linear regression for $\hat{Y}$ with predictor term:

$\eta = \beta_1 x_1 + \beta_2 x_2 + ...$ so that $\hat{Y} = \eta + \epsilon$ where ϵdescribes the error term of the regression. Consequently, $\hat{Y}$ is split into two parts. The first one (η) represents variation in $\hat{Y}$ that can be explained by the predictors, and the second one (ϵ) represents variation that remains unexplained.

C VIDEO STIMULI

Table 1 describes the stimuli videos used.

Genre	Video Description
News	Video segment from a PBS News broadcast on forest fires.
Documentary	Video segment from a National Geographic documentary on a rock-climbing goat crossing a river with her kid.
Late night show	Video segment from a Conan O'Brian show with a guest containing sarcasm and humor.
Sports	Video segment from an American football game between Los Angeles Rams and Cincinnati Bengals.
Movies	Video segment from the movie V for Vendetta showing a tense conversation between V and Evey Hammond.
Short-form Video	A segment of a TikTok from a dataset shared on Kaggle showing a girl singing Upside Down by JVKE.

Table 1: Descriptions of the videos used in the user study for all six genres.