Caption Royale: Exploring the Design Space of Affective Captions from the Perspective of Deaf and Hard-of-Hearing Individuals

Caluã de Lacerda Pataca
Computing and Information Science
Rochester Institute of Technology
Rochester, NY, USA
calua.pataca@gmail.com

Saad Hassan
Department of Computer Science
Tulane University
New Orleans, LA, USA
saadhassan@tulane.edu

Nathan Tinker
Department of ASL and Interpreting Education
Rochester Institute of Technology
Rochester, NY, USA
ntt3752@rit.edu

Roshan L Peiris
Matt Huenerfauth
School of Information
Rochester Institute of Technology
Rochester, NY, USA
rxpics@rit.edu
matt.huenerfauth@rit.edu

ABSTRACT

Affective captions employ visual typographic modulations to convey a speaker's emotions, improving speech accessibility for Deaf and Hard-of-Hearing (DHH) individuals. However, the most effective visual modulations for expressing emotions remain uncertain. Bridging this gap, we ran three studies with 39 DHH participants, exploring the design space of affective captions, which include parameters like text color, boldness, size, and so on. Study 1 assessed preferences for nine of these styles, each conveying either valence or arousal separately. Study 2 combined Study 1's top-performing styles and measured preferences for captions depicting both valence and arousal simultaneously. Participants outlined readability, minimal distraction, intuitiveness, and emotional clarity as key factors behind their choices. In Study 3, these factors and an emotion-recognition task were used to compare how Study 2's winning styles performed versus a non-styled baseline. Based on our findings, we present the two best-performing styles as design recommendations for applications employing affective captions.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility technologies**; *Empirical studies in accessibility*.

KEYWORDS

Accessibility, Caption, Emotion, Accessibility for people who are Deaf and Hard-of-Hearing

ACM Reference Format:
Caluã de Lacerda Pataca, Saad Hassan, Nathan Tinker, Roshan L Peiris, and Matt Huenerfauth. 2024. Caption Royale: Exploring the Design Space of Affective Captions from the Perspective of Deaf and Hard-of-Hearing Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642258

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642258

1 INTRODUCTION

Captions are a widely adopted strategy to make speech accessible [12, 46, 48, 49]. Recent advancements in automatic speech recognition have marked significant progress in enhancing the quality of computer-generated captions, thereby expanding their range of applications [46]. However, even when transcriptions are accurate and synchronization precise, captions frequently lack paralinguistic cues. This means that they present speech – an expressive and nuanced form of communication – in a flat manner, conveying salient verbal cues but falling short of depicting nonverbal elements.

Previous research has highlighted the adverse impact that this absence of paralinguistic cues has on the viewing experience of captioned content among DHH individuals [24, 32, 39]. Addressing this, different researchers have explored approaches to convey paralinguistic information through stylistic modulations in typography [9, 22, 23, 62, 79]. Much of this prior work has focused on conveying aspects of speech such as pitch, rhythm, or loudness, i.e., prosody. However, a recent empirical study involving DHH individuals showed that a more effective approach involves directly representing the emotional content of a speaker's paralanguage, rather than solely focusing on its acoustic qualities [24].

Yet, little is known about how to actually *design* affective captions. This stems from there being a gap in systematic explorations of their design space, which is different from that of prosodic captions, e.g., [22, 79]. What little research there is (e.g., [39]) does not focus primarily on DHH individuals' perspectives on what caption styles are preferred and perform better at conveying emotions. This is especially true for scrolling captions, i.e., those that are written one word at a time, as are commonly used for automatically generated captions [24]. In this paper, we address this gap by reporting on three studies that investigated the preferences of DHH caption users for different caption styles and how effective they were in conveying a speaker's emotions to DHH participants.

Our contributions are empirical:

- In Study 1, we compared nine different caption styles for their ability to independently represent valence or arousal. Our statistical analysis identified two styles tied for first place for valence (font-color and shadow-color) and four for arousal (shadow-color, font-size, font-color, and font-weight).
- In Study 2, we examined six combinations of styles (based on the winning styles from the first study) for representing valence and arousal simultaneously. Our results indicated a four-way tie for first place (font-color with font-weight, font-color with font-size, font-color with shadow-color, and shadow-color with font-size).
- Based on participants' feedback, we outlined EASE OF READING, LOW DISTRACTION, INTUITIVENESS, and CLARITY OF EMOTIONAL REPRESENTATION as key factors for deciding whether a given caption style would be preferred or not.
- In Study 3, we compared the four top-performing styles from the previous study against an unstyled baseline condition. We found that both font-color with font-weight and font-color with font-size perform well, objectively and subjectively, and offer both as design recommendations for researchers and designers of affective captioning applications.

The three studies presented herein can be likened to a Battle Royale-style competition, where we systematically filtered an initial pool of 72 possible combined styles down to a final selection of just 2 winning options. Figure 1 illustrates this process.

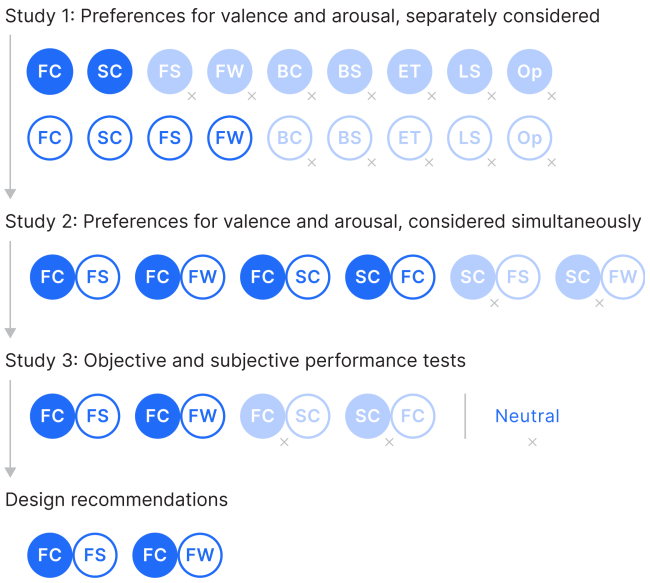


Figure 1: Map of the three studies and our design recommendations. Valence styles are represented by filled circles and arousal by outlined circles. The typographic modulations are abbreviated as follows: BC: background-color; BS: baseline shift; ET: emotional typeface. FC: font-color; FS: font-size; FW: font-weight; LS: letter spacing; OP: opacity; and SC: shadow-color; Each study eliminated certain styles from consideration, marked by faded colors and a cross symbol (×).

2 BACKGROUND AND RELATED WORK

2.1 Visualizing Speech Properties Using Stylistic Representations in Text

Over the years, researchers have explored how visual modifications in typography can serve as a means of infusing text with an additional layer of information, encompassing not only its conventional linguistic content but also dimensions of prosody and emotions. These alterations include adjustments in letter forms and typesetting parameters [14, 59, 79], as well as the incorporation of complementary visual elements. Research in this field has shown that readers can effectively assimilate some of these changes into their reading patterns [9], suggesting that these modulations can be a useful approach to overlay meaning on written text.

A recent study proposes two models of speech-modulated typography aimed at allowing the conversion of expressive speech into written text. Using pairwise comparisons, it gauged participants' preferences for various typographic modulations, including font weight, letter width, letter slant, and baseline-shift, used as a means of depicting utterances with five emotions. The study found that font-weight was favored for intense utterances, while baseline shift was preferred for quieter ones [23]. These findings suggest that participants' preferences can converge around the use of specific typographic parameters to convey paralinguistic aspects of speech. In that vein, previous studies have also explored the use of color-emotion associations in written text to convey emotions [42, 62], such as employing red-colored text to express anger [64].

Combining multiple modulations, the Kinedit system allows users to manipulate typographic attributes like font-size, color, position, and rotation to convey emotions, prosody, direction of attention, characters, and more [27]. This system was later expanded to accommodate instant messaging scenarios [10].

Not all studies are necessarily based on modulating preexisting typographic parameters. Promphan, for instance, designed a font where the actual shapes of each character interpolated along a continuous spectrum from negative to positive valence, shifting from harsh and spiky shapes for the former to soft and rounded ones for the latter [54].

Other approaches aim to assist specialized audiences in prosodic analysis. In these cases, authors have developed precise visualizations of speech that mix traditional textual transcription with graphical elements representing acoustic features such as pitch, energy, and rhythm. Despite their accuracy, these approaches might prove challenging to comprehend for those unversed in the specialized conventions of fields such as linguistics [1, 51].

While these studies provide an initial starting point when it comes to representing a speaker's prosody and emotions, they mostly focus on text for print or online settings. Captions have unique requirements and constraints, such as the need to synchronize the text with corresponding audio or visual content, limited space and time for displaying text, distinct needs for legibility and readability, and the potential for distractions or occlusion caused by other on-screen elements. These factors constrain the generalizability of findings from research on prosodic and affective representations in conventional long-form text, which calls for more caption-focused research.

2.2 Enhancing Caption Text to Improve Viewer's Experience of Captioned Media

In addition to improving the transcription accuracy of captions, there has been a recent focus in the fields of HCI and accessibility research on enhancing the usability and presentation of captions. This includes exploring ways to make captions more informative, visually engaging, and easy to read, while still maintaining their primary function of conveying spoken content to DHH individuals. For example, recent studies have investigated the importance of identifying the current speaker in a panel discussion [4], as well as the benefits of inserting correct punctuation or pauses in captioned videos to improve readability for DHH viewers [29, 73].

Previous research has investigated the benefits and various approaches to highlighting important words in caption texts, as well as examining the most effective styles to achieve this [37]. Additionally, researchers have explored how various aspects of caption text appearance, including styles, font, and background, can influence DHH users' subjective impression of caption quality and readability [7, 19]. Proper segmentation, which aligns caption boundaries with syntactic boundaries, has also been found to improve caption readability [73]. Researchers have explored captioning approaches that place captions in regions that cause the least interference with important on-screen information, in order to mitigate the occlusion of caption text [2, 3].

In general, it has been shown that DHH individuals appreciate caption enhancements that improve the informativeness and engagement of the content, provided that they do not hinder the primary function of the caption text. This highlights the importance of running empirical studies with DHH participants to investigate key performance and design variables.

Affective captions have been investigated as a means of enhancing the overall viewing experience by conveying the emotional tone of the speaker in addition to their spoken words. Rashid et al. presented a study in which artists collaborated to create animated closed captions that visually represented categorical emotions present in speech. Although the enhanced captions did not lead to better emotion recognition compared to a control group with traditional captions, both hearing and DHH participants expressed their preference for the new enhanced captions [58].

A recent CHI study investigated the limitations of standard captions in live meetings by interviewing DHH individuals. It showed that standard captions lack emotional depth, resulting in a loss of emotional cues, leaving DHH individuals feeling alienated from their hearing peers. It then compared captioning models incorporating prosody, emotions, and a combination of both, and found that the emotion-based model was preferred by DHH viewers [24]. In this same work, the authors recommended the use of multi-dimensional models of emotion, like the circumplex model, which categorizes emotions along two dimensions: valence, indicating pleasantness or unpleasantness, and arousal, reflecting intensity [60]. This model, they argued, provides a comprehensive way to represent emotions, capturing a continuum of emotional experiences that still allows for the mapping of discrete emotions as specific coordinates. For example, anger can be mapped to low valence and high arousal, while relief maps to high valence and low arousal, etc [34].

The most closely related study to ours is a recent CHI Late-Breaking work that explored the use of color, typography, and their combination to visualize pleasure, arousal, and dominance in speech, based on the PAD (circumplex with an additional dominance dimension) emotional model. However, the study did not find any significant differences among the styles tested, nor did it measure its affective captioning approaches versus a non-styled baseline. A qualitative analysis of the data collected showed a preference for color-based affective caption approaches, while also highlighting concerns regarding legibility, distraction, and interpretability of affective captions [32].

Despite the recent research on affective captions, there is still no consensus on the best way to represent emotional properties in caption text. Research in this area often borrows from more general studies on paralinguistic representation in text, but captions have enough specificities for this extrapolation of knowledge to be challenging, e.g., captions are an animated medium with unique challenges regarding legibility, readability, how much they might suffer from or cause distractions with other on-screen content, etc. Our research bridges this gap by developing a range of typographic styles specifically designed for use in captions that depict affective dimensions of speech. We then have these styles compete among themselves for the preference of DHH viewers. In doing so, we seek to address the following research questions:

RQ1. Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal…
  A *… when depicted individually?*
  B *… when depicted in combination?*

We divided the first research question into A and B parts because, while we ultimately are seeking a caption style able to convey *both* valence and arousal, there would be a combinatorial explosion if we were to test all styles combinations depicting valence and arousal. Thus, answering RQ1.A can help us narrow down viable styles for each aspect before combining them to address RQ1.B. In order to enrich our understanding of DHH individuals' preferences and the reasons for their choices, we also ask:

RQ2. What factors influence DHH viewers' preference for specific caption text styles conveying valence and arousal in speech?

Once we have established a roster of caption styles with high favorability ratings among DHH viewers, we will further refine the selection by putting it through a series of evaluations that, in answering the research questions below, will allow us to prepare design recommendations for researchers and designers interested in employing affective captions.

RQ3. Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ1.B, outperform a baseline caption text when DHH participants…
  A *… engage in an emotion-recognition task when watching captioned videos?*
  B *… report on their subjective impressions of how each caption style performs according to the factors outlined in the answering of RQ2?*

3 STUDY 1: EVALUATING CAPTIONS STYLES THAT DEPICT VALENCE OR AROUSAL *INDIVIDUALLY*

In Study 1, we aimed to understand the preferences of DHH participants for caption styles that depict either valence or arousal, but not the two combined. Participants viewed examples of affective captions presented in various caption styles (see Figure 3). They were tasked with comparing them, choosing the styles that they saw as having a clearer representation of the depicted emotion. These choices were compiled into a ranking of preferences across all participants (RQ1.A). Since we were also interested in uncovering the reasons behind these choices, questions about subjective impressions about the selected styles were included as well (RQ2).

3.1 Methods

3.1.1 *Stimuli Generation.*

Text and Audio Processing. For our evaluation, we needed a set of videos to which we could add affective captions depicting valence and arousal using different visual styles. We used videos from the Stanford Emotional Narratives Dataset (SEND) [52]. These are short videos featuring individuals retelling personal stories with a strong emotional component. Examples include one person's reflection on their mother's battle with tuberculosis, another person's experience of a breakup on a school trip, a third person's unexpected victory at a school race, among others.

Included with the dataset are each video's transcriptions, which we fed, along with their audio channel, through an instance of Gentle [50], a Kaldi-based force-alignment toolkit set at a word-based granularity level. This gave us a timestamp for when each word in the transcript starts and ends. With these timestamps, we isolated the audio excerpts for each word, which were processed through a transformer-based neural network [70] to obtain values of valence and arousal, emotional components as defined by the circumplex model of emotion [60].[1] These two values, which are the final output of this process, as illustrated in Figure 2, were then normalized [22] and annotated in a caption file [21].

Typographic Styles. To generate the videos, we processed their annotated caption files. We mapped the valence and arousal values assigned to each word to specific typographical parameters. In essence, this means that as emotional values increase, so do the associated typographic parameters. Since there is a lack of systematic exploration regarding the mapping between emotions and typography, we gathered a diverse range of typographic modulations from the literature — although in their original use-cases many were not specifically used to represent either valence or arousal, or even applied to captions, this approach allowed us to cover a wide spectrum of possibilities during our evaluation. To ensure the accessibility of text output in each style, we adhered to the WCAG guidelines [18] to the best of our ability.

The *font-color*, *background-color*, and *shadow-color* styles are based on using changes in hue to represent the chosen affective dimension. Color has been used to represent emotions and moods by researchers in different fields [6, 15, 16, 24, 28, 32, 42, 43, 63]. Frequently, this is used to depict valence, with red commonly associated with negative values [28, 42] and blue [63] or green [42] used for positive ones. We used the color palette defined in Hassan et al. [32], designed to ensure that individuals with color vision deficiency are able to distinguish the different values.

Although all three styles use color, they differ in their application. The *font-color* (Figure 3a) style involves changing the color of the word itself [24]. *Background-color* (Figure 3b) involves adding a colored box behind each word. This is similar to visual experiments done for instant messaging interfaces [15, 16]. Finally, the *shadow-color* (Figure 3c) style applies a blurred halo behind each word. While we have not found examples of this use in the literature, it is based on exaggerating the common drop shadow effect used in conventional captions to increase their figure-ground contrast.

The *font-weight*, *baseline-shift*, and *letter-spacing* typographic parameters (shown in Figures 3d, 3e, and 3f, respectively) have been used by various authors to represent elements of prosody, such as loudness, pitch, and duration [9, 14, 22, 23, 59, 61, 79], or arousal and intensity of valence [32].

Font-size (Figure 3g) has been used to depict arousal [24]. Hassan et al. [32] used brightness to depict dominance, which we here adapted as *opacity* (Figure 3h). Lastly, Promphan [54]'s *emotional-typeface* (Figure 3i) has letter shapes that translate negative, neutral, and positive valence as jagged, balanced, or rounded strokes.

With videos rendered at 960 pixels wide, valence and arousal values were shown as follows: *Background-color* & *font-color*: 0.0 as ■ `#ff8979`, 0.5 as ■ white; 1.0 as ■ `#00ffff`; *Baseline-shift*: 0.0 as -20 PX, 1.0 as 20 PX; *Font-size*: 0.0 as 18 PX, 1.0 as 34 PX; *Font-weight*: 0.0 as 200, 0.5 as 400, 1.0 as 900; *Letter spacing*: 0.0 as -0.2 CH, 1.0 as 0.6 CH; *Opacity* 0.0 as 30 %, 1.0 as 100 %; and *Shadow-color*: 0.0 as ■ `#ff8979`, 0.5 as ■ black; 1.0 as ■ `#00ffff`. Intermediary values were interpolated linearly. PX and CH units are presented as were defined in our CSS stylesheets. For the *Emotional typeface*, the five discrete font shapes applied evenly between 0.0 and 1.0. For all other styles, the Inter typeface was used [5].

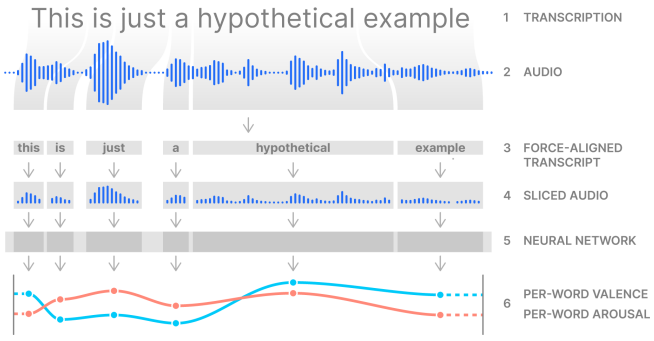
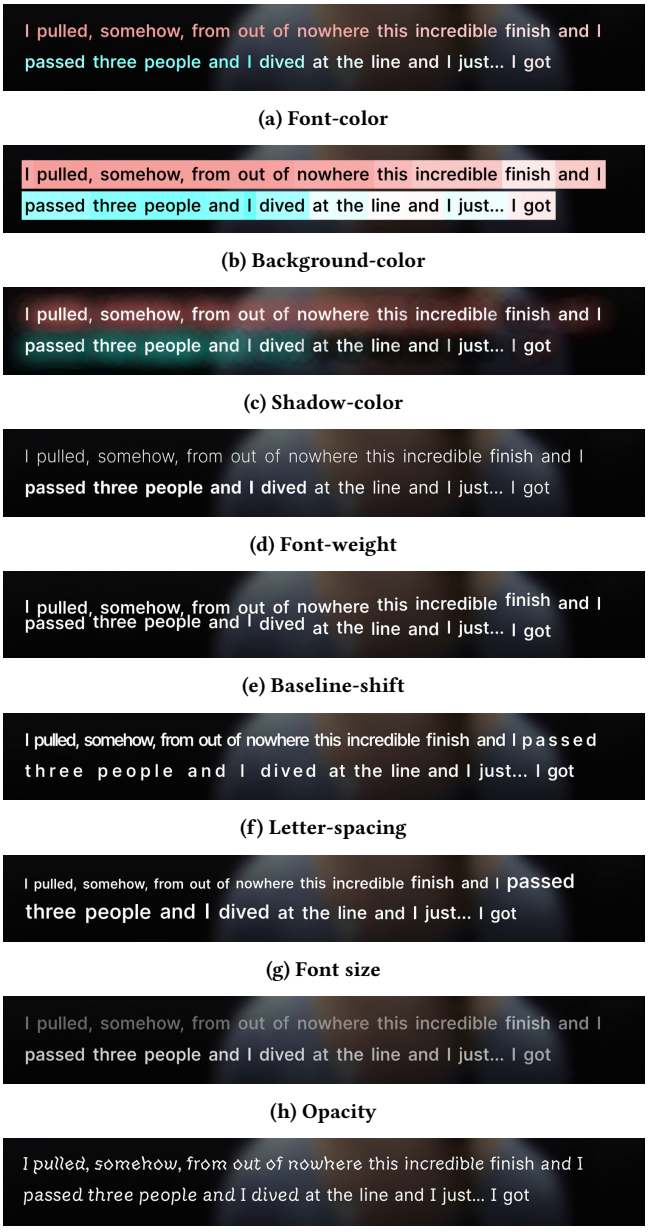


Figure 2: Diagram of how the transcription of a spoken utterance (1), together with its audio file (2), are used to generate a force-aligned transcript (3), which allows for the slicing of each word's audio (4), which is then fed into a neural network (5) that outputs its valence and arousal levels (6).

[1] We acknowledge the scholarly debate challenging the possibility of there actually being a meaningful ground *truth* for such a system to deduce [11, 35], but feel that this discussion is best left to a more focused inquiry on the subject and, as such, will sidestep it as a tangential matter considering our goals with this paper.



(a) Font-color

(b) Background-color

(c) Shadow-color

(d) Font-weight

(e) Baseline-shift

(f) Letter-spacing

(g) Font size

(h) Opacity

(i) Emotional-typeface

Figure 3: The nine caption styles used in the first evaluation. All images are screenshots from one of the videos that were used as stimuli.

Video Selection. The selection of videos within the SEND dataset had to meet several criteria. First, each video needed to cover a range of valence and arousal levels, including low, medium, and high values, so as to show a representative example of each caption style. Second, the videos had to be brief, as participants would need to evaluate nine caption styles for each input dimension in a session no longer than 75 minutes, as per our approved research protocol. Lastly, we aimed to represent the diversity of stories and participants in the SEND dataset by selecting videos that included a variety of ethnicities, genders, and positively and negatively toned stories. This diversity helps to account for how different caption styles might be more or less favored depending on the subject or theme of the video.

We selected short extracts from multiple videos to maximize diversity and minimize video length, while also ensuring significant variation in valence and arousal levels. This involved a careful analysis of the videos' valence and arousal levels to identify brief moments with notable variations.

3.1.2 Evaluation Design and Analysis Plan. Answering RQ1.A[2] required a method capable of assessing participants' preferences for each of the nine chosen caption styles, whether applied to depicting valence or arousal. To do so, we opted for Best-worst scaling (BWS).

In this method, participants are presented with a set of options and asked to choose the *best* and *worst* based on specific criteria. In our case, the options were the caption styles, and the criteria was whether each style did a good or bad job at depicting either valence or arousal. This process is shown in Figure 4. We can leverage the best-worst choices to obtain *explicit* and *implicit* ranking data for each style. For instance, if a participant selects A as the best and D as the worst from the caption styles A, B, C, and D, they explicitly indicate $A > D$, but implicitly suggest that $A > B$, $A > C$, $B > D$, and $C > D$. The only missing ranking information in this example pertains to the non-selected options B and C. With repeated rounds, especially if there are many options, we can obtain good coverage without overloading participants at any particular round.

For our setup, BWS offers advantages over Likert-rating scales, pairwise comparisons, or integer rankings. For one, it suits scenarios where participants might overlook small differences[3] between items [8]. By having participants compare only a small subset of options at a time, BWS strikes a balance between the simplicity of pairwise comparisons (easier, but requiring too many rounds) and the efficiency of integer rankings (harder, but with shorter tests).

This is coupled with evidence suggesting that BWS performs well for experiments that measure caption-appearance preference among DHH users [8] or, more broadly, that measure typographic style preferences [72]. Lastly, BWS is more robust than Likert-rating scales against inconsistencies[4] in participants' ratings across multiple rounds [20, 40].

To analyze the data, we used an ELO-rating system that incorporates all obtained pairings, whether explicit or implied, modelling participants' preferences as a set of likelihoods of choosing one option over another. The system operates under the assumption that each style has an underlying *strength* S, such that if $S_A > S_B$, style A is expected to be chosen more frequently than style B, with the difference in strengths directly influencing the frequency of this choice.

[2] RQ1.A: Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal *when depicted individually?*

[3] We prioritized legibility when designing the caption styles. This at times constrained visual expression, leading to subtle differences between some styles.

[4] Repeated showings of the same caption style across multiple videos can lead to inconsistencies in participants' ratings, an effect stemming from participants not knowing the full range of choices until they have been through many rounds of the test.

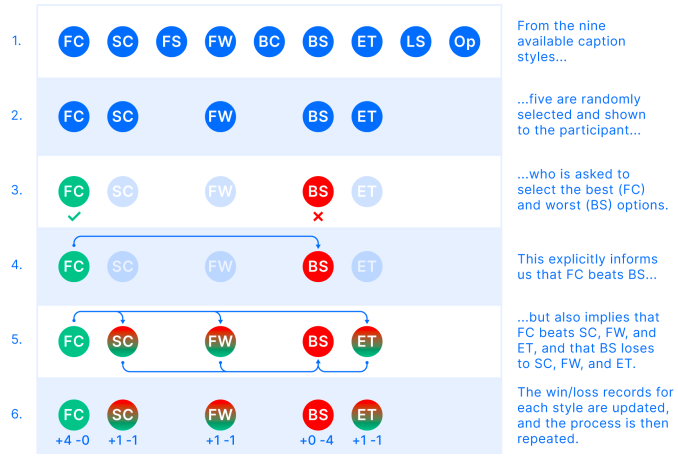


Figure 4: Example of one round in our Best-worst scaling setup. Caption style names are abbreviated as in Figure 1.

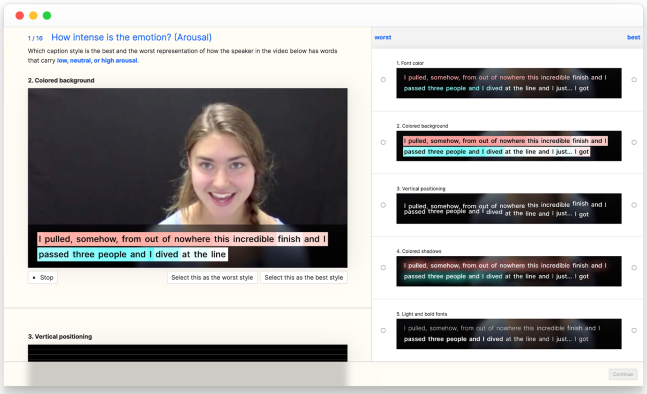


Figure 5: Screenshot of the experiment's platform. On the left side of the image, an example video is shown with the background-color caption style. On the right side, five different caption styles are displayed, which were presented to participants in a particular Best-worst scaling (BWS) round. The image illustrates instructions for the arousal segment of the test. For the valence portion, the text read 'What type of emotion? (Valence) Which caption style is the best and the worst representation of how the speaker in the video below has words that carry negative, neutral, or positive valence.'

Every match updates the ratings of the two styles, but notably, ELO algorithms are self-correcting, i.e., a choice that confirms the expectations of the model will not cause large changes in ratings, whereas an upset victory of a 'weaker' style would [26]. This feature allows us to define a preference ranking that captures potential dissensus among participants. Put differently, the ranking does not rigidly impose a 'winner-takes-all' structure; rather, it adapts to accommodate diverse participant responses.

The actual implementation used was Herbrich et al. [33]'s TrueSkill method. This is an ELO-like algorithm that models each style's strength as a normal distribution with mean value μ and standard deviation σ. This quantifies both a style's expected performance and the level of uncertainty around this estimate, which can help us moderate the degree to which we trust the results. Larger or smaller values of σ reflect more or less uncertainty as to the overall rankings, which is expected to diminish as more matches are run. To determine a given caption style's true relative strength, we use $\mu \pm 2\sigma$, providing a 95% confidence interval.[5]

3.1.3 Experimental Set-up. Both Study 1 and 2 shared a similar overall structure. In Study 1, described in this section, participants were randomly assigned to start with videos showing either valence or arousal. Over the course of the session, they completed a total of 16 rounds, with eight rounds dedicated to valence and eight to arousal. Every round consisted of five videos with the same scene, each one showing a different caption style selected from a pool of nine options. The videos had an average duration of 18 seconds (σ = 4s). Figure 5 shows a screenshot of the interface used during the first study. As with the following two studies, this experiment was developed as a website using jsPsych [25].

In Study 1, after the BWS/videos portion of the test, participants were asked to answer two questions about their most favored and disfavored style for both valence and arousal (so eight questions in total). These questions were open-ended, and phrased as *Why did you think this caption style worked [or 'did not work'] as a representation of each word's valence, or emotional tone [or 'arousal level']?*, and *Do you have any suggestions about what could be made to make this particular style work better?*

3.2 Findings from Study 1

Participants were recruited by sending out Institutional Review Board-approved ads to social network groups and university-related student groups. Participants qualified to participate in this experiment if they identified as d/Deaf or Hard-of-Hearing. For Study 1 we recruited a total of 10 participants, 7 of which identified as female and 3 as male, 7 of which identified as d/Deaf and 3 as Hard-of-Hearing, with a mean age of 29.5 years (σ = 11.9).

3.2.1 Caption Style Reference Rankings. Study 1 had 10 participants evaluating 5 videos per round for 8 rounds for each affective dimension. A 5-way BWS generates 7 data pairs, so $10 \times 8 \times 7 = 560$ pairwise comparisons for both valence and arousal. Table 1 shows the results from the study, including both the raw answers — i.e., what participants explicitly chose (or 'N/A', for the times a style was shown but neither won nor lost) —, and the choices implied by the BWS setup.

Note that the numbers presented here show the frequency with which each caption style was favored over the other styles it was compared against. These numbers alone do not reflect the final ranking of the strengths of each style, as understanding the relative strength of each comparison is crucial — beating a weaker opponent will result in fewer ranking points being earned than if you were

[5] A caveat of ELO systems is their dependence on match order, i.e., the outcomes of subsequent matches are affected by previous ones. While this is logical for competitive games, in our experiment the sequence of matches lacks any inherent order, so we follow Clark et al. [17]'s suggestion of averaging ELO-outputs from randomly ordered iterations of the data until consistent outcomes emerge. In our case, stability was achieved after 1,000 iterations.

Style	RAW ANSWERS						IMPLIED ANSWERS			
	VALENCE			AROUSAL			VALENCE		AROUSAL	
	WON	LOST	N/A	WON	LOST	N/A	WINS	LOSSES	WINS	LOSSES
Background-color	40%	20%	40%	31%	25%	44%	62%	38%	54%	46%
Baseline shift	15%	85%	0%	31%	69%	0%	25%	75%	40%	60%
Emotional typeface	12%	88%	0%	16%	84%	0%	24%	76%	27%	73%
Font-color	48%	0%	52%	33%	14%	53%	83%	17%	63%	37%
Font-size	17%	10%	73%	25%	4%	71%	56%	44%	67%	33%
Font-weight	14%	2%	84%	20%	7%	72%	60%	40%	60%	40%
Letter spacing	9%	37%	53%	9%	23%	68%	31%	69%	39%	61%
Opacity	6%	21%	73%	2%	27%	71%	39%	61%	31%	69%
Shadow-color	24%	3%	74%	44%	8%	48%	67%	33%	74%	26%

Table 1: Raw and implied (as per the BWS method) results for each one of the 9 styles, applied either for depicting valence or arousal. In the raw results columns, choosing a style as the best option counts as a win, and choosing it as the worst option counts as a loss. 'N/A' columns indicate the percentage of times a given style was shown in a round but was not chosen as the best or worst option.

to beat a stronger opponent. Nevertheless, they serve as a useful initial reference point for further analysis of the data.

To assess the internal consistency of participant responses, a Split-Half Reliability test was conducted by calculating the Spearman rank correlation coefficient between two randomly divided segments of the complete dataset [41]. The data, transformed into scores using the counting procedure outlined by Orme [53], revealed high correlations for both the valence ($\rho = 0.92$, $p < 0.001$) and arousal ($\rho = 0.90$, $p < 0.01$) datasets.

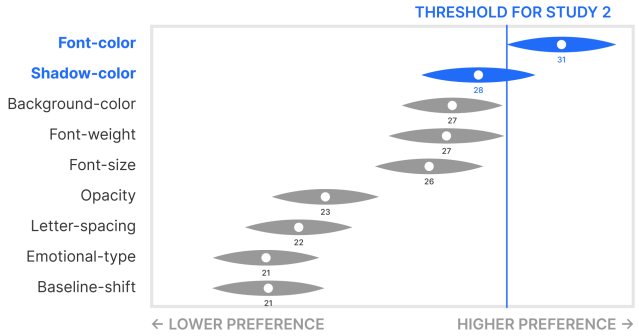
We ran the pairings through the TrueSkill algorithm, obtaining the relative strengths of each style. We used the Python library with default initial parameters.[6] Of note, μ (the strength of each caption style) started at 25, so styles that ended above or below this gained or lost points after all the matches were processed. The final values obtained were:

- Valence: **font-color** ($\mu = 30.6, \sigma = 0.9$), **shadow-color** ($\mu = 27.9, \sigma = 1.0$), background-color ($\mu = 27.0, \sigma = 0.9$), font-weight ($\mu = 26.9, \sigma = 1.0$), font-size ($\mu = 26.2, \sigma = 0.9$), opacity ($\mu = 22.9, \sigma = 0.9$), letter-spacing ($\mu = 22.0, \sigma = 0.9$), emotional-type ($\mu = 20.9, \sigma = 0.9$), and baseline-shift ($\mu = 21.0, \sigma = 1.0$).
- Arousal: **shadow-color** ($\mu = 29.0, \sigma = 0.9$), **font-size** ($\mu = 27.7, \sigma = 0.9$), **font-color** ($\mu = 26.8, \sigma = 0.9$), **font-weight** ($\mu = 26.6, \sigma = 0.9$), background-color ($\mu = 25.6, \sigma = 0.8$), baseline-shift ($\mu = 23.7, \sigma = 0.9$), letter-spacing ($\mu = 23.0, \sigma = 0.9$), emotional-type ($\mu = 21.4, \sigma = 0.9$), and opacity ($\mu = 21.5, \sigma = 0.9$).

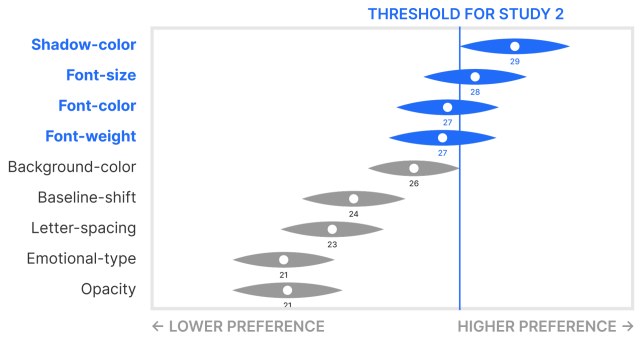
Styles in bold were included for Study 2 if their higher bound was greater than the lower bound of the top-scoring style. TrueSkill results are also shown in figure 6.

3.2.2 Open-Ended Feedback from Participants. After the BWS part of the test, participants were shown, for each of valence and arousal, the caption styles that best and worst performed according to their

[6] See Lee [44] for documentation on installing and using the library. Parameters were set at their default values of $\mu = 25$, $\sigma = \mu/3$, $\beta = \sigma/2$, and $\tau = \sigma/100$.



(a) Preferences for valence-representing caption styles



(b) Preferences for arousal-representing caption styles

Figure 6: Charts showing the relative strength and confidence interval for each caption style in relation to valence and arousal, using data collected in Study 1. The numbers shown are the TrueSkill output of each caption style after all matches were run. The initial value was set at 25, so values above that indicate caption styles that ended up gaining skill. The blue vertical line highlights the lower bound of the top choice for each input dimension, which was used as a cut-off point to select styles for inclusion in Study 2.

individual votes. They were then asked to comment on the reasons they thought these styles were (or were not) able to convey valence or arousal. (Quotes edited for brevity and clarity).

Participants commented on the unsuitability of certain styles to depict the two affective dimensions. Commenting on the emotional-typeface style (Fig. 3i), P1 said it didn't work for either arousal (*'I don't see how jagged letters will help me associate negativity'*) or valence (*'it's confusing to recall whether the jags were a positive or negative association'*). P4, on its use for valence: *'when they were jagged it was hard to read — it just wasn't helpful'*. P5, for arousal: *'sometimes the neutral looks similar to positive'*.

P9 noted on letter-spacing: *'[I] didn't understand how space is associated with positive arousal level, and it's a bit hard to read.'* P10 echoed the sentiment but regarding valence: *'for negative words the letters are too close to each other, making it harder to read.'* For P5 Baseline-shift (Fig. 3e) didn't work, since there was no visual reference against which changes could be seen. *'If someone came up being positive all the way until the end, it looks like almost nothing changed,'* an effect compounded by how line breaks also create changes in vertical spacing: *'also, if there are two lines, it's hard to tell if there was a change in the speaker's tone or just a new line.'*

Some styles posed readability issues due to limitations in design and mapping, even though participants deemed them suitable for representing an affective dimension. P2, on opacity (Fig. 3h) for arousal: *'I think it did a good job showing one's emotions, but was harder to read when it was really light.'* Again for arousal, P5 said that *'it's not that this caption style didn't necessarily work, but its transparency makes it a bit hard to read.'* P3 suggested font-weight be more spacious, but thought that it *'helped to see the person's tone by identifying [highlighted] words.'*

P2, on font-size (Fig. 3g) for valence, thought that *'at some points it was too small to read,'* a concern also echoed by P7, although they thought it *'worked well because the font size aligns with the level of the arousal.'* P9 agreed, stating that *'because we are taught that capital letters are associated with speaking loudly, I can easily associate a larger font with a speaker's positive arousal level and vice-versa.'*

Comments about font-color and background-color (Fig. 3a) were predominantly favorable, with participants commenting positively on the readability and interpretability of these styles. P2 commented on font-color: *'It was easy to read and super clear whereas the others were harder. It did a good job of expressing one's emotions using different shades.'*

Some participants shared their interpretation of the colors used. P3: *'Maybe green represents positive, and red represent angry?'* P4: *'It was helpful because in general, red is known to be more negative, while blue is more positive.'* P10 summarized the sentiment with *'colors help recognize tones.'* P1 preferred the background-color (Fig. 3b) style for arousal, saying *'it was the most visible option because I could see the colors and that helped me see the differences in arousal levels.'* P5 said that the use of colors has a learning slope, since *'blue is like the sky, which is good, and red is like anger, so it's bad, but when you are sad it's also blue and when you are happy it can be red, so it's a bit confusing.'* Still, they thought the style worked *'because its colors are way more obvious than the other styles, where changes in tones where harder to recognize.'*

Participants felt shadow-color (Fig. 3c) had legibility challenges despite being easier to interpret. P4: *'captions were difficult to read with the shadows behind them.'* P2 suggested *'making the shadows a little smaller around the words because it could get to be a little much at some points, with the shadows sometimes running over other words.'* Still, they said that *'once again, this is a color one so I really liked these. It was really clear and obvious to see how one was feeling.'*

4 STUDY 2: EVALUATING CAPTION STYLES THAT DEPICT VALENCE AND AROUSAL *IN COMBINATION*

After completing Study 1, we used the results to identify a small number of combinations between the most preferred caption styles for valence and arousal. These combinations included the top choice for each of valence and arousal, as well as any styles that overlapped with the winning styles' 95% confidence range. In total, we identified eight styles: font-color and shadow-color for valence, and font-weight, color, size, and shadow-color for arousal, which we combined in pairs of two. However, due to the way these styles were defined, font-color could not be combined with itself, and neither could shadow-color. This left us with a total of six styles to evaluate in Study 2, as seen in Figure 7.

The BWS portion of the test was similar to that of Study 1. Excerpts were slightly longer ($\mu = 27$s, with $\sigma = 5$s), and accounting for how the comparisons themselves were more complex — two styles per video, with subtle differences between some combinations — we reduced the number of videos per round from five to four, and the total number of rounds from 16 to 12.

After completing the BWS/videos portion of Study 2, participants were asked to provide open-ended feedback on their most and least preferred caption styles. They were then presented with the winning and losing caption styles according to their choices in the BWS portion of the test, i.e., each participant would see a different *best* and *worst* option.

Using an *inductive open coding* method [77], this data, along with notes taken during the study, and the open-ended data collected in Study 1, was separately analyzed and clustered by two authors to answer RQ2 regarding the factors that influence DHH viewers' preference for specific caption text styles.

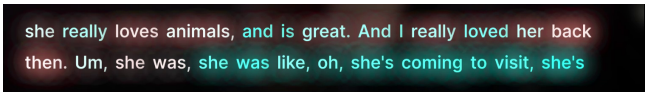
4.1 Findings from Study 2

Participants were recruited by sending out IRB-approved advertisements to social network groups and university-related student groups. Participants were identified as qualified to participate in this experiment if they identified as d/Deaf or Hard-of-Hearing. For Study 2 we recruited a total of 11 participants, 8 of which identified as female and 3 as male, 5 of who identified as d/Deaf and 6 as Hard-of-Hearing, with a mean age of 28.5 years ($\sigma = 9.7$).

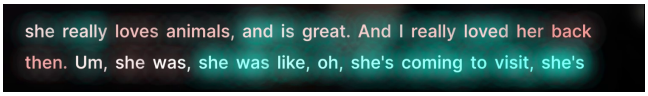
4.1.1 Caption Style Reference Rankings. Study 2 involved 11 participants evaluating 4 videos per round for 12 rounds. With 4 videos, the number of implied pairings generated at each round was 5, so we had $11 \times 12 \times 5$, resulting in 660 pairwise comparisons for the 6 combined styles. Table 2 shows both the raw and implied answers from participants. The Split-Half Reliability test for this dataset showed a strong correlation, with $\rho = 0.83$ and $p < 0.051$.

Style (valence with arousal)	RAW ANSWERS			IMPLIED ANSWERS	
	WON	LOST	N/A	WINS	LOSSES
Font-color with font-weight	42%	23%	35%	61%	39%
Font-color with font-size	31%	18%	51%	58%	42%
Font-color with shadow-color	29%	22%	49%	55%	45%
Shadow-color with font-color	24%	24%	52%	51%	49%
Shadow-color with font-weight	9%	28%	63%	38%	62%
Shadow-color with font-size	12%	35%	53%	35%	65%

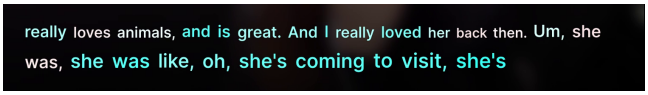
Table 2: Raw and implied (as per the BWS method) results for each one of the 6 font style combinations. In the raw results columns, choosing a style as the best option counts as a win, and choosing it as the worst option counts as a loss. 'N/A' columns indicate the percentage of times a given style was shown in a round but was not chosen as the best or worst option.



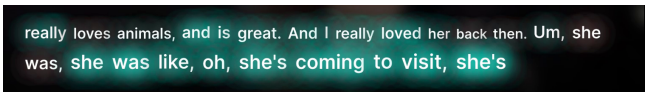
(a) Font-color for valence, shadow-color for arousal



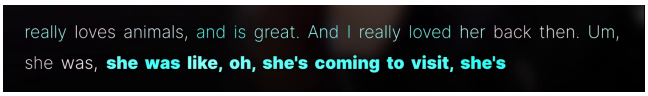
(b) Shadow-color for valence, font-color for arousal



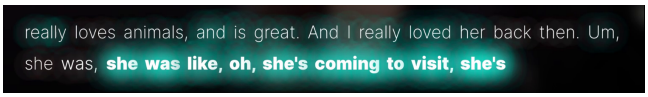
(c) Font-color for valence, font-size for arousal



(d) Shadow-color for valence, font-size for arousal



(e) Font-color for valence, font-weight for arousal



(f) Shadow-color for valence, font-weight for arousal

Figure 7: The six caption styles used in Study 2 of the evaluation. All images are screenshots from one of the videos used as stimuli.

Running the pairings through the TrueSkill, again with the same parameters as used in Study 1, gave us the following values: **Font-color with font-weight** (μ = 26.5, σ = 0.8), **font-color with font-size** (μ = 26.2, σ = 0.8), **font-color with shadow-color** (μ = 25.8, σ = 0.8), **shadow-color with font-color** (μ = 25.2, σ = 0.8),

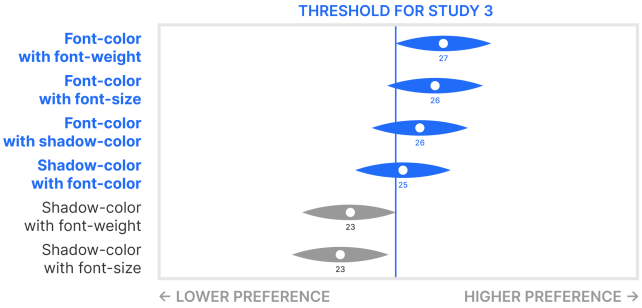


Figure 8: Relative strength and confidence interval for each caption style tested in Study 2. Style names have the first style depicting valence, and the second arousal. Styles in bold blue have their TrueSkill μ value overlapping the top choice when considering the 95% confidence interval. As with Study 1, the initial μ value was set at 25, meaning that styles that finished the matches above that value gained skill points.

shadow-color with font-weight (μ = 23.4, σ = 0.8), and shadow-color with font-size (μ = 23.1, σ = 0.8).

4.1.2 Open-Ended Feedback from Participants. Nine participants commented favorably on using font-size or font-weight to depict arousal. P18, for instance, noted that for depicting arousal, font-size was the most appropriate way but in a real-world application font-weight would be better. They believed that font-size might be a CLEARER and more INTUITIVE choice for conveying arousal, but font-weight could be a less disruptive alternative. They commented: *'[font-weight] represents arousal well while keeping it minimalist. Bold fonts naturally convey intensity — when we want to emphasize something, we use bold fonts.'* P17 echoed this sentiment: *'I liked how with font-color and weight there wasn't much factor or adaptation to the new changes for captioning. It is a subtle yet good change.'* Font-size, on the other hand, *'won't work as it can cause our eyes to «juggle» throughout the captioning, making it an effort to read.'*

Six participants in total, including P21 liked font-color paired with font-weight: *'Both of these make the general point, but are not [overwhelming]. The other ones were either too hard to read (font-size) or just too much for me (shadow-color).'* Like P18, they thought that just being able to clearly convey valence or arousal might not be

enough: *'I think shadow-color does a really good job in getting the point across but is just too distracting.'* This is a subtle point, though, since they also made an argument about how different parameters could make those styles work: *'I can read the small font sizes, but it could potentially be harder to read. So maybe making the shadows less showy and changing up the sizes a bit could help.'*

Three participants expressed their preference for different tones and saturation ranges for representing affective dimensions. P12: *'I think if the red color was darker or noticeable in some way, or having the neutral statements in a certain color rather than a muted red or blue, it would stand out more.'* P18 thought colors work well for valence, *'as long as the person using the caption gets used to the representation of each color. Maybe it would work better to put colors that refer more to negative/positive things?'* P21 added that *'maybe making some of the shades a little darker would help?'*

Concerns with LEGIBILITY and DISTRACTION were also common, particularly when shadow-color was used. P11: *'shadow-color with font-color is very distracting and hard to read. My eyes get strained while trying to pay attention, and I do not like how there is too much overlapping of shadow with the letters.'* P17 agreed, saying that *'the glow [on the shadow-color] is a nice idea, but too much can be too bearing for us to read,'* although they did think a *'little glow could help [the font-color with font-weight style].'* P13 thought there might be challenges *'for people with poor vision, or Deaf-Blind, seniors, and such — reading becomes really challenging if the glow is over-used, as it was for me a few times.'*

P15 echoed this: *'The glow in the caption kind of confuses me. Also, when the emotion is low and the font decreases in size it makes it hard for me to read.'* They still complimented font-size, though, because *'it is clear when showing the emotions of the speaker.'* P16 agreed, saying that *'font-size represents arousal the best, giving an insight about the level of an emotion — high, low, excited, etc.'*

5 STUDY 3: SUBJECTIVE AND OBJECTIVE PERFORMANCE OF THE COMBINED STYLES AGAINST A NEUTRAL BASELINE

At this point, Studies 1 and 2 had given us two important insights. First, they allowed us to narrow an initial set of 72 possible caption style combinations[7] into only 4. Second, they provided us with a set of criteria that DHH participants judged to be relevant when selecting an affective caption style. We will go over them in detail in the discussion session but, briefly, they were as follows: EASE OF READING, LOW DISTRACTION, INTUITIVENESS, and CLARITY OF EMOTIONAL REPRESENTATION. With this in hand, we designed and ran Study 3, aimed at helping us answer RQ3.A and RQ3.B.[8]

5.1 Methods

5.1.1 Stimuli and Experimental Design. We once again utilized the SEND resource to gather 10 videos of speakers recounting stories, both positive and negative. Each video was prepared in one of five caption styles, i.e., a total of 50 videos were rendered in the following conditions: (1) baseline, featuring non-stylized captions; (2) font-color with font-weight (Figure 7e); (3) font-color with font-size (Figure 7c); (4) font-color with shadow-color (Figure 7a); and (5) shadow-color with font-color (Figure 7b).

One of the factors outlined in answering RQ2 was that affective captions should have a CLEAR EMOTIONAL REPRESENTATION, done so INTUITIVELY. *Intuitive* is a fraught term in HCI [76], but here we use it to mean an artifact that, because it matches its users' expectations, is *easy to learn.* Thus, while intuitiveness might be too abstract a notion to objectively measure, we can quantify how quickly participants learn how a caption style works as an indirect proxy for it.

To do so, we divided the test into two blocks, with each of the five conditions being presented once per block, but twice overall. This study design allowed us to compare task performance for each style across the two blocks and, in finding meaningful differences, deduce the presence of a learning effect, i.e., participants were getting better (or worse) in decoding the caption styles. To maximize this effect, and in contrast to the previous two studies, we presented the videos in their entirety, giving participants more time to familiarize themselves with each caption style ($\mu = 141$ s, $\sigma = 36$ s, versus $\mu = 21$ s, $\sigma = 7$ s, for the combined set of videos used previously). As before, the test was implemented as a website with a mix of custom and off-the-shelf jsPsych plugins.

5.1.2 Effectiveness at Conveying Speakers' Emotions. To effectively measure how well each caption style was able to convey the speaker's emotions we used two approaches. The first was a subjective self-report instrument adapted from previous affective caption research [24, 39]. In it, participants signaled their level of agreement with the statement: *I could discern the speaker's emotions.* While we expect participants to also consider elements such as a speaker's facial expressions and body language [45], by comparing each novel caption style with the neutral, emotion-free baseline condition, we can infer that any observed differences were related to differences in caption styles.

As with other Likert-rating scales employed in this test, to compare answers we conducted statistical significance testing on responses using a Kruskal-Wallis test. If significant, we ran a post hoc Mann-Whitney U test between each caption type, with p-values adjusted using Holm-Šídák corrections.

The second approach, previously explored by Hassan et al. [32], involved having participants annotate single words from a captioned video. We expanded on this method by having participants annotate four distinct ten-word[9] groups per video. The selection of these four groups aimed to include examples of words with positive valence and arousal, negative valence and arousal, positive valence and negative arousal, and negative valence and positive arousal. In other words, an illustrative example was sought for each of the four quadrants in the circumplex plane. This entailed examples where valence and arousal values were either convergent or divergent, and either positively and/or negatively oriented.

[7] Permutation of 9 items into subsets of 2, $P(n,r) = \frac{n!}{(n-r)!} = \frac{9!}{7!} = 72$

[8] Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ1.B, outperform a baseline caption text when DHH participants (RQ3.A) *engage in an emotion-recognition task when watching captioned videos?*, and (RQ3.B) *report on their subjective impressions of how each caption style performs according to the factors outlined in the answering of RQ2?*

[9] The choice of ten words struck a balance between having too many words, which occasionally exceeded two lines of captioned text, and having too few, which might lack contextual understanding when isolated.

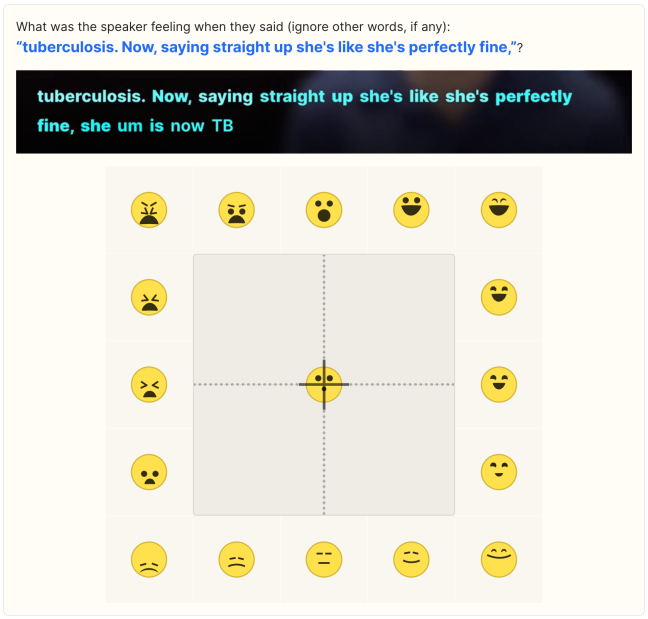


Figure 9: Our EmojiGrid implementation, with figures from Toet [66]. In this example, participants had to label an image that — as we know beforehand, but they had to figure out — has valence, depicted by font-color, and arousal, depicted by font-weight, both placed on the top-right quadrant.

To select portions of text from our video captions that would be positioned within each quadrant, valence and arousal values for groups of 10 words across each video were averaged, ordering them by how distant these average x (valence) and y (arousal) coordinates were to the extreme points in each quadrant, e.g., $(1, 1)$ for positive valence and arousal, $(-1, -1)$ for negative, and so forth. The top choice for each quadrant was then selected. To prioritize groups of words with greater homogeneity, in ordering the selection by distance to the extreme points we rounded values to one place after the decimal point and broke the ties by selecting the group with the lowest standard deviation.

To effectively measure participants' interpretations of the ten-word groups, we implemented an EmojiGrid [67]. This instrument asks participants to select a coordinate within a Cartesian plane, mapping valence along the horizontal axis and arousal along the vertical axis. This mirrors the representation of emotions in the circumplex emotional framework, with valence on the x-axis, and valence on the y-axis. Rows and columns of emojis were positioned along the edges of the plane, hinting at corresponding emotions at each position. Originally designed for labeling affective responses to images of food, it has since been widely employed to annotate diverse stimuli, including self-experiences, videos, and so on, e.g., [68, 69]. An advantage of the EmojiGrid as a measurement tool is that its use of graphic elements reduces the risk of differences in written literacy affecting the interpretation of the instrument.

With it in place, we expected participants to generate four coordinate pairs for each video, totaling eight for each caption style and 40 in total. To analyze how effective each caption style is in translating affective information, we would need to measure how distant each of these participant-provided coordinates was from a 'ground truth' provided by the neural network that analyzed each video's audio. In measuring this correlation, however, a typical approach such as finding Pearson's correlation coefficient might fall short of our needs, given that it is limited to considering two variables at a time [74], i.e., correlating ground-truth valence versus participants' valence while ignoring the corresponding pair of arousal values.

As an alternative, Székely et al.'s *distance correlation* measures the degree of dependence between two random vectors by evaluating the similarity of pairwise distances within each vector [65]. It captures both linear and non-linear correlations, although it is worth mentioning that it does not indicate the direction of the correlation, i.e., it quantifies the degree of dependence on a scale from 0 (independence) to 1 (high degree of dependence).

With distance correlation, we quantify how strongly each caption style is informing participants' perceptions of the depicted affective signal, i.e., its CLARITY OF EMOTIONAL REPRESENTATION. Moreover, by independently applying the method to the first and second times each caption style was shown we can capture differences in performance that are related to each style's ease-of-learning which, as exposed above, we will use as a proxy for their INTUITIVENESS.

5.1.3 Ease of Reading and Processing. Another set of criteria that came out as important for affective caption styles is that they should be EASY TO READ and NOT DISTRACTING. To measure this, we adapted three Likert-rating scale items from Kim et al. [39], to gauge legibility and cognitive load. These items were themselves constructed based on the NASA-TLX framework and prior research focused on caption accessibility for DHH individuals [31, 38]. In our study, participants were asked to rate their level of agreement with the following statements: *I felt hurried / rushed while I was watching the video, I found watching the captions and video *simultaneously* mentally demanding*, and *I found these captions easy to read.*

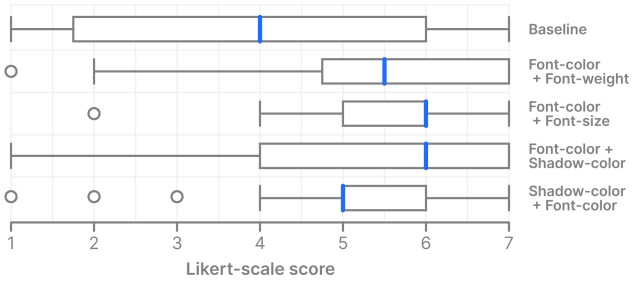
5.2 Findings from Study 3

Participants were recruited through IRB-approved posts made to social media and university-related student groups. Participants qualified to participate if they identified as d/Deaf or Hard-of-Hearing. For Study 3 we recruited a total of 18 participants, 10 of which identified as female and 8 as male, 11 of who identified as d/Deaf and 7 as Hard-of-Hearing, with a mean age of 27 years ($\sigma = 8.1$). In this section, we again follow the convention: the first typographic style represents valence, and the second represents arousal in naming a combined caption style.

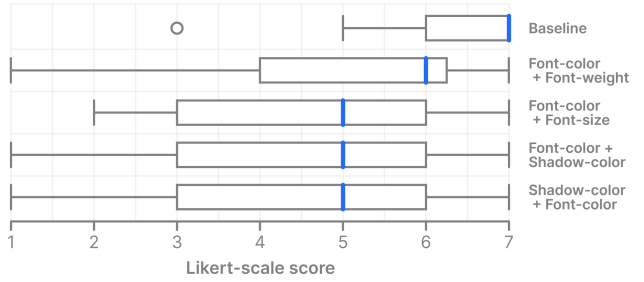
5.2.1 Effectiveness at Conveying the Speaker's Emotions. Median responses for agreement with the *I was able to understand the speaker's emotions* statement, shown in Figure 10a, were: 4 for the baseline condition; 5.5 for font-color with font-weight; 6 for font-color with font-size, with significant difference versus the baseline ($U = 385.0$, $p < 0.05$, *medium effect*)[10]; 6 for font-color with shadow-color; and 5 for shadow-color with font-color.

We calculated the distance correlation between ground-truth and participant-provided coordinates using Ramos-Carreño and

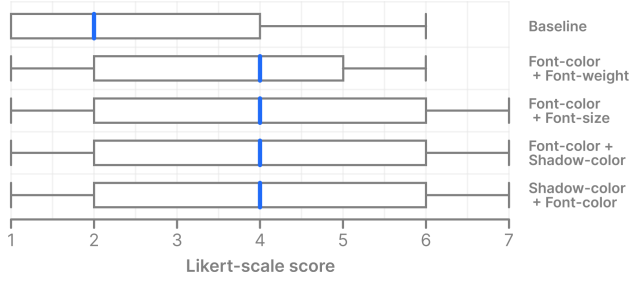
[10]P-values presented adjusted using Holm-Šídák corrections.



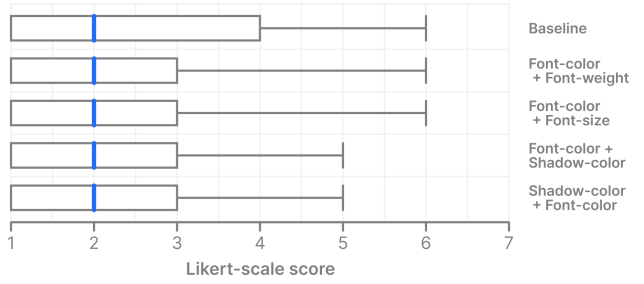
(a) Agreement with *'I was able to understand the speaker's emotions.'* Significant differences were found between the baseline and the font-color with font-size caption styles.



(b) Agreement with *'I found these captions easy to read.'* Significant differences were found between the baseline and the four novel caption styles.



(c) Agreement with *'I found watching the captions and video *simultaneously* mentally demanding.'* Significant differences were found between the baseline and the font-color with font-size styles, and the baseline and font-color with shadow-color caption styles.



(d) Agreement with *'I felt hurried / rushed while I was watching the video.'* No significant differences were found between the baseline and the four novel caption styles.

Figure 10: Box-whisker plot showing the spread of answers between the five conditions for different Likert scales.

Torrecilla's implementation [56, 57]. The resulting correlation coefficients are presented in Table 3. Notably, font-color with font-weight and font-color with font-size exhibited a significant distance correlation with participants' answers compared to the ground truth, while the other two styles and the baseline did not. Additionally, both top-performing styles demonstrated performance differences between rounds 1 and 2, suggesting a learning effect.

CONDITION	ROUND 1	ROUND 2	ROUND 1 & 2
Baseline	0.14	0.09	0.07
font-color + font-weight	0.14	**0.32*****	**0.21*****
font-color + font-size	0.10	**0.23****	**0.14***
font-color + shadow-color	0.10	0.10	0.06
shadow-color + font-color	0.10	0.10	0.07

Table 3: Distance correlations between participants' valence and arousal measures and the ground truth for each of the five conditions. The columns slice the data into three groups. Columns 2 and 3 show, respectively, the first and the second time participants saw each condition. Column 4 includes the whole data. Significant correlations are highlighted by * for $p < 0.05$, ** for $p < 0.01$, and * for $p < 0.001$. P-values were calculated using a 10,000-round permutation test.**

5.2.2 Ease of Reading and Processing. Median responses for agreement with the *I found these captions easy to read* statement, shown in Figure 10b, were: 7 for the baseline condition; 6 for font-color with font-weight, with significant difference versus the baseline ($U = 938.0$, $p < 0.01$, *large effect*); 5 for font-color with font-size, with significant difference versus the baseline ($U = 1003.0$, $p < 0.001$, *large effect*); 5 for font-color with shadow-color, with significant difference versus the baseline ($U = 1021.0$, $p < 0.001$, *large effect*); And, lastly, 5 for shadow-color with font-color, with significant difference versus the baseline ($U = 1039.0$, $p < 0.001$, *large effect*).

Median responses for agreement with the *I found watching the captions and videos *simultaneously* mentally demanding* statement, shown in Figure 10c, were: 2 for the baseline condition; 4 for font-color with font-weight; 4 for font-color with font-size, with significant difference versus the baseline ($U = 390.5$, $p < 0.05$, *medium effect*); 4 for font-color with shadow-color, with significant difference versus the baseline ($U = 391.0$, $p < 0.05$, *medium effect*); And, lastly, 4 for shadow-color with font-color.

No significant differences were found in participants' agreement to the *I felt hurried / rushed while I was watching the video* statement. All conditions reported a median value of 2, as shown in Figure 10d

6 DISCUSSION

While prior caption presentation research had investigated the benefits of affective captions and their effectiveness at conveying emotions [21, 23, 24, 32, 39], no prior study had compared caption styles that represent emotions in scrolling captions from the perspectives of DHH users. Indeed, a comprehensive empirical investigation to identify the preferred typographic modulations was a suggestion for future studies in some of these works and served to inspire this paper. Our research findings provide evidence regarding the preference of DHH participants for particular caption styles in representing valence and arousal in captions (RQ 1.A & RQ 1.B), the factors that influence these choices (RQ2), and how well the preferred styles performed under a set of meaningful objective (RQ3.A) and subjective (RQ3.B) quality criteria.

6.1 Caption-Style Preferences (RQ 1.A & RQ 1.B)

Findings from Study 1 and 2 revealed marked differences in participants' preferences for using different styles to depict valence and arousal. However, preferences for styles depicting valence were more cohesive than those for styles depicting arousal. Regarding RQ1.A, the outcome of Study 1 showed two styles emerging as the most preferred for depicting valence: font-color and shadow-color, with font-color holding a slight advantage. For arousal, there was a closer tie between four styles: shadow-color, font-size, font-color, and font-weight.

Combining Study 1's styles for Study 2 yielded six new styles. Answering RQ1.B, participants' choices showed a four-way tie. For depicting valence, the top three choices featured font-color, while the fourth option used shadow-color. In contrast, the top-four choices had arousal depicted as follows: font-weight, font-size, shadow-color, and font-color. This ranking substantiates design choices made by previous authors in the affective captioning space [24, 32], while also showing that many typographic parameters considered for prosodic captions were not as effective [14, 22, 23, 79].

6.2 Factors Influencing Participants' Choices of Caption Styles (RQ 2)

Participants' reasons for choosing the winning and losing caption styles are noteworthy in that they seem nearly identical, regardless of justifying a winning or losing choice. Echoing Hassan et al. [32], these factors included EASE OF READING, LOW DISTRACTION, CLEAR EMOTIONAL REPRESENTATION, and an INTUITIVE VISUAL DESIGN, i.e., participants' expectations of how a visual attribute should map to an emotion corresponds to how the style actually implements this modulation. These concepts appeared throughout the answers, but different participants applied them to different caption styles. For instance, more participants claimed that the font-weight style for arousal was easier to read than font-size. However, there was no consensus, and a few participants found the opposite to be true. Therefore, the answer to the question of what makes an affective caption style readable and intuitive is not straightforward, as it depends on the expectations of the user, which can vary from person to person. This implies that, in answering RQ2, one can cite these overall factors — readability, low distraction, intuitiveness, etc. — with the caveat that their applicability can depend on context and group of users.

6.3 Objective Measures of Performance (RQ 3.A)

Table 3 shows that styles using text shadows, whether when conveying valence or arousal, did not appear to influence how participants interpreted speakers' emotions. As such, we do not recommend those styles for affective captions.

Conversely, styles with font-color for valence and either font-weight or font-size for arousal had significant correlations. This suggests a degree of EMOTIONAL CLARITY. Furthermore, the observed increase in these correlations when comparing participants' initial exposure to each style with their second exposure hints at a learning effect. This effect can be attributed to an INTUITIVE utilization of typographic modifications to convey affective dimensions.

In sum, and in answering RQ3.A, these results show that, while there were no significant differences between participants' *preferences* for the four caption styles we tested, the performance in an emotion-recognition task is notably higher for the two styles that combined font-color for valence with either font-weight or font-size styles for arousal when compared to the two styles that combined font-color with shadow-color for each affective dimension.

6.4 Subjective Measures of Performance (RQ 3.B)

We also measured how much participants *felt* each caption style helped their understanding of the speaker's emotions (see Figure 10a). For this, only the style with font-color for valence and font-size for arousal had a *significant* difference versus the neutral baseline. This observation aligns with the outcomes of the emotion-recognition task and positions font-size ahead of font-weight for depicting arousal, given how it had high marks in both objective and subjective measures at helping DHH viewers understand a captioned speaker's emotions.

However, it is important to note that the winning style's CLEAR EMOTIONAL REPRESENTATION and INTUITIVENESS appear to come at the cost of higher distraction levels, as indicated by participants agreeing that watching captions with the style along the video was mentally demanding (see Figure 10c). This aspect stands as a notable drawback of the winning caption style, given how LOW DISTRACTION was also a factor guiding participants' choices of caption styles. P18's comment from Study 2 reinforces this, noting that font size changes, though effective for depicting arousal, also introduce disruption. In this sense, using font-weight to depict arousal shows an edge over font-size. In both cases, font-color performed well in its depiction of valence.

Also of note, all four affective caption styles scored lower than the baseline in LEGIBILITY (see Figure 10b).

6.5 Design Recommendations

The findings from our studies provide design guidance for researchers and designers of affective captioning applications and reveal some remaining open questions, which may be a basis for future research studies.

(1) Combining font-color for valence with either font-size or font-weight for arousal leads to compelling and effective caption styles for depicting affective dimensions of speech. Both styles were shown to be viable options for affective captioning applications and can be presented as choices for users of such systems.

(2) In scenarios where providing users with these options isn't feasible, a balancing must be made between mitigating cognitive load (favoring font-color with font-weight) versus enhancing user perception of caption efficacy (favoring font-color with font-size). For instance, if we expect users to be distracted by other parallel tasks, e.g., a remote meeting, using font-weight might be more appropriate; in settings where we can expect their wholehearted attention, e.g., watching a movie, font-size might be a better choice.
(3) Though we acknowledge that further work could refine the range of each style's variation, participants' subjective feedback highlighted the need for customizable ranges for each style. Differences in individual preferences, legibility, ease of understanding, and contextual appropriateness are important considerations that can potentially be addressed with personalizable styling and ranges for selected styles.

7 LIMITATIONS & FUTURE WORK

The emotional richness of the SEND videos used in these studies allowed a comprehensive exploration of the visual ranges of each caption style, with their pre-recorded nature providing a high degree of experimental control. However, future work should investigate how affective captions behave under more diverse contexts. How effective would it be with more nuanced stimuli, e.g., those featuring smaller fluctuations in valence and arousal levels, especially in instances of linguistic ambiguity? Would it accommodate multiple speakers? Would it work in video-conferencing settings? In answering these questions, future studies could broaden the generalizability of our findings.

Working with pre-processed videos allowed us to run our studies on any computer participants had available. Thus, real-time performance was not a primary concern during the development of our caption rendering pipeline. Nevertheless, in initial tests on a computer with a high-end GPU (>12GB VRAM), we achieved less than 2s latency for a single stream of audio using OpenAI's Whisper speech recognition model [55], accompanied by modules for word-level timestamping [47], voice activity detection [78], and emotion recognition [70]. However, developing a real-time system capable of processing user-provided audio is a crucial step toward enabling real-world applications of affective captions. Such a system could help study edge cases in these applications, shedding light on potential challenges in settings with people from diverse cultural backgrounds and contexts, such as those who speak too loudly, too quietly, with a non-native accent, etc.

Another important aspect for consideration is the duration of the videos used in our studies. While they were generally short, affective captions may be used for longer periods in settings such as online meetings. Therefore, future work could investigate how participants' reactions to the two top-performing caption styles may be influenced by extended use. One can speculate that longer video durations could also call for adjustments in how each typographic parameter is modulated. For instance, longer viewing periods might warrant subtler and less disruptive visual alterations. Having longer exposure times to each caption style can also impact the learning effects we saw, and the legibility and mental demand measures. These considerations, coupled with how different genres of videos might work better with different caption styles, limit our claims but also inspire future work.

In choosing colors for our studies, we prioritized those resilient to common color vision deficiencies while serving as clear representations of negative, neutral, or positive values. We note that, although participants generally agreed with these choices, our study was conducted within a specific socio-cultural context. While some color-emotion associations may have cross-cultural applicability, they are shaped by linguistic and regional factors. For instance, red's negative connotations in our context could differ in China, where it often carries positive sentiments [36]. This variability extends to other design choices as well. Prosodic and affective captions in Hanzi (Chinese characters) and Hangul (Korean script) share similarities with those in Latin alphabets [30, 39], but differences exist in perception, such as a smaller legibility drop for Hangul [24, 39]. Given these considerations, we highlight that our study was situated within the specific context of ASL/English-speaking, North American DHH culture. While our participants resonated with some choices, caution should be exercised when extrapolating our findings beyond this specific context. Future research could compare design choices across similar conditions within diverse cultural contexts. However, until then, it is prudent to recognize the limitations of generalizability beyond our specific cultural setting.

Although we found strong evidence of subjective preference differences for caption styles, we could not identify the underlying factors that drive these differences. Some of these factors could be tied to demographics, as has been seen in previous studies on typographic preferences [13, 71], but alas our study population was generally too young to uncover if this was the case here. This underscores the need for further research to determine the factors contributing to these preferences. In the interim, we suggest providing users with the option to personalize their caption preferences. For instance, as mentioned earlier, investigating the role of color as a user preference is warranted. Additionally, exploring settings such as minimum and maximum font size and font weight would also be valuable.

Finally, we are aware that affective cues can come from various channels, such as facial expressions, body language, and lip-reading. Having affective captions as an additional channel was perceived by our participants as beneficial. However, it is important to consider whether this addition positively enhances the existing array of affective cues or introduces a certain level of dissonance. This can be compounded if one considers captions employing not only visual elements, as we have explored here, but other channels, such as haptic feedback, e.g., [75]. Gaining insight into this aspect could not only aid the refinement of affective captions but also provide clarity on the contexts where they can be most effective.

8 CONCLUSION

In Study 1, we developed nine distinct caption styles that use typographic modulations to convey emotional dimensions of speech. We asked participants to evaluate how effectively each style depicted either valence or arousal independently. While our primary aim was to discover styles capable of representing both valence and arousal, this initial phase enabled us to rule out seven styles for valence and five for arousal, streamlining our follow-up investigation.

In Study 2, the remaining styles were combined and once again participants assessed their preferences. This time, we focused on styles that could communicate both valence and arousal simultaneously. The winning combinations were font-color with font-weight, font-color with font-size, font-color with shadow-color, and shadow-color with font-color.

Participants indicated that their preferences were guided by their evaluation of whether a given caption style was EASY TO READ, NON-DISTRACTING, INTUITIVE, and provided a CLEAR REPRESENTATION OF EMOTIONS.

In Study 3, we combined these four factors with an emotion-recognition task to collect objective and subjective performance metrics for each of the four winning styles identified in Study 2. We compared these metrics against a neutral baseline. Notably, two styles, font-color with font-weight and font-color with font-size, emerged as well-performing options. The former exhibited lower cognitive load, while the latter was perceived as more effective at conveying emotions. As a result, we recommend both styles as design choices for affective captions.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Nº 1954284, 2125362, 2212303, and 2235405; by the Department of Health and Human Services under Award 90DPCP0002-0100; and by the Fulbright Commission (Fulbright-CAPES Scholarship, ME / CAPES Nº 8 / 2020).

The authors acknowledge the use of OpenAI's ChatGPT and Google's Bard as editorial tools to refine the clarity and style of this manuscript's presentation.

REFERENCES

[1] Aviad Albert, Francesco Cangemi, and Martine Grice. 2018. Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In *Proceedings Speech Prosody*, Vol. 9. International Speech Communication Association (ISCA), International Speech Communications Association, Poznań, Poland, 804–808. https://doi.org/10.21437/SpeechProsody.2018-162

[2] 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

[3] Akhter Al Amin, Saad Hassan, Sooyeon Lee, and Matt Huenerfauth. 2023. Understanding How Deaf and Hard of Hearing Viewers Visually Explore Captioned Live TV News. In *Proceedings of the 20th International Web for All Conference* (<conf-loc>, <city>Austin</city>, <state>TX</state>, <country>USA</country>, </conf-loc>) *(W4A '23)*. Association for Computing Machinery, New York, NY, USA, 54–65. https://doi.org/10.1145/3587281.3587287

[4] Akher Al Amin, Joseph Mendis, Raja Kushalnagar, Christian Vogler, Sooyeon Lee, and Matt Huenerfauth. 2022. Deaf and Hard of Hearing Viewers' Preference for Speaker Identifier Type in Live TV Programming. In *Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 200–211.

[5] Rasmus Andersson. 2023. The Inter typeface family. https://rsms.me/inter/. [Online; accessed 22-November-2023].

[6] Lyn Bartram, Abhisekh Patra, and Maureen Stone. 2017. Affective Color in Visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1364–1374. https://doi.org/10.1145/3025453.3026041

[7] Larwan Berke. 2017. Displaying Confidence from Imperfect Automatic Speech Recognition for Captioning. *SIGACCESS Access. Comput.* 117 (feb 2017), 14–18. https://doi.org/10.1145/3051519.3051522

[8] Larwan Berke, Matthew Seita, and Matt Huenerfauth. 2020. Deaf and Hard-of-Hearing Users' Prioritization of Genres of Online Video Content Requiring Accurate Captions. In *Proceedings of the 17th International Web for All Conference* (Taipei, Taiwan) *(W4A '20)*. Association for Computing Machinery, New York, NY, USA, Article 3, 12 pages. https://doi.org/10.1145/3371300.3383337

[9] Ann Bessemans, Maarten Renckens, Kevin Bormans, Erik Nuyts, and Kevin Larson. 2019. Visual prosody supports reading aloud expressively. *Visible Language* 53, 3 (2019), 28–49.

[10] Kerry Bodine and Mathilde Pignol. 2003. Kinetic Typography-Based Instant Messaging. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 914–915. https://doi.org/10.1145/765891.766067

[11] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2005. Affect: From Information to Interaction. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* (Aarhus, Denmark) *(CC '05)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/1094562.1094570

[12] Janine Butler, Brian Trager, and Byron Behm. 2019. Exploration of Automatic Speech Recognition for Deaf and Hard of Hearing Students in Higher Education Classes. 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, 32–42. https://doi.org/10.1145/3308561.3353772

[13] Tianyuan Cai, Shaun Wallace, Tina Rezvanian, Jonathan Dobres, Bernard Kerr, Samuel Berlow, Jeff Huang, Ben D. Sawyer, and Zoya Bylinskii. 2022. Personalized Font Recommendations: Combining ML and Typographic Guidelines to Optimize Readability. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1–25. https://doi.org/10.1145/3532106.3533457

[14] João Couceiro e Castro, Pedro Martins, Ana Boavida, and Penousal Machado. 2019. «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

[15] Qinyue Chen, Yuchun Yan, and Hyeon-Jeong Suk. 2021. Bubble Coloring to Visualize the Speech Emotion. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 361, 6 pages. https://doi.org/10.1145/3411763.3451698

[16] Qinyue Chen, Yuchun Yan, and Hyeon-Jeong Suk. 2022. Designing voice-aware text in voice media with background color and typography. *Journal of the International Colour Association* 28 (2022), 56–62.

[17] Andrew P Clark, Kate L Howard, Andy T Woods, Ian S Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PloS one* 13, 1 (2018), e0190393.

[18] Michael Cooper. 2021. W3C Accessibility Guidelines (WCAG) 3.0. https://www.w3.org/TR/wcag-3.0/

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

[20] Bruyne L. De, De Clercq Orphée, and Hoste Véronique. 2021. Annotating affective dimensions in user-generated content. *Language Resources and Evaluation* 55, 4 (12 2021), 1017–1045. https://ezproxy.rit.edu/login?url=https://www.proquest.com/scholarly-journals/annotating-affective-dimensions-user-generated/docview/2580827900/se-2 Copyright - © The Author(s), under exclusive licence to Springer Nature B.V. part of Springer Nature 2021; Last updated - 2021-12-24.

[21] Caluã de Lacerda Pataca. 2023. *Speech-modulated typography.* Master's thesis. University of Campinas School of Electrical and Computer Engineering. https://doi.org/10.31237/osf.io/yu5dn

[22] 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

[23] 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

[24] Caluã de Lacerda Pataca, Matthew Watkins, Roshan Peiris, Sooyeon Lee, and Matt Huenerfauth. 2023. Visualization of Speech Prosody and Emotion in Captions: Accessibility For Deaf And Hard-of-Hearing Users. , 15 pages. https://doi.org/10.1145/3544548.3581511

[25] Joshua R. de Leeuw, Rebecca A. Gilbert, and Björn Luchterhandt. 2023. jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software* 8, 85 (May 2023), 5351. https://doi.org/10.21105/

joss.05351
[26] Arpad E Elo. 1978. *The rating of chessplayers, past and present.* Arco Pub., New York.
[27] Jodi Forlizzi, Johnny Lee, and Scott Hudson. 2003. The Kinedit System: Affective Messages Using Dynamic Texts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 377–384. https://doi.org/10.1145/642611.642677
[28] Sandrine Gil and Ludovic Le Bigot. 2016. Colour and emotion: children also associate red with negative valence. *Developmental science* 19, 6 (2016), 1087–1094.
[29] Michael Gower, Brent Shiver, Charu Pandhi, and Shari Trewin. 2018. Leveraging Pauses to Improve Video Captions. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 414–416. https://doi.org/10.1145/3234695.3241023
[30] Kaixin Han, Weitao You, Heda Zuo, Mingwei Li, and Lingyun Sun. 2023. Glancing back at your hearing: Generating emotional calligraphy typography from musical rhythm. *Displays* 80 (2023), 102529. https://doi.org/10.1016/j.displa.2023.102529
[31] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. North-Holland, Amsterdam, 139–183. https://doi.org/10.1016/s0166-4115(08)62386-9
[32] 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* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 125, 10 pages. https://doi.org/10.1145/3544549.3585880
[33] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill(TM): A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 20* (advances in neural information processing systems 20 ed.). MIT Press, Cambridge, Massachusetts, 569–576. https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/
[34] Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht, Harald C. Traue, and Henrik Kessler. 2012. Mapping discrete emotions into the dimensional space: An empirical approach. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Seoul, South Korea, 3316–3320. https://doi.org/10.1109/ICSMC.2012.6378303
[35] Kristina Höök, Anna Ståhl, Petra Sundström, and Jarmo Laaksolaahti. 2008. Interactional Empowerment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 647–656. https://doi.org/10.1145/1357054.1357157
[36] 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.
[37] 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
[38] 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 *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
[39] JooYeong Kim, SooYeon Ahn, and Jin-Hyuk Hong. 2023. Visible Nuances: A Caption System to Visualize Paralinguistic Speech Cues for Deaf and Hard-of-Hearing Individuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 54, 15 pages. https://doi.org/10.1145/3544548.3581130
[40] Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. *CoRR* abs/1712.01765 (2017), 465–470. arXiv:1712.01765 http://arxiv.org/abs/1712.01765
[41] Svetlana Kiritchenko and Saif M. Mohammad. 2017. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best-Worst Scaling. arXiv:1712.01741 http://arxiv.org/abs/1712.01741
[42] Christof Kuhbandner and Reinhard Pekrun. 2013. Joint effects of emotion and color on memory. *Emotion* 13, 3 (2013), 375.
[43] Daniel G Lee, Deborah I Fels, and John Patrick Udo. 2007. Emotive captioning. *Computers in Entertainment (CIE)* 5, 2 (2007), 11.
[44] Heungsub Lee. 2018. Computing Your Skill. https://trueskill.org/ [Online; accessed 29-April-2023].
[45] Einat Liebenthal, David A Silbersweig, and Emily Stern. 2016. The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. *Frontiers in neuroscience* 10 (2016), 506.
[46] Fernando Loizides, Sara Basson, Dimitri Kanevsky, Olga Prilepova, Sagar Savla, and Susanna Zaraysky. 2020. Breaking Boundaries with Live Transcribe: Expanding Use Cases Beyond Standard Captioning Scenarios. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Virtual Event Greece, 1–6. https://doi.org/10.1145/3373625.3417300
[47] Jérôme Louradour. 2023. whisper-timestamped. https://github.com/linto-ai/whisper-timestamped.
[48] James R. Mallory, Michael Stinson, Lisa Elliot, and Donna Easton. 2017. Personal Perspectives on Using Automatic Speech Recognition to Facilitate Communication between Deaf Students and Hearing Customers. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 419–421. https://doi.org/10.1145/3132525.3134779
[49] Emma J. McDonnell, Ping Liu, Steven M. Goodman, Raja Kushalnagar, Jon E. Froehlich, and Leah Findlater. 2021. Social, Environmental, and Technical: Factors at Play in the Current Use and Future Design of Small-Group Captioning. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–25. https://doi.org/10.1145/3479578
[50] Robert M Ochshorn and Max Hawkins. 2015. Gentle: a robust yet lenient forced aligner built on Kaldi. https://lowerquality.com/gentle/
[51] Alp Öktem, Mireia Farrús, and Leo Wanner. 2017. Prosograph: a tool for prosody visualisation of large speech corpora. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*. International Speech Communication Association (ISCA), ISCA, Stockholm, Sweden, 809–810.
[52] Desmond C. Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2021. Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing* 12, 3 (2021), 579–594. https://doi.org/10.1109/TAFFC.2019.2955949
[53] Bryan Orme. 2009. *Maxdiff analysis: Simple counting, individual-level logit, and hb.* Technical Report. Sawtooth Software.
[54] Suksumek Promphan. 2017. *Emotional Type: Emotional expression in text message.* Master's thesis. Basel School of Design, Switzerland.
[55] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arXiv:2212.04356
[56] Carlos Ramos-Carreño. 2022. dcor: distance correlation and energy statistics in Python. https://doi.org/10.5281/zenodo.3468124
[57] Carlos Ramos-Carreño and José L. Torrecilla. 2023. dcor: Distance correlation and energy statistics in Python. *SoftwareX* 22 (2 2023), 101326. https://doi.org/10.1016/j.softx.2023.101326
[58] 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.
[59] 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
[60] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
[61] Tim Schlippe, Shaimaa Alessai, Ghanimeh El-Taweel, Matthias Wölfel, and Wajdi Zaghouani. 2020. Visualizing voice characteristics with type design in closed captions for arabic. In *2020 International Conference on Cyberworlds (CW)*. IEEE, IEEE, Caen, France, 196–203.
[62] Jocelyn J Shen, Kathryn Jin, Ann Zhang, Cynthia Breazeal, and Hae Won Park. 2023. Affective Typography: The Effect of AI-Driven Font Design on Empathetic Story Reading. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 26, 7 pages. https://doi.org/10.1145/3544549.3585625
[63] Hyeon-Jeong Suk and Hans Irtel. 2010. Emotional response to color across media. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 35, 1 (2010), 64–77.
[64] 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 (2016), 686–728.
[65] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35, 6 (Dec. 2007), 2769–2794. https://doi.org/10.1214/009053607000000505
[66] Alexander Toet. 2022. *EmojiGrid.* Open Science Framework. https://doi.org/10.17605/OSF.IO/H82YB
[67] Alexander Toet, Daisuke Kaneko, Shota Ushiama, Sofie Hoving, Inge de Kruijf, Anne-Marie Brouwer, Victor Kallen, and Jan B. F. van Erp. 2018. EmojiGrid: A

2D Pictorial Scale for the Assessment of Food Elicited Emotions. *Frontiers in Psychology* 9 (Nov. 2018), 1–21. https://doi.org/10.3389/fpsyg.2018.02396

[68] Alexander Toet and Jan B. F. van Erp. 2019. The EmojiGrid as a Tool to Assess Experienced and Perceived Emotions. *Psych* 1, 1 (Sept. 2019), 469–481. https://doi.org/10.3390/psych1010036

[69] Alexander Toet and Jan B. F. van Erp. 2021. Affective rating of audio and video clips using the EmojiGrid. *F1000Research* 9 (April 2021), 970. https://doi.org/10.12688/f1000research.25088.2

[70] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 1–13. https://doi.org/10.1109/TPAMI.2023.3263585

[71] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 38 (mar 2022), 56 pages. https://doi.org/10.1145/3502222

[72] Shaun Wallace, Rick Treitman, Jeff Huang, Ben D. Sawyer, and Zoya Bylinskii. 2020. Accelerating Adult Readers with Typeface: A Study of Individual Preferences and Effectiveness. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382985

[73] James M. Waller and Raja S. Kushalnagar. 2016. Evaluation of Automatic Caption Segmentation. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 331–332. https://doi.org/10.1145/2982142.2982205

[74] Jianji Wang and Nanning Zheng. 2020. Measures of Correlation for Multiple Variables. arXiv:1401.4827 [math.ST]

[75] Yiwen Wang, Ziming Li, Pratheep Kumar Chelladurai, Wendy Dannels, Tae Oh, and Roshan L Peiris. 2023. Haptic-Captioning: Using Audio-Haptic Interfaces to Enhance Speaker Indication in Real-Time Captions for Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 781, 14 pages. https://doi.org/10.1145/3544548.3581076

[76] Alex Wennberg, Henrik Åhman, and Anders Hedman. 2018. The Intuitive in HCI: A Critical Discourse Analysis. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (Oslo, Norway) *(NordiCHI '18)*. Association for Computing Machinery, New York, NY, USA, 505–514. https://doi.org/10.1145/3240167.3240202

[77] David Wicks. 2017. The coding manual for qualitative researchers. *Qualitative research in organizations and management: an international journal* 12, 2 (2017), 169–170.

[78] John Wiseman. 2021. py-webrtcvad. https://github.com/wiseman/py-webrtcvad.

[79] 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, IEEE, Bucharest, Romania, 1–9.